NBER WORKING PAPER SERIES

NOISY EXPERTS? DISCRETION IN REGULATION

Sumit Agarwal
Bernardo C. Morais
Amit Seru
Kelly Shue

Working Paper 32344
http://www.nber.org/papers/w32344

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2024

We thank Nick Barberis, Jonathan Berk, John Cochrane, James Choi, Claudia Robles-Garcia, Matt Gentzkow, Chad Jones, Pete Klenow, Eddie Lazear, Lasse Pedersen, Andrei Shleifer, Adi Sunderam, Richard Thaler, Ali Yurukoglu, Jeff Zwiebel, and participants in the Econometric Society (Behavioral Finance) meetings as well as seminars at the American Enterprise Institute, 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 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. First Version: November 2019. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Noisy Experts? Discretion in Regulation
Sumit Agarwal, Bernardo C. Morais, Amit Seru, and Kelly Shue
NBER Working Paper No. 32344
April 2024
JEL No. G28,G4

ABSTRACT

While reliance on human discretion is a pervasive feature of institutional design, human discretion can also introduce costly noise (Kahneman, Sibony, and Sunstein 2021). We evaluate the consequences, determinants, and trade-offs associated with discretion in high-stake decisions assessing bank safety and soundness. Using detailed data on the supervisory ratings of US banks, we find that professional bank examiners exercise significant personal discretion—their decisions deviate substantially from algorithmic benchmarks and can be predicted by examiner identities, holding bank fundamentals constant. Examiner discretion has a large and persistent causal impact on future bank capitalization and supply of credit, leading to volatility and uncertainty in bank outcomes, and a conservative anticipatory response by banks. We identify a novel source of noise: weights assigned to specific issues. Disagreement in ratings across examiners can be attributed to high average weight (50%) assigned to subjective assessment of banks' management quality, as well as heterogeneity in weights attached to more objective issues such as capital adequacy. Replacing human discretion with a simple algorithm leads to worse predictions of bank health, while moderate limits on discretion can translate to more informative and less noisy predictions.

Sumit Agarwal
National University of Singapore
Mochtar Raidy Building
15 Kent Ridge Drive
Singapore
ushakri@yahoo.com

Amit Seru
Stanford Graduate School of Business
Stanford University
655 Knight Way
and NBER
aseru@stanford.edu

Bernardo C. Morais
Federal Reserve Board
Mail Stop 18
20th & C St. NW
Washington, DC 20551
bernardo.c.morais@frb.gov

Kelly Shue
Yale School of Management
165 Whitney Avenue
P.O. Box 208200
New Haven, CT 06520-8200
and NBER
kelly.shue@yale.edu

1. 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)

Reliance on human discretion is a pervasive feature of institutional design. Judges, doctors, scientific reviewers, patent examiners, HR committees, and government regulators are not fully rule bound. Institutions are designed so that these decision-makers have the flexibility to draw upon their personal insights and interpretation of evidence when making decisions. Allowing for human discretion has obvious benefits: humans have a unique ability to process soft information, i.e., real information that is difficult to quantify and open to interpretation (Petersen and Rajan, 1994, 2002; Stein, 2002; Liberti and Mian, 2009 and Rajan, Seru, and Vig, 2015). Replacing human discretion with rules or machines may lead to decisions that fail to account for the nuanced merits of each case.

On the flip side, Kahneman, Sibony, and Sunstein (2021) argue that reliance on human discretion leads to the problem of "noise," defined as disagreement among professionals considering the same information. The same case, assigned to a different decision-maker with a different set of prior beliefs, biases, and abilities could receive an entirely different decision. Thus, discretion can inject volatility and unpredictability into the system. Moreover, if each case has an optimal decision, distortions away from the optimal outcome due to human discretion can be very costly.[1] As emphasized in Kahneman et al., the costs associated with judicial decisions that are too strict for one case and too lenient for another case do not average away, but rather add up. A similar concern applies to supervision of the US banking sector, as we will show in this paper.

The potential for discretion to create costly noise is underscored by research in psychology showing that humans attach too much weight to their personal insights and interpretation of soft information (e.g., research on "algorithm aversion" by Meehl (1954), Dawes (1979), Dawes, Faust, and Meehl (1989), and Huang and Pearce (2015)). These studies find that simple algorithmic forecasts outperform human forecasts.[2] In addition, human decision-makers have been shown to suffer from a variety of behavioral biases, limitations in skill, and agency conflicts which can lead to mistakes and disagreement (e.g., Tversky and Kahneman, 1974; Chan, Gentzkow, and Yu, 2022; Agarwal et al. 2014). Finally, algorithmic forecasts may be harder to beat in modern times due to advances in machine learning (Kleinberg et al., 2017; Hoffman, Kahn, and Li, 2018; Mullainathan and

[1] See the Appendix for a simple framework illustrating why noise can be costly.

[2] Reliance on human judgment over algorithms can also lead to relatively more discrimination against disadvantaged minorities (e.g., Yang, 2015; Kleinberg, Ludwig, Mullainathan, and Rambachan, 2018; Benson, Li, and Shue, 2021; Arnold, Dobbie, and Yang, 2018; and Arnold, Dobbie, and Hull, 2022).

Obemeyer, 2022; Angelova, Dobbie, and Yang, 2022 and Fuster et al., 2022).

Empirical research on human discretion by professionals in high-stakes field settings has so far been limited and focuses primarily on judicial or medical settings.[3] In this paper, we evaluate the consequences, determinants, and trade-offs associated with discretion in a high-stakes setting in which professional bank examiners evaluate the safety and soundness of US banks. Government regulation and supervision of the US banking sector is a key determinant of bank capitalization, credit supply, and ultimately the amount of growth and risk in the macro-economy. An important pillar of banking rules and regulation around them is the supervisory ratings assigned by bank examiners during routine bank examinations. Using detailed confidential data on the supervisory ratings decisions that govern these banks, we measure examiner discretion as the extent to which bank supervisory decisions deviate from algorithmic benchmarks or can be predicted by examiner identities and past experiences, holding bank fundamentals constant. We find that examiner discretion is quantitatively large and injects a significant degree of noise into the US banking system. We then provide evidence that this noise is potentially costly.

Our analysis uses micro-data on the decisions made by bank examiners who conduct annual "safety and soundness" on-site examinations of US banks. As part of the examination, a lead examiner along with her team reviews documents from the bank, evaluates its loan portfolio, and meets with the bank's management. The examinations result in the assignment of a composite CAMELS rating, which reflects the overall condition of the bank. 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 discretion 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 concerns. Composite CAMELS ratings have significant consequences for future bank activity. They are a key determinant of each bank's FDIC insurance premium, access to the Federal Reserve discount window, and licensing, branching, and merger approval.[4] These ratings also serve as a key variable in our analysis.

To trace the causal impact of discretion on outcomes, we limit our empirical tests to the subsample of banks that are subject to regular examiner rotation. These banks together have assets worth 6.5 trillion dollars (~32% of total assets in the sector). While the largest banks are excluded from our analysis because they are not subject to regular examiner rotation, our sample does include

[3] See, e.g., Ramji-Nogales, Schoenholtz, Schrag (2010), and Yang (2015). For a review, see Kahneman, Sibony, and Sunstein (2021).

[4] Banks with unsatisfactory composite ratings of 3 or above generally increase capital ratios and reduce lending, actions that can ultimately affect the supply of credit in the economy.

important regional banks such as Silicon Valley Bank and First Republic Bank, whose recent collapse in 2023 triggered a regulatory crisis. Banks in our sample are rotated across examiners associated with their state or region, subject to examiner workload limitations (Agarwal et al., 2014). Our analysis relies on the assumption that the assignment of lead examiners to banks is unrelated to the bank's true quality, within a state and time period. Indeed, we empirically confirm that an examiner's tendency to be more lenient or tough (calculated excluding the current examination), is uncorrelated with observable bank health as well as the future performance of the bank's existing loan portfolio.

We measure the discretion associated with each exam as the residual component of the rating decision that cannot be predicted by observable bank characteristics. Defined in this way, discretion measures the bank's soft information, as well as the examiner's personal influence on the interpretation and synthesis of all case information, soft and hard.[5] Since a higher rating represents greater safety and soundness concern, a positive residual implies that the examiner gave a tougher decision than predicted by bank observables, and vice versa. A non-zero measure of discretion does not necessarily mean the decision was not well-matched to case merits. Discretion can reflect the examiner's interpretation of case-specific soft information. However, quasi-random rotation implies that the soft information for each bank should be uncorrelated with examiner assignment within a state-time period. Therefore, the extent to which our measure of discretion varies predictably across examiners is evidence of noise introduced by them.

Using the exam-level measure of discretion, we then construct two measures of discretion at the examiner level: directional and absolute examiner discretion. These two measures of discretion are related, yet conceptually distinct, and can be tested in different ways. We define examiner "directional discretion" as the average of the signed values of discretion across all exams conducted by each examiner. Directional discretion captures how some examiners are predictably more lenient than others. We define examiner "absolute discretion" as the average of the *absolute value* of discretion across all exams conducted by each examiner. Absolute discretion measures of the extent to which the examiner relies on case-specific soft information, as well as any biases, gut feelings, or intuition. Notably, it is possible for an examiner to exercise zero directional discretion (so she is not more lenient than other examiners) and have high absolute discretion (because she heavily weighs soft information or gut feelings in either direction when evaluating each case).

Overall, we find economically large levels of and variation in absolute discretion among examiners, as well as wide variation in directional discretion. For instance, the average level of

[5] While soft information can be thought of as real information about bank quality, we classify soft information as part of discretion because its usage is dependent upon human interpretation. Indeed, it is not possible to empirically measure soft information in the absence of discretion because soft information's key feature is that it is open to human interpretation.

absolute discretion by examiners is 0.12, which is 20% larger than the difference in ratings induced by Federal versus State regulatory agencies which was shown to be consequential in Agarwal et al. (2014). In simulations, we estimate that healthy banks that would receive a rating of 2 (which compose the majority of our sample) in the absence of discretion, are exposed to a 4.2% probability per exam of being rated an unsatisfactory 3 or higher due to examiner assignment. This magnitude is large compared to the overall transition probability of 6.7% of a bank moving from a rating of 2 to a 3. Affected banks would be subjected to substantial additional regulatory oversight, due purely to examiner discretion. Likewise, 5.0% of banks that would have gotten a rating of 2 absent discretion receive a rating of 1. These banks would end up not being supervised as intensively as they should. This illustrates why noise introduced in supervision across banks may not cancel each other.

Next, we exploit the examiner rotation system to estimate the causal impact of examiner discretion on bank behavior. We instrument for the rating of each exam using the leave-out-mean residual rating of each examiner. The intuition is that an examiner who assigns high ratings to other banks is more likely to assign a high rating to the current bank, in a way that is uncorrelated with the current bank's fundamentals. Using this variation, we estimate that an exogenous one-point increase in ratings due to examiner discretion causes a 24% increase in a bank's capitalization relative to the sample median and a large reduction in loan growth equivalent to 0.88 sample standard deviations within one year after the exam.[6] These effects persist and impact bank capitalization and loan growth two and three years into the future. These causal estimates also imply that any quasi-random assignment system of examiners to banks will increase volatility and uncertainty in bank outcomes, as bank capitalization and lending will vary depending on which examiner was assigned.

