Weighted Noise: Discretion in Regulation

September 2025

Sumit Agarwal Bernardo Morais
Amit Seru Kelly Shue

The views expressed here are those of