In addition to influencing bank behavior ex post, we provide suggestive evidence that examiner discretion induces an anticipatory response. Banks located in states where examiners exercise a high degree of absolute discretion relative to the national average (such that examiner ratings are relatively difficult to predict) appear to take precautionary measures by maintaining more capital and lower loan growth than other banks with similar observable fundamentals. Notably, we condition on the state's level of past ratings. Doing so accounts for these effects being driven by past supervision which may have required banks to increase their capitalization. By engaging in these conservative actions, banks reduce the probability of getting a bad rating decision should they be matched to a tough examiner in the future. This conservative response in anticipation of uncertainty is related to the mechanism in Bloom (2009), in which firms reduce investment and hiring due to

[6] These causal effects of ratings contrast sharply with the non-causal predictive power of ratings: An endogenously higher rating today predicts 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, which we trace using our IV strategy, is uncorrelated with current bank fundamentals and causes a lower rating next year because the bank responds to the higher supervisory rating by taking conservative actions.

macroeconomic uncertainty. It is also related to evidence from Gissler, Oldfather, and Ruffino (2016) showing that uncertainty over regulatory rules caused banks to reduce lending. We show that human discretion by regulators can introduce a different type of uncertainty that may dampen credit supply.

After establishing the consequences of examiner discretion, we then assess the determinants of discretion. We explore the question of why disagreement occurs. Examiners are trained professionals who perform on-site examinations as part of their full-time jobs. When exposed to equivalent information,[7] why do they disagree? We find that the weights assigned to specific issues are a major contributor to disagreement. The process of reaching a final decision can be modeled as a set of judgments on specific issues as well as choices over weights assigned to each issue. A decision-maker may assign some issues more weight because she believes that the issue is more important or is estimated with greater precision. In our setting, examiners have discretion over how component ratings are aggregated to form the composite rating. A key advantage of our data is that we observe composite CAMELS ratings as well as each of the six component ratings, allowing us to explore how weights assigned to specific issues contribute to disagreement. Notably, our estimates of weights do not require any assumptions about the randomness of examiner assignment to banks or measurement of bank observables.

Our analysis reveals two sources of disagreement. First, examiners place the most weight, approximately 50 percent, on a highly subjective component of the CAMELS rating: management quality. The high weight placed on perceptions of management quality is consistent with psychology research showing that people place too much weight on face-to-face interactions or perceptions drawn from facial expressions and language (Levine, McCornack and Park, 1999; Ludwig and Mullainathan, 2023). Relatedly, examiners may overweight salient interactions with management (Bordalo, Gennaioli, and Shleifer, 2015). Second, examiners exhibit heterogeneity in how they weight different components, even relatively objective ones such as capital adequacy. Heterogeneity in weights across examiners can lead to disagreement in the final decision even if the examiners agreed on all component ratings. These findings imply that efforts to reach consensus on specific issues may not lead to consensus for the final verdict, as long as decision-makers continue to differ in how they weigh specific issues and/or continue to heavily weight a highly subjective issue.

Our results showing that examiners assign high weight to management quality is particularly interesting in light of calls after the failure of Silicon Valley Bank (SVB) for bank examiners to hold

[7] We never observe cases in which multiple examiners review identical bank information, because only one lead examiner is assigned to a bank per examination. We instead estimate disagreement by exploiting examiner rotation. Within a region over time, no examiner should be systematically assigned to worse banks in terms of observable or unobservable characteristics. Therefore, examiners within a region observe "equivalent" information plus random noise. If one examiner predictably issues higher ratings or ratings with greater variance, then that examiner disagrees with another examiner after observing equivalent information.

even greater discretionary powers. Prior to its collapse, SVB received an unsatisfactory CAMELS rating of 3 precisely because of high weight assigned to its management rating of 3 (SVB received a fair rating of 2 on all other CAMELS components).[8] To the extent that it was desirable for examiners to provide warnings about SVB, this showcases an instance when it was beneficial to allow examiners discretion to heavily weight their subjective assessments of management quality.

In the final part of the paper, we assess the trade-offs associated with limiting or replacing examiner discretion with more rule-based or machine-driven decision-making. We begin by examining whether discretion in ratings is useful in predicting future bank problems—an important purpose of these ratings. Identifying the predictive power of endogenous ratings is empirically challenging because CAMELS rating can both (i) affect bank outcomes by forcing banks to take conservative actions, as identified through our instrumental variables method and (ii) predict bank problems by reflecting real safety concerns. To identify predictive power, we examine whether the discretionary component of ratings can predict bank outcomes that are unlikely to be directly affected by ratings. We do so by focusing on near-term changes in the performance of loans that were made prior to the current examination. We also test if ratings predict whether the bank will receive a high rating in the next examination. This test exploits the fact that the predictive power and the causal effect of the current rating operate in opposite directions: a higher rating should predict a higher rating in the future but, due to supervisory actions, can cause a reduction in the future rating.

We find that the discretionary component of ratings, i.e., the residual that cannot be projected on to observable bank characteristics, effectively predicts near-term changes in bank outcomes in the direction of the bank becoming less safe. It also predicts increases in the bank's future rating. These results imply that discretion allows examiners to use soft information to better predict changes in bank outcomes. We can also reject strong conclusions in Meehl (1954) and Dawes (1979) about the inability of human professionals to use soft and hard information to make effective forecasts. Contrary to their analyses which showed that humans underperform even simple unit-weighted linear models (simple linear regressions with coefficients constrained to equal unity), we find that examiner ratings outperform a simple algorithm that uses only observable bank characteristics to predict bank outcomes.[9] Thus, there seems to be a tradeoff when deciding how much discretion to

[8] 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

[9] We purposely do not perform a "race" between examiner ratings and more advanced and complex algorithmic models. Given the extremely rapid rate of progress in AI and machine learning methods, such a test is unlikely to be informative. 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 (indeed there are no current policy proposals under practical consideration to replace human bank examiners or judges

allow: discretion introduces noise in bank outcomes but also allows for the processing of soft information, leading to more predictive ratings.[10]

Although discretion adds value on average relative to a simple algorithmic benchmark, our analysis also suggests that modest restrictions on discretion can decrease noise without sacrificing predictive power. We show that examiners who exercise high discretion do not outperform others in their ability to predict future bank outcomes. Thus, high-discretion examiners introduce noise without adding predictive power. We also find that a reweighted average of component ratings—that effectively puts restrictions on discretion—outperforms the actual composite rating in predicting changes in loan performance. The adoption of restrictions or guidelines on weights in bank supervision would be similar in spirit to US federal sentencing guidelines which constrain judicial decision-making without entirely removing judge discretion (Yang 2015). Our results suggest that placing moderate limits on examiner discretion can have a similar effect in producing less noisy and more predictive ratings. This evidence also speaks to the vast literature studying governance of banking institutions (e.g., see Laeven and Levine, 2009 and Huber, 2021).

Our paper complements the very large body of research showing that decision-making is influenced by psychological biases, mistaken beliefs, inherited traits, personal experiences, and agency incentives (e.g., Tversky and Kahneman ,1974; Bohren et al., 2023). In contrast to studies that zero in on a specific determinant of human decision-making, we posit that there may be substantial heterogeneity across individuals in the large set of potential factors that can influence decision-making. This heterogeneity can lead to disagreement and introduce costly noise. By examining a high stakes setting involving supervision of trillions of dollars of US banking assets, we shed light on the consequences and tradeoffs associated with this heterogeneity. We also approach the question of what determines disagreement from a new angle. Instead of focusing on a particular psychological, inherited, experience, or agency factor, we show that the channel through which disagreement manifests is through discretion over weights on issues: disagreement in the final decision arises because of heavy weight on highly subjective issues as well as heterogeneity in weights across issues.

Our analysis contributes to the related literatures on algorithm aversion and human decision-making biases in three ways. First, our empirical setting does not have to deal with 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 et al. (2018), and Benson, Li, and Shue (2019). In those studies, the econometrician only observes outcomes for affirmative decisions. For example, we only observe

with machines). Thus, we instead test whether imposing modest restrictions on human discretion can improve predictive power, while reducing noise.

[10] Allowing human decision-makers the *option* to rely on an algorithm may not produce the desired results. For example, Hoffman et al. (2018) show that hiring managers choose to overrule algorithmic recommendations even when they do not have superior information.

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 component and composite ratings allows us to explore how discretion over weights can lead to disagreement. Third, we show how human discretion interacts with government regulation and can potentially influence both the ex post and ex ante credit supply of US banks. While behavioral biases among households are often cited as a justification for increased regulatory oversight, our analysis suggests that government regulators may also exhibit a variety of biases in their personal judgments which can significantly affect how regulations are implemented.

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 guarantees *fairness*, in that no individual or entity systematically faces more lenient judgment.[11] However, random assignment does not address the deeper issue that human decision-makers can make noisy judgments that are not well matched to case merits (Lipsky, 2010). At the extreme, one can imagine that substantial welfare losses would occur under a regime in which all case outcomes were determined by fair but arbitrary coin flips. Our estimation of the causal impact of examiner discretion is also related to empirical research using judge fixed effects as an instrumental variable to estimate the causal effect of bankruptcy protection on household and firm outcomes (e.g., Dobbie and Song, 2015; Bris, Welch, and Zhu, 2006; Chang and Schoar, 2006; and Sampat and Williams 2019). The fact that judge assignment is such a powerful instrument for case decisions is evidence that human discretion can be very distortionary.

Finally, our paper speaks to the broad literature that has studied the optimal design of regulation and its enforcement (e.g., Shleifer and Vishny 1999). Within the context of banking, much of the existing research has focused on the role regulatory arbitrage plays in such design (e.g., Agarwal et al. 2014; Calomiris and Gorton 1991; Calomiris 2006 and Garicano 2012). We illustrate that human discretion of regulators is another important dimension that can potentially impact regulatory outcomes and should be a central consideration in the design of regulation.

2. Institutional background

2.1 Related literature on CAMELS ratings

This paper is most closely 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

[11] Perhaps because of these advantages, assessment of the quality of institutions has largely focused on the extent to which case assignment succeeds or fails 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/.

focus and findings differ in several important ways. First, Agarwal et al. (2014) zoom in on one specific reason driving one group of examiners to give tougher 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 examiner) decision-making. Thus, we examine the determinants, consequence, and trade-offs associated with examiner discretion. Moreover, we investigate all forms of discretion, including absolute discretion which can manifest through increased uncertainty rather than predictable differences in means across examiners. Our findings on both of these aspects set us apart from Agarwal et al. (2014).[12] Notably, we also show that, *even within the set of state or federal examiners*, there exists large individual variation in examiner discretion. 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).

2.2 An overview of US banking regulation

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

[12] These effects from individual discretion that we document are also quantitatively larger than those in Agarwal et al. (2014).

conducted every 18 months.[13]

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 3, 4, and 5 ratings present moderate to extreme levels of regulatory concerns and face much stricter supervision than those with ratings of 1 or 2. In this paper, we refer to composite ratings of 3, 4, or 5 as unsatisfactory ratings.

Aside from providing a central 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.

2.3 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

[13] 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.

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 components. 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, subject to examiner workload limitations and qualifications. 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 argue that conditional on a state-year-quarter, the assignment of lead examiners to banks should be approximately uncorrelated with a bank's true quality (both observable and unobservable). Because assignment is not completely random due to considerations for examiner workload, we also empirically support this assumption by showing that observable measures of bank quality, as well as future changes in the performance of the bank's existing loan portfolio, 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 and concurrent examinations 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 with no signs of supervision rotation do not show up systematically within the sample—they are spread out across states and over time. 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 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 whether male examiners exercise greater discretion than female examiners. 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 weights these examiners put on specific issues is a major contributor to this heterogeneity.

3. Results

3.1 Summary statistics

We begin by presenting summary statistics of our data in Table 1. Our data covers the period from 1998 to the first quarter of 2020. Panel A shows that we observe 2,272 distinct lead examiners, and 13,020 examinations in our cleaned data sample.[14] 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 11% 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 7.9 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 use this timing convention to ensure that the bank variables used in our analysis are indeed observable to the examiner at the time that they are assessing the bank's safety and soundness. Note that if we instead used bank observables as of the quarter end after the rating is released, the variables could reflect information that was released after the examination was completed. *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-

[14] 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 purposely excluded from our sample because they fall outside the regular rotational schedule, and (3) examiners have a relatively high rate of rate turnover.

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 1.4 billion US$ and median assets of 184 million US$. Note that because we limit our sample to banks subject to regular examiner rotation, we exclude the largest set of banks in the US economy. Having said that, 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 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 the largest set of 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.

3.2 Measuring discretion

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

$$Rating_{ijrt} = X_{it} + \gamma_{rt} + \varepsilon_{ijrt} \quad (1)$$

Observations are at the exam level, and is indexed for bank *i*, examiner *j*, state *r*, and year-quarter *t*. $Rating_{ijrt}$ is the composite CAMELS rating, X_{it} represents the set of 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. Specifically, we regress the composite camels rating on the set of bank variables described in Panel D of Table 1. The results of this regression are presented in Column 1 of Table 4.

We define $RatingPred_{ijrt}$ as the predicted value from this regression, which represents the expected rating given bank observables and region-time trends in the data. $RatingPred_{ijrt}$ does not necessarily represent the "correct" decision for each examination. It nevertheless provides a useful benchmark because deviations in ratings from $RatingPred_{ijrt}$ represent the extent to which examiners deviate from the average behavior of the examiner sample, conditional on bank observables.

For each observation, we define $Directional\_Discretion_{ijrt} = Rating_{ijrt} - RatingPred_{ijrt}$. A positive value for $Directional\_Discretion_{ijrt}$ implies the examiner gave a tougher rating than average

conditional on observables, and a negative value implies the examiner gave a more lenient rating than average conditional 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 implies the examiner exercised more discretion in either the tougher or more lenient directions.

We aggregate to the examiner level by taking the average of each measure across all exams conducted by each examiner to form $Examiner\_Directional\_Discretion_{ijt}$ and $Examiner\_Absolute\_Discretion_{irt}$. An examiner with high $Examiner\_Absolute\_Discretion_{irt}$ is a person who tends to deviate from the predicted average rating in either direction while an examiner with high $Examiner\_Directional\_Discretion_{ijt}$ is a person who issues tougher ratings on average, conditional on observables. We prefer to use $Examiner\_Absolute\_Discretion_{irt}$ as our primary measure of examiner discretion because it is more inclusive: 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 examiners on average, but may more heavily weight soft information or gut feelings in either direction. However, a modified version of $Examiner\_Directional\_Discretion_{ijt}$ will be useful for later analysis in which we estimate the causal effect of an exogenous change 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 (1) 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 (1) to the nearest integer value, 1 through 5, and measure directional and absolute discretion relative to the integer-rounded version of predicted rating.

Panels A and B of Table 2 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 0.16 points, or 0.19 after adjusting for integer rounding. As described in the Appendix, these estimates of the variance of examiner-level directional biases will be biased due to measurement error. After applying an Empirical Bayes shrinkage adjustment, we estimate that a

standard deviation in examiner-level directional discretion is 0.135 points. This variation is large considering that a standard deviation in overall ratings across all bank exam observations in our sample is 0.66, and the within-bank standard deviation in ratings is 0.48. In other words, holding observable bank characteristics and region-time constant, the variation in ratings due to random rotation across examiners is one-quarter as large as the total variation in ratings in our sample.

We also find a substantial degree of examiner absolute discretion. The average examiner deviates from the predicted rating in either direction by 0.24, or 0.13 after adjusting for integer rounding. This is approximately 20% larger than the difference in ratings induced by Federal versus State regulatory agencies which was identified and shown to be consequential in Agarwal et al. (2014). 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 0.22, or 0.13 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 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 very large compared to the overall transition probability of 6.7% and 9.0% 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 For Online Publication

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. For example, the standard deviation of examiner-level directional and absolute discretion is 0.16 and 0.12, respectively, in the full sample. A standard deviation of examiner-level directional and absolute discretion is also 0.16 and 0.12 in the federal sample, and 0.16 and 0.12 in the state sample. We further show in Panels E and F of For Online Publication

Appendix Table 1 that our findings relating to the causal impact of examiner discretion, as presented in the next section, remain similar if we control for the examiner's agency, and therefore only exploit variation across examiners in the same agency. Altogether, these results imply that exercise of discretion varies significantly across examiners, over and above differences due to agency affiliation (state versus federal) and is quantitatively large.

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 (which includes individual interpretation of the hard information as well as processing of soft information). While we cannot control for each examiner's true model of observable bank hard information, 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.

3.3 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.

Our empirical strategy exploits the fact that examiners are rotated across the banks within each region, subject to examiner workload limitations and qualifications (see Agarwal et al. 2014). Therefore, assignment of bank examination to a tougher or more lenient examiner should be 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*$_{j,-it}$, equal to the average directional discretion across all exams for examiner *j*, excluding the current exam. This approach is akin to a jack-knife measure of each examiner's directional discretion, where we exclude the current observation so that unobserved bank quality for the current examination does not affect our measure of *Directional_Discretion_LO*$_{j,-it}$

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

$$Rating_{ijt} = Directional\_Discretion\_LO_{j,-it} + X_{it} + \gamma_{rt} + \varepsilon_{ijrt} \quad (2)$$

$$BankOutcome_{i,t+1} = \widehat{Rating}_{ijt} + X_{it} + \gamma_{rt} + \eta_{ijrt} \quad (3)$$

We use *Directional_Discretion_LO*$_{j,-it}$ as the excluded instrument, and measure bank outcomes four quarters after the quarter of the current examination. Note that, while we use variation in *directional* discretion across examiners to estimate the causal effect of a CAMELS rating, this causal magnitude applies broadly, including to ratings that vary due to examiner *absolute* discretion. 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.

Before proceeding to the estimation of the instrumental variables strategy, 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 empirically find that examiner leave-out-mean ratings are uncorrelated with observable measures of bank health and 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. Finally, in Appendix Figure 1, 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.[15]

We then move on to estimating the instrumental variables procedure. Column 2 of Table 4 shows that the leave-out-mean directional discretion strongly predicts the composite rating assigned by the examiner for the current exam. Column 3 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 $Directional\_Discretion\_LO_{j,-it}$ strongly predicts the composite rating, conditional on bank observables. A one-unit increase in $Directional\_Discretion\_LO_{j,-it}$ is associated with a 0.19 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-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 5 presents our estimates of the causal effect of exogenously 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

[15] Recent work by Chan, Gentzkow, and Yu (2022) raised the concern that the monotonicity assumption may be violated if the costs of false negatives and false positives are highly asymmetric, and decision-makers vary in skill. In their setting, false negative diagnoses by radiologists are more costly than false positives, leading radiologists with lower skill to choose a lower diagnostic threshold for a positive diagnosis. While we cannot reject this critique, we note that it is less obvious that a strongly asymmetric cost function applies to our setting. First, the cost of a false positive by a radiologist is low partly because radiologists act as a first screen and positive diagnoses are sent for further review by other doctors. In contrast, CAMELS decisions by bank examiners are a final screen in the sense that they directly lead to regulatory restrictions. Second, whether overly harsh or lenient bank regulation is more costly is a hotly debated policy issue, with the two main political parties in the US often taking opposite sides.

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.

An exogenous one-point increase in ratings for the current exam causes a 1.79 unit increase in the tier1 capital ratio (equivalent to a change of 0.72 standard deviations or a 24% increase in a bank's capitalization relative to the sample median) and a 9.68 unit decline in loan growth (equivalent to a decrease of 0.88 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 0.60 point decrease in the bank's rating during the following exam. The magnitude of the effect is economically meaningful, equivalent to approximately 1.5 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 exogenous variation in ratings induced by CAMELS ratings indeed has 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 should not affect the performance of loans made prior to the ratings decision. We also find that ratings have a negative, albeit insignificant, effect on the bank's return on assets (ROA). This negative effect is consistent with the bank's conservative response to ratings resulting in reduced profitability.

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.

3.4 Bank anticipatory response

Discretion can affect bank behavior in two ways. First, as shown in the previous section, discretion introduces variation in ratings simply due to examiner assignment, which then has a causal impact on ex post bank capital ratios and lending behavior. Second, discretion introduces uncertainty, which can influence ex ante bank behavior. 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 due to discretion, we exploit the fact that the level of absolute discretion exercised by examiners varies across states and over time. We then test whether banks in states-years in which examiners have recently exercised a high degree of discretion engage in precautionary measures. The idea behind this test is as follows: while bank management cannot directly observe the CAMELS ratings of other banks in their state, they can observe the publicly disclosed actions of other banks such as their capitalization and lending activity. Thus, if bank management observes substantial shifts in these activities for other banks from year to year, bank management can partially infer that regulation within their state is uncertain and volatile.

We measure state-level discretion as the average exam-level absolute discretion for all exams in the state in the past five years, excluding the current bank. To isolate variation due to uncertainty, we control directly for the bank's own most-recent rating, bank observables, state fixed effects, and year-quarter fixed effects. These controls account for the possibility that examiners tend to exert more discretion in states where banks are riskier and thus banks are better capitalized due to past regulatory oversight. 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.*

In Table 6 Panel A, we show that higher average state absolute discretion is associated with significantly higher tier 1 capital ratios and lower loan growth. We also explore whether banks with lagged ratings greater than one are more sensitive to uncertainty, because moving from a rating of 2 to 3 is associated with greater negative consequences than moving from a rating of 1 to 2. In column 4, we find a slightly larger response for this sample of banks in terms of reduced 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 0.74 units and increasing capital ratios by 0.37 units, a 11% decline and a 3% 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 6 suggests that discretion can impact both the ex post and ex ante behavior of banks.

3.5 Why do examiners disagree? Weights on components

So far, we have shown that examiners vary widely in directional discretion and that some examiners exhibit significant absolute discretion. We have also shown that examiner discretion can influence ex post and ex ante bank outcomes. However, bank examiners are experienced professionals who are trained to assess bank safety and soundness. Why would two examiners, upon observing equivalent bank information, decide on different ratings? In this section, we focus on weights as a cause of disagreement. In supplementary results in the Appendix, we show that variation in examiner experience and recent rating experience can also contribute to disagreement.

Before proceeding, we first explicitly define what it means for multiple examiners to disagree after reviewing equivalent bank information. In our sample, only one lead examiner is assigned to review a bank at each point in time, so we never observe cases in which multiple lead examiners review identical bank information and disagree on ratings. Rather, we estimate disagreement by exploiting regular examiner rotation. Examiners with a region are regularly rotated across the banks in a region. Within a region over time, one examiner should not be systematically assigned to worse banks. Therefore, examiners within a region observe banks of equal expected quality. If one examiner predictably issues higher ratings or ratings with greater variance, then that examiner disagrees with another examiner after observing equivalent information.

Examiners in our setting separately rate each component of the CAMELS rating. In creating the composite rating, examiners are expected to use the component ratings but are not constrained to using a pre-specified weighting function of the components. We can model the composite rating as a weighted average of the component ratings, where examiners have discretion in choosing the weights.

$$R = w_C C + w_A A + w_M M + w_E E + w_L L + w_S S \quad (4)$$

Weights can lead to disagreement through two main channels. First, suppose examiners agree on all weights. Disagreement in the final decision will increase if examiners assign greater weight to components for which there is more disagreement across examiners. Second, suppose examiners agree on all components. Disagreement in the final decision will increase with disagreement in

weights across examiners.

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. Thus, the coefficients, under plausible assumptions, can be thought of as weights on the component ratings.[16] Notably, these estimates of weights do not require any assumptions about the randomness of examiner assignment to banks or measurement of bank observables.

We find that examiners on average place a substantial 49% of the total weight on the Management component rating when forming the composite CAMELS rating. Examiners also place significant weight on the Assets, Capital, and Earnings components, at 15%, 12%, and 11% respectively. Meanwhile, examiners place little weight on the Sensitivity and Liquidity components. These patterns are illustrated in Figure 2, which shows the graphical counterpart to the regression in Column 2 of Table 7.

We also find that examiners place substantial weight on assessments of management quality across a variety of subsamples. Columns 3 and 4 of Table 7 Panel A separate the sample into healthy banks with ratings of 1 or 2 and unhealthy banks with ratings of 3, 4, or 5, respectively. For both healthy and unhealthy banks, the management rating accounts for approximately half of the composite rating. In Panel B, 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 56% in the post-crisis period.

The high weight placed on the management component is especially remarkable in that assessment of management quality is necessarily subjective. In 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."

Panel C of Table 7 shows the extent to which each component rating can be explained by observable bank characteristics and region-time fixed effects. Each reported number represents the

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

R-squared from a regression of the component rating on bank observables and location-quarter fixed effects. A low R-squared implies that a component category is associated with greater examiner subjectivity, conditional on the observable hard information. We find that Management, Liquidity, and Sensitivity have low R^2s while Assets and Earnings have relatively high R^2s. By combining the results in Panels A and B of Table 7, we arrive at the first reason why examiners may disagree with respect to composite ratings: they assign relatively greater weight to components for which there is more disagreement, in particular the Management component rating. The high weight placed on perceptions of management quality is suggestive of a particular form of discretion noted in the psychology literature, in which decision-makers place very high weight on inferences drawn from face-to-face interactions. For example, Levine et al. (1999) show that people place too much weight to information gleaned from facial expressions, language, and face-to-face interactions.

Next, we explore how the weighting function over component ratings varies with directional discretion and absolute discretion. The results are reported in Table 8, 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 positive directional discretion corresponds to significantly greater weight placed on Management (an increase of 37% relative to quartile 1) and Sensitivity (an increase of 81%), and significantly lower weight placed on Earnings (a decrease of 22%). Greater absolute discretion corresponds to significantly greater weight placed on Assets (an increase of 26%) and Management (an increase of 18%), and significantly lower weight placed on Capital (a decrease of 29%), Liquidity (a decrease of 31%), and Sensitivity (a decrease of 3%). These patterns point to a second important reason for disagreement in reaching the final composite rating decision. Examiners disagree in how much weight to assign to components, and these weights vary predictably with directional and absolute discretion.

3.6 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 how discretion affects the quality of examiner 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. We partly address these shortcomings by measuring the ability of the CAMELS rating to predict future CAMELS ratings and future changes in bank observable characteristics that are known to be strong predictors of the CAMELS rating. However, we acknowledge that this is imperfect estimate of what the CAMELS rating is supposed to predict, and caution that our results should be viewed only as suggestive.

Second, we are interested in the ability of a CAMELS rating to predict future bank outcomes. However, as we have established in Section 3.3, CAMELS ratings causally affect future bank outcomes in addition to predicting them. For example, a high CAMELS rating may predict that a bank's loan portfolio will underperform but may also cause an improvement in loan performance 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 5.

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 9 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 who exercise the most discretion do not outperform examiners who exercise moderate or low levels of discretion. In other words, examiners who exercise high discretion introduce additional uncertainty and noise without adding predictive power.

Finally, we compare the ability of the composite rating chosen by examiners and two 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 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, 0.115, and 0.099, respectively. 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 (see e.g., Dawes (1979) and Dawes et al. (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. As discussed in the Introduction, we purposely do not "race" examiner ratings against an advanced and complex algorithmic model. Given the extremely rapid rate of progress in AI and machine learning methods, such a test is unlikely to be informative. 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 second counterfactual removes examiner discretion over how to weight each component rating. In particular, this reweighted composite rating reduces the weight examiners place on subjective assessments of management quality from approximately 50% to 29%. Panel A of Table 10 shows the correlations between the real and two counterfactual ratings. While positively correlated, there exists substantial variation between the real and counterfactual ratings.

Panels B and C of Table 10 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 (≥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 three 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 relative predictive power of the actual and reweighted composite ratings is ambiguous. The reweighted composite rating is a stronger predictor of near-term changes in the performance of the bank's existing loan portfolio, whereas the actual rating is a stronger predictor of the bank's CAMELS rating in the following year. An important consideration when evaluating these results is that the bank's future rating is chosen by a (different) human examiner who may anchor her decision to the bank's previous rating, thereby giving the actual composite rating an advantage in predicting the bank's future rating. To the extent that the goal of a bank examiner is to assess the health of the bank's loan portfolio, our results suggest that the reweighted composite rating outperforms the actual rating chosen by examiners. The performance gain of the reweighted composite rating over the actual rating amounts to only a small increase in AUC. Nevertheless, like our earlier results in Table 9, these results suggest that limiting examiner discretion can lead to predictions that are at least as powerful, while simultaneously reducing noise.

A drawback of using AUCs to assess predictive power is that outcomes must be measured as binary values. To address this limitation, we assess the ability of the actual and counterfactual ratings to predict continuous measures of the bank's change in loan portfolio performance, as well as the bank's full range of future ratings (1-5). Panel C shows the mean squared error (MSE) from flexible regressions of future bank outcomes on indicators for each level, 1 through 5, of the actual or counterfactual ratings. A separate linear regression is estimated for each of the actual and counterfactual ratings. Dependent variables are the bank's next rating, the change in the non-performing loan ratio in the quarter after the exam relative to the quarter before the exam, and the change in the loan delinquency ratio in the quarter after the exam relative to the quarter before the exam. Lower MSE values correspond to greater predictive power.

The MSE values in Panel C imply a similar ranking of predictive power as the AUC values in Panel B. We again find that the actual rating is a better predictor of changes in bank quality than the predicted rating based on observable bank characteristics. The reweighted composite rating is a stronger predictor of near-term changes in the performance of the bank's existing loan portfolio, whereas the actual rating is a stronger predictor of the bank's CAMELS rating in the following year.

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 better forecasts of changes in loan performance. Likewise, our earlier results in Table 9 showed that examiners who choose to exercise less discretion generate ratings that are just as predictive as those who exercise more discretion.

4. 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 strict regulation. However, human discretion injects a large degree of noise and uncertainty into the system. Discretion has a causal impact on future bank capitalization and supply of credit, leading to increased uncertainty and volatility in bank outcomes, as well as a conservative anticipatory response on the part of banks. We explore the causes of disagreement across examiners and find that heterogeneity in the weights attached to specific issues, particularly management quality, leads to disagreement in final rating decisions. Finally, we show that human discretion can be valuable because it allows for

the effective processing of soft information. However, placing moderate limits on the degree of human discretion can translate into more informative predictions.

Our goal in this paper was to evaluate the consequences, determinants, and trade-offs associated with human discretion in bank regulation. It is worth noting that the optimal amount of noise is unlikely to be zero. 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., see 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., see 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.

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), 467-75.

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

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

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer. 2018. "Extrapolation and bubbles." *Journal of Financial Economics*, 129(2), 203-227.

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.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2015. "Salience Theory of Judicial Decisions." *Journal of Legal Studies*, 44(S1): S7–S33.

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

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

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.

Chernenko, Sergey, Samuel Hanson, and Adi Sunderam. 2016. "Who Neglects Risk? Investor Experience and the Credit Boom." *Journal of Financial Economics*, 122(2), 248-269.

Chilton, Adam S., and Marin K. Levy. 2015. "Challenging the randomness of panel assignment in the Federal Courts of Appeals." *101 Cornell Law Review*, 1.

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.

Epstein, Lee, Andrew D. Martin, Kevin M. Quinn, and Jeffrey A. Segal. 2007. "Ideological drift among Supreme Court justices: Who, when, and how important." *Nw. UL Rev.* 101: 1483.

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.

Gennaioli, Nicola, Yueran Ma and Andrei Shleifer. 2015. "Expectations and Investment." NBER Macroeconomics Annual.

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.

Goetzmann, William N., and Nadav Peles. 1997. "Cognitive Dissonance and Mutual Fund Investors." *Journal of Financial Research*, 20(2), 145-158.

Greenwood, Robin, and Andrei Shleifer. 2014. "Expectations of Returns and Expected Returns." *Review of Financial Studies*, 27(3), 714-746.

Griffin, Dale, and Amos Tversky. 1992. "The Weighing of Evidence and the Determinants of Confidence." *Cognitive Psychology*, 24(3), 411-435.

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. *Noise: A Flaw in Human Judgment*. Little, Brown Spark, 2021.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. "Algorithmic Fairness." *AEA Papers and Proceedings*, 108: 22-27.

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.

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

Ludwig, Jens, and Sendhil Mullainathan. 2023. "Algorithmic behavioral science: Machine learning as a tool for scientific discovery." *Chicago Booth Research Paper*.

Malmendier, Ulrike, and Stefan Nagel. 2016. "Learning from Inflation Experiences." *Quarterly Journal of Economics*, 131(1), 53-87.

Martin, Andrew D., and Kevin M. Quinn. 2007. "Assessing Preference Change on the US Supreme Court." *The Journal of Law, Economics, and Organization*, 23(2), 365-385.

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.

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

Stajkovic, Alex. 2006. "Development of a Core Confidence-Higher Order Construct." *Journal of Applied Psychology*, 91(6), 1208-1224.

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.

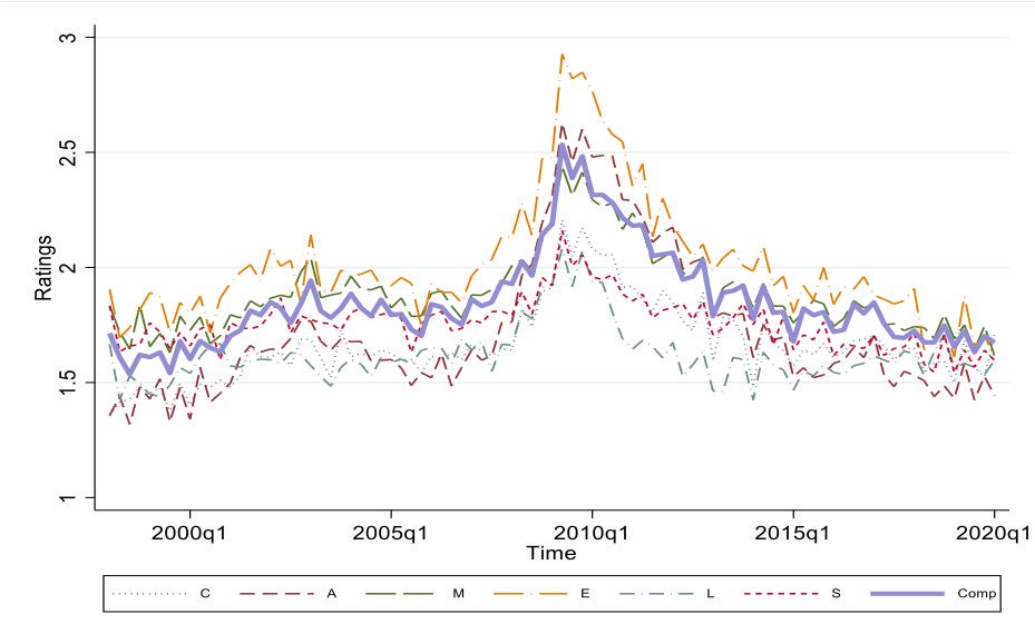


Figure 2: Weights on component ratings

This figure shows the composite rating as a weighted average of the component ratings, using the estimates from Table 7 Panel A Column 2.

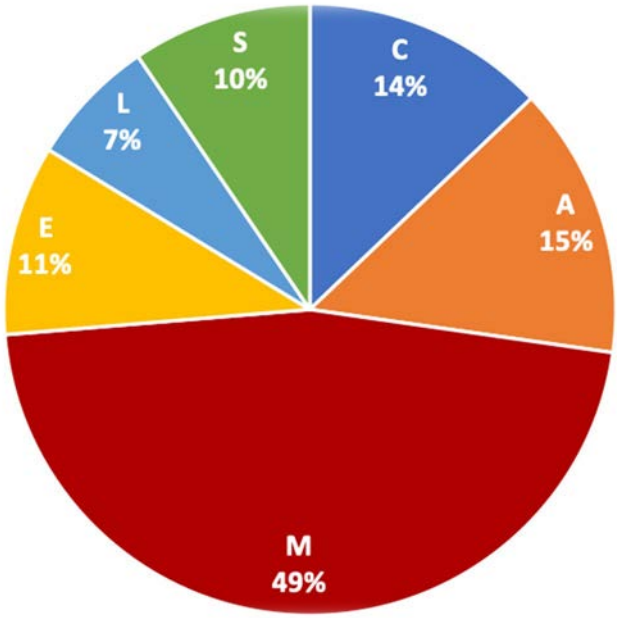
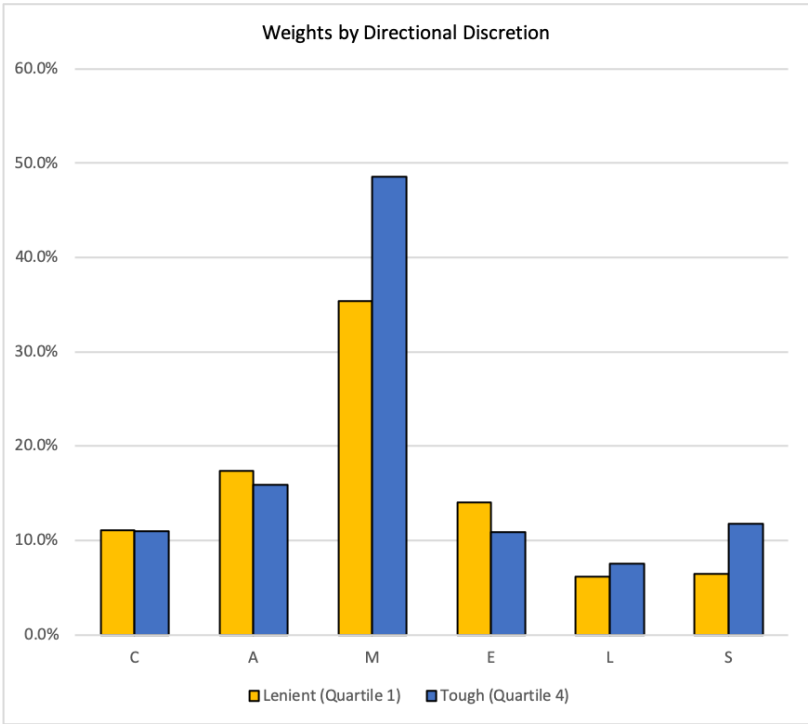


Figure 3: Weights on CAMELS component ratings and discretion

This figure shows the composite rating as a weighted average of the component ratings, separately for observations in the top and bottom quartiles of direction discretion (top panel) and top and bottom quartiles of absolute discretion (bottom panel). Detailed corresponding regression estimates are reported in Table 8.



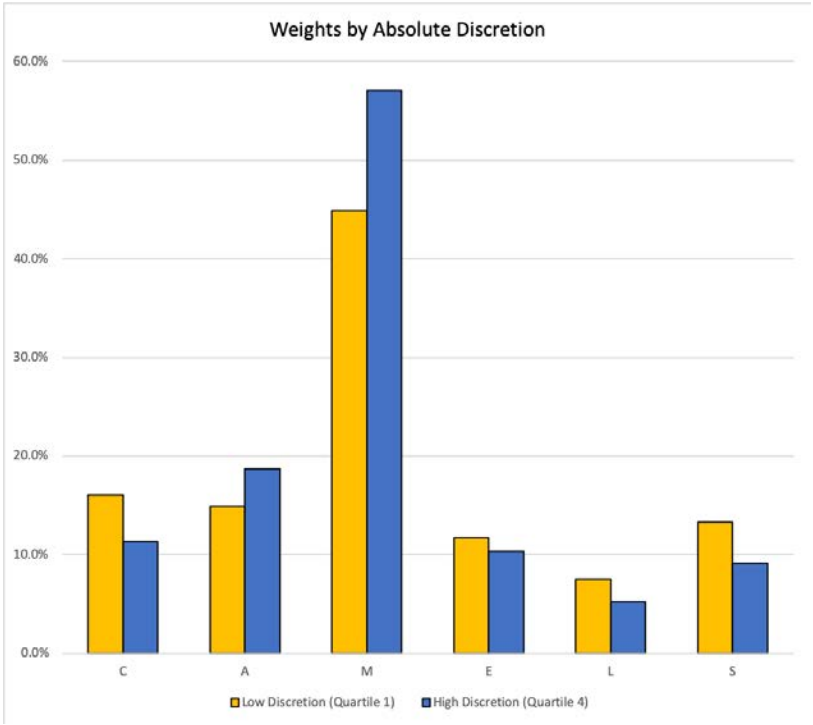
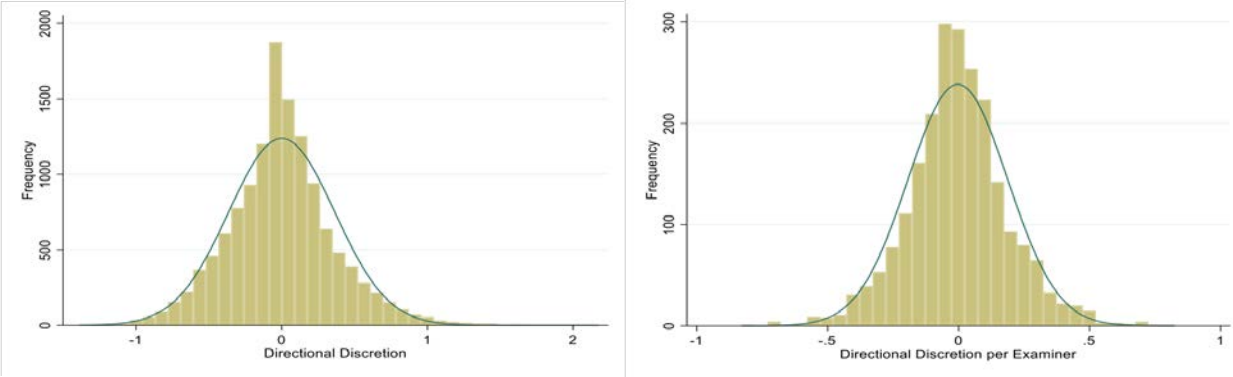


Figure 4: Distribution of directional discretion and absolute discretion

These figures show the distribution of estimates of directional discretion and absolute discretion at the exam level (left side) and examiner level (right side).

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



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

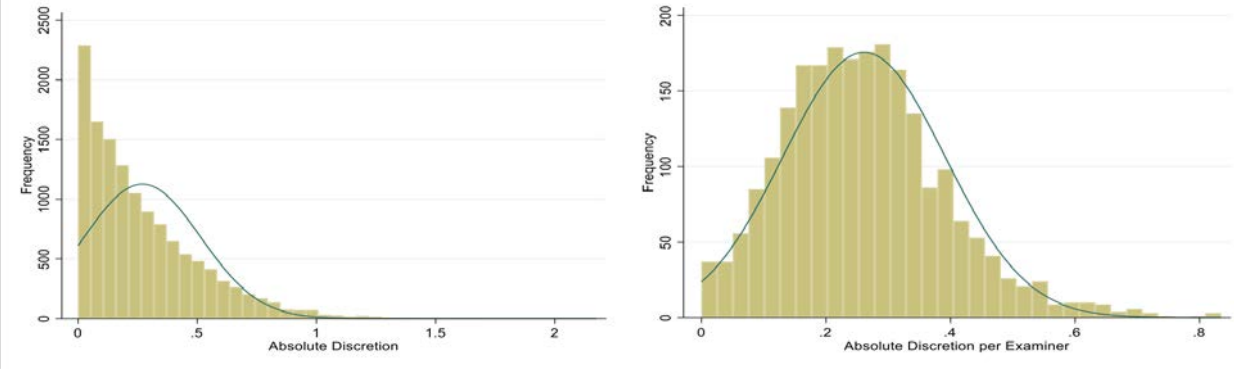


Table 1: Summary statistics

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 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	2272	5.7	5.2	2	2	4	7	67
Examiner-exam-level								
Years experience	13020	5.3	4.7	1	2	4	7	24
Number of exams so far	13020	5.7	6.1	1	2	4	7	67

Panel B: CAMELS rating components

Rating	N	Mean	S.D.	S.D. (Within bank)	Min	p25	p50	p75	Max	Frac ≥ 3
Capital	13020	2.00	0.64	0.45	1	1	2	2	5	0.06
Assets	13020	2.00	0.76	0.60	1	1	2	2	5	0.12
Management	13020	1.89	0.68	0.49	1	1	2	2	5	0.13
Earnings	13020	2.01	0.90	0.62	1	1	2	2	5	0.23
Liquidity	13020	1.62	0.61	0.43	1	1	2	2	5	0.05
Sensitivity	13020	1.76	0.57	0.41	1	1	2	2	5	0.06
Composite	13020	1.85	0.66	0.48	1	1	2	2	5	0.11

Panel C: Transitions probability matrix

	Next Rating					
Rating	1	2	3	4	5	Total
1	79.0	19.9	1.0	0.1	0.0	100
2	9.0	82.9	6.7	1.1	0.2	100
3	0.0	39.4	50.6	8.4	1.6	100
4	0.0	4.0	19.2	56.3	20.5	100
5	0.0	0.0	4.8	14.3	81.0	100
Total	27.7	60.1	9.2	2.3	0.7	100

Panel D: Bank observables

	N	Mean	S.D.	S.D. (within bank)	Min	p25	p50	p75	Max
Tier1 ratio	13020	14.73	5.43	2.49	8.84	10.97	13.16	16.65	33.07
Loan growth	13020	9.10	13.42	10.99	-12.58	0.85	6.75	14.07	54.25
Leverage	13020	9.78	2.48	1.31	6.36	8.05	9.24	10.83	17.45
ROA	13020	0.24	0.19	0.14	-0.38	0.16	0.25	0.34	0.61
NPL ratio	13020	1.14	1.27	1.03	0.00	0.25	0.71	1.53	5.49
Delinq ratio	13020	2.23	1.97	1.48	0.01	0.80	1.68	3.02	8.44
Efficiency	13020	63.79	19.03	16.08	25.81	48.86	63.43	80.47	95.08
Assets(millions)	13020	1381.0	10940.6	3696.8	5.9	82.6	183.9	486	621962

Table 2: 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 discretion	Non-integer	13020	0.00	0.32	-0.40	-0.19	0.00	0.17	0.40
Directional discretion	Integer	13020	0.00	0.36	0.00	0.00	0.00	0.00	0.00
Absolute discretion	Non-integer	13020	0.24	0.22	0.01	0.07	0.18	0.35	0.54
Absolute discretion	Integer	13020	0.13	0.33	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 discretion	Non-integer	2272	0.00	0.16	-0.20	-0.10	0.00	0.09	0.19
Directional discretion	Integer	2272	0.00	0.19	-0.20	0.00	0.00	0.00	0.20
Absolute discretion	Non-integer	2272	0.23	0.12	0.08	0.15	0.22	0.30	0.38
Absolute discretion	Integer	2272	0.12	0.17	0.00	0.00	0.00	0.20	0.36

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. If examiners are rotated across banks within regions on a regular schedule, bank characteristics at the time of the exam and future loan performance should be uncorrelated with measures of examiner discretion, holding the region and time period fixed. 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 directional discretion

	(1) Tier1 ratio	(2) Loan growth	(3) Leverage	(4) ROA	(5) NPL ratio	(6) Delinq ratio	(7) Efficiency	(8) Chg NPL Ratio	(9) Chg Delinq ratio
Examiner dir disc (LO)	-0.592	0.650	-0.121	-0.004	0.006	-0.079	-0.269	-0.018	-0.074
	(0.383)	(0.791)	(0.169)	(0.010)	(0.081)	(0.130)	(0.607)	(0.045)	(0.074)
Observations	13020	13020	13020	13020	13020	13020	13020	13020	13020
R-squared	0.286	0.317	0.286	0.335	0.393	0.369	0.835	0.274	0.269
Location-qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

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

	(1) Tier1 ratio	(2) Loan growth	(3) Leverage	(4) ROA	(5) NPL ratio	(6) Delinq ratio	(7) Efficiency	(8) Chg NPL Ratio	(9) Chg Delinq ratio
Examiner abs disc (LO)	0.373	-1.939	-0.006	-0.018	0.170	0.236	1.356	0.033	0.132
	(0.572)	(1.238)	(0.253)	(0.016)	(0.125)	(0.199)	(0.981)	(0.074)	(0.117)
Observations	13020	13020	13020	13020	13020	13020	13020	13020	13020
R-squared	0.286	0.317	0.286	0.335	0.393	0.369	0.835	0.274	0.269
Location-qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel C: Autocorrelation in examiner assignment to banks

	(1) Examiner dir disc (LO)	(2) 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: Predictors of ratings, first stage regression

Column 1 represents a regression of the composite rating on bank characteristics at the time of the exam. Column 2 represents a regression of the composite rating for the current exam on the examiner's leave-out-mean directional discretion. Column 3 represents the same regression as column 1, with the addition of the examiner's leave-out-mean directional discretion as an explanatory variable. Column 3 also represents the first stage regression in the instrumental variables estimation presented in Table 5, with the examiner leave-out-mean directional discretion as the excluded instrument. 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)	(3)
Examiner directional discretion(LO)		0.233***	0.186***
		(0.046)	(0.037)
Tier1 ratio	-0.018***		-0.018***
	(0.003)		(0.003)
Loan growth	-0.002***		-0.002***
	(0.000)		(0.000)
Leverage	-0.854***		-0.851***
	(0.033)		(0.046)
ROA	0.050***		0.050***
	(0.008)		(0.009)
NPL ratio	0.057***		0.057***
	(0.005)		(0.006)
Delinq ratio	-0.003***		-0.003***
	(0.001)		(0.001)
Efficiency	-0.017***		-0.016***
	(0.005)		(0.006)
Observations	13,020	13,020	13,020
Bank FE	Yes	No	Yes
Location-quarter FE	Yes	Yes	Yes

Table 5: 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 format of the first stage regression is as reported in Column 3 of Table 4). 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. 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	Problem bank (next rating)	Tier1 ratio - 1yr	Loan growth - 1yr
Pred composite rating	-0.671***	-0.167***	1.862***	-9.755**
	(0.302)	(0.082)	(0.702)	(3.883)
Observations	10,489	10,489	11,068	11,122
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	1.900**	-9.217	1.917*	-19.241**
	(0.930)	(6.497)	(1.028)	(8.866)
Observations	10,085	10,119	9,201	9,243
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)	(3)	(4)
Future bank outcome	ROA	NPL ratio	Delinq ratio	Efficiency
Pred composite rating	0.035	0.047	0.216	0.768
	(0.050)	(0.367)	(0.515)	(1.750)
Observations	11,124	11,123	11,123	11,128
Bank controls	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes	Yes

Table 6: Pre-emptive bank response

This table regresses bank outcomes on the average of the absolute examiner discretion for all exams in the state over the past five years, calculated excluding the current bank. 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		Loan growth	
	All obs	Lagged rating > 1	All obs	Lagged rating > 1
Avg state abs disc, past 5 yrs	2.62***	2.42**	-5.26**	-8.20***
	(0.96)	(1.20)	(2.11)	(2.97)
Observations	12,253	6,188	11,528	6,171
R-squared	0.57	0.55	0.10	0.11
Quarter FE	Yes	Yes	Yes	Yes
Lagged Bank Controls	Yes	Yes	Yes	Yes

Panel B: Auxiliary outcomes

	(1)	(2)	(3)	(4)
	ROA	NPL ratio	Delinq ratio	Efficiency
Avg state abs disc, past 5 yrs	0.03	0.03	0.41	0.52
	(0.03)	(0.30)	(0.47)	(1.49)
Observations	12,312	12,304	12,304	12,312
R-squared	0.22	0.20	0.25	0.86
Quarter FE	Yes	Yes	Yes	Yes
Lagged bank controls	Yes	Yes	Yes	Yes

Table 7: Weights on CAMELS rating components

Panel A shows how the composite rating is related to the component ratings. Column 1 presents an unconstrained OLS regression and Column 2 estimates the same regression subject to the constraint that the coefficients sum to one. Column 3 and 4 restricts the sample to banks with satisfactory and unsatisfactory ratings, respectively. 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 C shows the R-squared measure of extent to which each component rating can be explained by observable bank characteristics and region-time fixed effects. Bank observables consist of variables reported in Table 1 Panel D.

Panel A: Weights

Composite rating	(1) All banks	(2) All banks	(3) Lagged rating < 3	(4) Lagged rating≥3
C-rat	0.137***	0.123***	0.125***	0.103***
	(0.006)	(0.005)	(0.005)	(0.017)
A-rat	0.153***	0.149***	0.137***	0.205***
	(0.005)	(0.004)	(0.004)	(0.013)
M-rat	0.494***	0.491***	0.486***	0.499***
	(0.008)	(0.005)	(0.006)	(0.015)
E-rat	0.107***	0.113***	0.106***	0.103***
	(0.004)	(0.003)	(0.004)	(0.011)
L-rat	0.072***	0.053***	0.063***	0.023
	(0.005)	(0.004)	(0.005)	(0.015)
S-rat	0.101***	0.070***	0.082***	0.067***
	(0.006)	(0.005)	(0.005)	(0.015)
Constrained regression	No	Yes	Yes	Yes
Observations	13,020	13,020	11,747	1,273
R-squared	0.851			

Panel B: Weights over time

Composite rating	(1) 1998Q1-2007Q4	(2) 2008Q1-2012Q4	(3) 2013Q1-2020Q1
C-rat	0.110***	0.131***	0.132***
	(0.007)	(0.011)	(0.009)
A-rat	0.139***	0.197***	0.111***
	(0.006)	(0.009)	(0.008)
M-rat	0.473***	0.470***	0.557***
	(0.007)	(0.012)	(0.010)
E-rat	0.131***	0.101***	0.074***
	(0.005)	(0.007)	(0.007)
L-rat	0.073***	0.023***	0.044***
	(0.006)	(0.010)	(0.008)
S-rat	0.075***	0.078***	0.081***
	(0.007)	(0.010)	(0.009)
Constrained regression	Yes	Yes	Yes
Observations	6,559	2,741	3,720

Panel C: Explanatory power of bank observables

R-squared	(1) C-rat	(2) A-rat	(3) M-rat	(4) E-rat	(5) L-rat	(6) S-rat
Bank observables	0.392	0.429	0.285	0.529	0.265	0.160
Bank observables + location-qtr FE	0.581	0.612	0.490	0.659	0.474	0.403

Table 8: Weights by CAMELS rating components and 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	0.111***	C-rat	0.160***
	(0.011)		(0.011)
C-rat x high direc disc	-0.000	C-rat x high abs disc	-0.047***
	(0.013)		(0.015)
A-rat	0.174***	A-rat	0.149***
	(0.009)		(0.010)
A-rat x high direc disc	-0.015	A-rat x high abs disc	0.038***
	(0.013)		(0.014)
M-rat	0.354***	M-rat	0.449***
	(0.014)		(0.017)
M-rat x high direc disc	0.132***	M-rat x high abs disc	0.082***
	(0.022)		(0.022)
E-rat	0.140***	E-rat	0.117***
	(0.007)		(0.008)
E-rat x high direc disc	-0.031***	E-rat x high abs disc	-0.014
	(0.010)		(0.011)
L-rat	0.062***	L-rat	0.075***
	(0.009)		(0.009)
L-rat x high direc disc	0.013	L-rat x high abs disc	-0.023*
	(0.013)		(0.013)
S-rat	0.065***	S-rat	0.133***
	(0.010)		(0.011)
S-rat x high direc disc	0.053***	S-rat x high abs disc	-0.042***
	(0.013)		(0.014)
Observations	6,510		6,510
R-squared	0.898		0.871

Table 9: Does discretion lead to better predictions of future outcomes?

This table explores the extent to which the subjective component in ratings is able to predict future bank ratings and near-term loan performance, and how the predictive power varies with examiner-level discretion. We measure the subjective 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. The dependent variable in Column 1 is the rating in next exam, usually 4 quarters away. In the other columns, 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) Next rating	(2) NPL ratio	(3) Delinq Ratio
Exam-level directional discretion	0.298***	0.163***	0.257***
	(12.185)	(6.422)	(6.518)
Observations	11,420	12,856	12,856
R-squared	0.401	0.278	0.274
Location-quarter FE	Yes	Yes	Yes

Panel B:

	(1) Next rating	(2) NPL ratio	(3) Delinq Ratio
Resid rating x low disc	0.281***	0.147***	0.211***
	(5.767)	(3.232)	(2.929)
Resid rating x med disc	0.324***	0.174***	0.298***
	(8.360)	(4.199)	(4.220)
Resid rating x high disc	0.284***	0.163***	0.249***
	(6.453)	(3.368)	(3.535)
Observations	11,420	12,856	12,856
R-squared	0.401	0.278	0.274
Location-quarter FE	Yes	Yes	Yes
Low disc = high disc	0.959	0.821	0.712

Table 10: Predictive power of actual versus counterfactual ratings

This table compares the ability of the composite rating chosen by examiners and two counterfactual ratings to predict future bank ratings and near-term changes in loan performance. 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 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). Both counterfactual ratings are rounded to the nearest integer, 1-5. Panel A shows the correlation between these three ratings. Panel B shows the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Predicted outcomes are indicators for whether the next rating is unsatisfactory (≥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 quarter after the exam relative to the quarter before the exam is in the top 10% of the sample. Higher AUC values correspond to greater predictive power. Panel C shows the mean squared error (MSE) from flexible regressions of future bank outcomes on indicators for each level, 1 through 5, of the actual or counterfactual ratings. A separate linear regression is estimated for each of the actual and counterfactual ratings. Dependent variables are the next rating, the change in the non-performing loan ratio in the quarter after the exam relative to the quarter before the exam , and the change in the loan delinquency ratio in the quarter after the exam relative to the quarter before the exam. Lower MSE values correspond to greater predictive power.

Panel A: Correlations between actual and counterfactual ratings

Correlations	Composite rating	Reweighted comp rating	Predicted rating
Composite rating	1.00		
Predicted rating	0.85	1.00	
Reweighted comp rating	0.92	0.80	1.00

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

	(1) Next rating	(2) NPL ratio	(3) Delinq ratio
Composite rating	0.8204	0.6219	0.5858
Predicted rating	0.7942	0.6207	0.5858
Reweighted comp rating	0.8043	0.6270	0.5915

Panel C: Mean squared error of actual and counterfactual ratings

	(1) Next rating	(2) NPL ratio	(3) Delinq ratio
Composite rating	0.4959	0.7646	1.2327
Predicted rating	0.5123	0.7638	1.2314
Reweighted comp rating	0.5185	0.7637	1.2311

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 2, 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 E is similar to Table 4, with the addition of an indicator variable for whether the examiner is associated with a federal or state agency as an additional control variable. Panel F is similar to Table 5 Panel A, with the addition of an indicator variable for whether the examiner is associated with a federal or state agency as a non-excluded control variable in the IV estimation. Standard errors are in parentheses and clustered by examiner. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Exam-level, federal sample

Discretion type	Pred rating	Obs	Mean	S.D.	p10	p25	p50	p75	p90
Directional discretion	Non-integer	7139	0.04	0.32	-0.35	-0.15	0.00	0.20	0.45
Directional discretion	Integer	7139	0.03	0.36	0.00	0.00	0.00	0.00	0.00
Absolute discretion	Non-integer	7139	0.24	0.22	0.01	0.06	0.18	0.35	0.54
Absolute discretion	Integer	7139	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	5881	-0.04	0.32	-0.47	-0.23	-0.01	0.13	0.34
Directional discretion	Integer	5881	-0.04	0.36	0.00	0.00	0.00	0.00	0.00
Absolute discretion	Non-integer	5881	0.24	0.22	0.01	0.07	0.18	0.35	0.55
Absolute discretion	Integer	5881	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	1204	0.03	0.16	-0.15	-0.07	0.02	0.12	0.23
Directional discretion	Integer	1204	0.03	0.18	-0.14	0.00	0.00	0.06	0.25
Absolute discretion	Non-integer	1204	0.23	0.12	0.08	0.15	0.22	0.30	0.38
Absolute discretion	Integer	1204	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	1068	-0.04	0.17	-0.26	-0.14	-0.03	0.06	0.16
Directional discretion	Integer	1068	-0.04	0.18	-0.27	-0.06	0.00	0.00	0.11
Absolute discretion	Non-integer	1068	0.23	0.12	0.06	0.14	0.22	0.31	0.39
Absolute discretion	Integer	1068	0.12	0.17	0.00	0.00	0.00	0.20	0.40

Panel E: Predictors of ratings, first stage regression, controlling for federal versus state

Composite rating	(1)	(2)	(3)
Examiner bias(LO)		0.139***	0.104***
		(0.048)	(0.038)
Tier1 ratio	-0.018***		-0.018***
	(0.003)		(0.003)
Loan growth	-0.002***		-0.002***
	(0.000)		(0.000)
Leverage	-0.842***		-0.841***
	(0.033)		(0.046)
ROA	0.049***		0.049***
	(0.008)		(0.009)
NPL ratio	0.057***		0.057***
	(0.005)		(0.006)
Delinq Ratio	-0.003***		-0.003***
	(0.001)		(0.001)
Efficiency	-0.017***		-0.017***
	(0.005)		(0.006)
Fed	0.108***		0.107***
	(0.008)		(0.009)
Observations	13,020	13,020	13,020
Bank FE	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes

Panel F: Causal impact of discretion on ratings, controlling for federal versus state

	(1)	(2)	(3)	(4)
Future bank outcome	Next rating	Problem bank (next rating)	Tier1 ratio	Loan growth
Pred composite rating	-0.407	-0.316*	2.364*	-14.718*
	(0.584)	(0.190)	(1.424)	(8.150)
Observations	10,489	10,489	11,068	11,122
Bank controls	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes	Yes

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

This table replicates Table 2, 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 discrection	Non-integer	13020	0.00	0.32	-0.40	-0.18	0.00	0.17	0.39
Directional discrection	Integer	13020	0	0.34	0	0	0	0	0
Absolute discretion	Non-integer	13020	0	0.22	0.001	0.065	0.17	0.34	0.54
Absolute discretion	Integer	13020	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 discrection	Non-integer	2272	-0	0.16	-0.21	-0.095	0	0.09	0.19
Directional discrection	Integer	2272	0	0.18	-0.2	0	0	0	0.2
Absolute discretion	Non-integer	2272	0.22	0.12	0.723	0.136	0.21	0.29	0.38
Absolute discretion	Integer	2272	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 5 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.

Panel A: Full Sample

	(1)	(2)	(3)	(4)
Future bank outcome	Next rating	Problem bank (next rating)	Tier1 ratio - 1yr	Loan growth - 1 yr
Pred composite rating	-0.450*	-0.095*	1.835***	-9.254**
	(0.232)	(0.057)	(0.657)	(3.738)
Observations	10,489	10,489	11,068	11,122
R-squared	-0.169	0.021	0.424	0.054
Bank controls	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes	Yes

Panel B: Long run impact

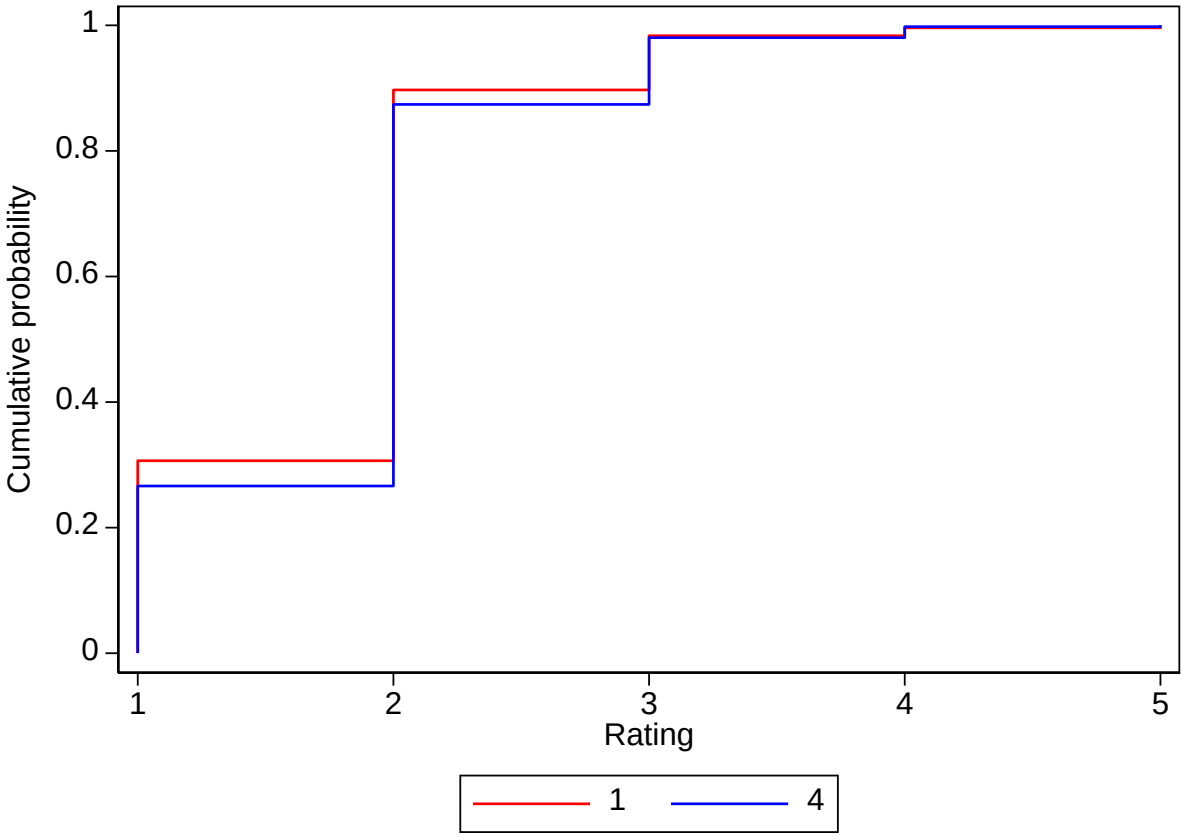
	(1)	(2)	(3)	(4)
Future bank outcome	Tier1 ratio - 2yr	Loan growth - 2 yr	Tier1 ratio - 3yr	Loan growth - 3 yr
Pred composite rating	1.806**	-9.033	2.046**	-19.829**
	(0.884)	(6.287)	(0.986)	(8.941)
Observations	10,085	10,119	9,201	9,243
R-squared	0.220	0.106	0.052	0.035
Bank controls	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes	Yes

Panel C: Auxiliary outomces

	(1)	(2)	(3)	(4)
Future bank outcome	ROA	NPL Ratio	Delinq ratio	Efficiency
Pred composite rating	0.020	0.136	0.315	2.059
	(0.047)	(0.346)	(0.486)	(1.755)
Observations	11,124	11,123	11,123	11,128
R-squared	-0.169	0.021	0.424	0.054
Bank controls	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes	Yes

Appendix Figure 1: 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: Framework for discretion and costly noise

General Framework

We present a framework to illustrate why human discretion may lead to costly noise. Suppose each case has an optimal outcome Z^*. Let $\mathrm{Z} = Z^* + b + e$ be the case outcome determined by the human decision-maker. b is a constant representing the extent to which the population of decision-makers is too harsh or too lenient in expectation relative to the optimal outcome. If we view decision-makers as attempting to estimate the optimal outcome, then b represents the bias of their estimate in a statistical sense. Let e be the additional error relative to the optimal outcome introduced by individual decision-makers, with $E[e] = 0$ and $Var[e] = \sigma^2$. σ^2 is a measure of noise, i.e., the extent to which decision-makers disagree with one another, conditional on reviewing the same case.

Let X represent the hard information for each case, and the expected decision for a case given its hard information be $E[Z|X]$. Empirically, we approximate $E[Z|X]$ as $\hat{Z}$, equal to the predicted value from a regression of Z on observable hard information. Note $E[E[Z|X]\,] = Z^* + b$. Let soft information s be defined such that $Z^* + b = E[Z|X] + s$, with $E[s] = 0$. 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 outcome Z^*. We measure discretion in each case decision as $d = Z - E[Z|X]$. Combining terms yields $d = s + e$. Thus, discretion for each case is equal to a linear combination of soft information and additional error. Discretion can be valuable because it allows for use of soft information, but discretion can also be costly because it allow for additional error.

This framework allows us to estimate the amount of noise σ^2 due to discretion in our sample. Consider a setting where decision-makers are randomly assigned to cases, and each decision-maker makes decisions on multiple cases. Random assignment guarantees that each decision maker i is matched to cases with the same soft information in expectation, so $E[d \,|\, decision\ maker\ i] = E[e|decision\ maker\ i]$. We further assume that each decision maker has a fixed e. We can estimate $E[e|decision\ maker\ i]$ as the examiner fixed effect (or the examiner mean directional discretion). The variance of the examiner fixed effects, after applying a shrinkage adjustment, is an estimate of σ^2. Note that this is a conservative estimate in that we only account for noise due to fixed directional discretion at the examiner level. Examiners can introduce additional noise due to absolute discretion.

Cost of Noise

Let $F(Z - Z^*)$ represent the cost of deviations from the optimal case outcome. We follow Kahneman et al. (2019) and assume a quadratic cost function. This allows use to decompose the cost of discretion into separate bias and noise terms.

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

$$E[(Z - Z^*)^2] = E[(b + e)^2] = b^2 + \sigma^2.$$

b^2 represents the cost due to bias and σ^2 represents the cost due to noise.[17] 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. 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.

Applying to Our Setting

Applying these results to our empirical setting, we assume that an optimal CAMELS rating exists for each bank examination. One way to think about this is that these ratings translate into optimal outcomes in terms of capitalization and lending. If examiner discretion moves the CAMELS rating away from its optimal level, it will also move capitalization and lending away from the optimum, creating potentially costly noise. As econometricians, we are unable to observe Z^* and we do not make any assumptions regarding b, the extent to which the population of examiners is too harsh or too lenient in expectation. Nevertheless, we can measure the cost of noise as σ^2. The model further implies that two potential social planners who disagree about whether bank examiners are too tough or lenient on average (i.e., whether b is positive or negative) would nevertheless agree that the cost of noise is still equal to σ^2.

Finally, we note that our paper is focused on examining the determinants of discretion and measuring the *ex post* consequences and trade-offs associated with discretion. As such, this simple framework only captures the ex post consequences of noise. Of course, anticipation of future discretion by human decision-makers could also impact the ex-ante behavior of agents associated with the cases being considered. Anticipation of noise could have the benefit of preventing banks from gaming the examination system. We also present suggestive evidence that banks in states where examiners make noisier decisions take conservative actions ex ante by maintaining higher capitalization and lower lending.[18]

[17] 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.

[18] 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.

Appendix: Shrinkage

In this section, we account for sampling error in the fixed effects estimates of examiner directional discretion reported in Table 2. In finite samples, examiner fixed effects are estimated with error, leading to upward bias in the estimate of the variance of directional discretion across examiners.

We apply an Empirical Bayes shrinkage estimator to estimate the variance of the true magnitude of examiner fixed effects. Following the methods developed in Morris (1983), with applications provided in Chandra, Finkelstein, Sacarny, and Syverson (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.

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

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).

The standard deviation of $\hat{u}_j$ is 0.1628. We estimate that the standard deviation of u_j is 0.1348.

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.

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:

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

$f(x)$ 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. As shown in Table 1 Panel C, 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 0.1348, 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.

Appendix: Examiner experience and recent exposure

Experience effects

In addition to weights, disagreement can arise due to heterogeneous experience levels across examiners. Experience may cause examiners to process information differently, and to place more or less weight on observable hard bank information relative to other types of soft information (e.g., Malmendier and Nagel (2016); Chernenko, Hanson and Sunderam (2016)). Experience may also cause examiners to believe that left tail bank outcomes are more or less likely to occur for a given set of bank observables, leading to differences in directional discretion. Finally, research in psychology has suggested that experience and feelings of power may exacerbate biases and increase self-confidence, leading to more extreme exercise of human discretion (e.g., Griffin and Tversky (1992) and Stajkovic (2006)).

The table below shows how directional and absolute discretion measured at the exam level varies with the number of exams conducted by the examiner. We find that experience is associated with greater exercise of directional discretion and absolute discretion, although these differences become more noisily estimated after the inclusion of examiner fixed effects. Overall, we do not find any evidence that examiners exercise less discretion as they become more familiar with their jobs. The evidence suggests that more experienced examiners, if anything, exercise more absolute discretion and become less lenient as they gain experience.

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.

Exam-level discretion	(1) Directional	(2) Directional	(3) Absolute	(4) Absolute
Log(num exams so far)	0.006*	0.005	0.011***	0.002
	(0.003)	(0.005)	(0.002)	(0.003)
Observations	13,020	13,020	13,020	13,020
R-squared	0.000	0.188	0.002	0.212
Examiner FE	No	Yes	No	Yes

Recent exposure -- Serial autocorrelation in decisions

Examiner decisions may also be affected by recent exposure, which can further lead to disagreements over ratings. In this section, we show that bank observables do not appear to be serially correlated in the panel data ordered within examiner over time.[19] In other words, the true quality of the current bank does not appear to be positively or negatively related to the quality of the previous bank that an examiner was assigned to examine. Thus, the ratings decisions should also be serially uncorrelated if examiners are unbiased. However, we find that examiners assign ratings that are strongly serially correlated. Overall, we find a strong effect of recency that cannot be explained by observable bank characteristics. This positive autocorrelation in examiner ratings is consistent with examiners having extrapolative beliefs bias or slow-changing moods.

The table below explores serial autocorrelation in examiner ratings. Panel A presents regressions of the current rating on the lagged rating given by the examiner to the previously examined bank. The regressions control for the examiner's average rating, calculated excluding the current, lagged, and two-period lagged exams. Columns 1-3 use data ordered by the real sequence in which examiners rated banks.

We find that, after controlling for the examiner's general tendency to be tough or lenient across all exams excluding the current exam, examiners are significantly more likely to give a high rating to the current bank if they gave a high rating to the previous bank. The current rating is more strongly associated with the most recent previous rating than the rating two exams ago. Further, the findings cannot be attributed to a fixed examiner tendency to be more tough or lenient. In Column 4, we summarize the results of 1,000 permutation simulations which function as placebo tests. In each placebo simulation, we randomly order the exam ratings within an examiner over time, and estimate the relation between the current rating and recent ratings. The mean and standard deviation of these coefficient estimates over the 1,000 placebo tests are reported in Column 4. We find close-to-zero and insignificant serial correlation in these randomly ordered samples.

Panels B, C, and D provide support for our assumption that examiners are not assigned to review banks with serially correlated observable quality. In Panel B, observations are ordered by examiner-time. We regress characteristics of the current bank on the examiner's rating given to the previously reviewed bank. In Panel C, observations are again ordered by examiner-time. We regress

[19] For example, extrapolative beliefs (e.g., Greenwood and Shleifer (2014); Gennaioli, Ma, and Shleifer (2015) and Barberis et al. (2018)), experience effects (e.g., Malmendier and Nagel (2016)), or slow-changing moods and ideological drift (Goetzmann and Peles (1997), Martin and Quinn (2007), and Epstein et al. (2007)) could lead an examiner to make decisions with more serial correlation than would be justified by the correlation in case merits.

characteristics of the current bank on the same characteristic for the previous bank reviewed by each examiner. In Panel D, observations are ordered by bank-time. We regress the leave-out-mean directional discretion and absolute discretion associated with the examiner conducting the current exam on the leave-out-mean directional discretion and absolute discretion of the examiner who conducted the bank's most recent exam. In general, we find that observable characteristics of the current bank being examined are unrelated to the rating or characteristic of the previous bank assigned to the same examiner. Similarly, banks do not appear to be assigned to serially correlated examiners in terms of discretion over time. The only exception is that the tier1 capital ratio of the current and lagged bank are positively autocorrelated in Panel C. Given the number of outcomes evaluated, it is plausible that one significant autocorrelation could occur by chance.

Appendix Table 5: Recency effects and autocorrelation in examiner discretion

This table explores serial autocorrelation in examiner ratings. Panel A presents regressions of the current rating on the lagged ratings given by the examiner to previously examined banks. The regressions control for the examiner's average rating, calculated excluding the current, lagged, and two-period lagged exams. Columns 1-3 use data ordered according to the real sequence in which examiners rated banks. Column 4 presents a placebo test in which we randomly order bank examinations within each examiner within each of 1,000 simulations. The results in Column 4 report the average coefficient estimate and standard deviation within the set of 1,000 randomly-ordered simulations. Panels B and C present placebo tests to explore whether examiners are assigned to review banks with serially correlated observable characteristics. In Panel B, observations are ordered by examiner-time, and we regress characteristics of the current bank on the rating each examiner assigned to the previously reviewed bank. In Panel C, observations are ordered by examiner-time, and we regress characteristics of the current bank on the same characteristic for the previous bank reviewed by each examiner. Standard errors in parentheses are clustered by examiner. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels.

Panel A: Autocorrelation in examiner ratings

Examiner rating(t)	(1) Real order	(2) Real order	(3) Real order	(4) Random Order
Examiner rating($t-1$)	0.116***	0.041***	0.041***	0.012
	(0.011)	(0.014)	(0.015)	(0.014)
Examiner rating($t-2$)			0.025*	0.014
			(0.015)	(0.017)
Examiner average rating(excl. $t, t-1$)		0.155***		
		(0.024)		
Examiner average rating(excl. $t, t-1, t-2$)			0.091***	0.093***
			(0.028)	(0.024)
Observations	6,588	6,588	5,512	5,512
R-squared	0.395	0.639	0.660	0.645
Bank Controls	Yes	Yes	Yes	Yes
Examiner FE	No	No	No	No
Location-quarter FE	No	Yes	Yes	Yes

Panel B: Autocorrelation in bank quality assignments to examiners

	(1) Tier1 ratio	(2) Loan growth	(3) Leverage	(4) ROA	(5) NPL ratio	(6) Delinq ratio	(7) Efficiency
Composite rating($t-1$)	-0.366	-0.454	-0.133	-0.003	0.038	0.043	-0.122
	(0.242)	(0.575)	(0.116)	(0.007)	(0.054)	(0.081)	(0.390)
Observations	6,214	6,231	6,231	6,231	6,231	6,231	6,231
R-squared	0.653	0.677	0.657	0.689	0.737	0.722	0.923
Location-quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Panel C: Autocorrelation in bank observable characteristics assignment to examiners

	(1) Tier1 ratio	(2) Loan growth	(3) Leverage	(4) ROA	(5) NPL ratio	(6) Delinq ratio	(7) Efficiency
Dependent variable($t-1$)	0.075***	0.018	0.045	0.013	-0.011	-0.031	-0.026
	(0.030)	(0.033)	(0.029)	(0.026)	(0.032)	(0.029)	(0.031)
Observations	6,213	6,231	6,231	6,231	6,231	6,231	6,231
R-squared	0.654	0.677	0.657	0.689	0.737	0.722	0.923
Location-quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes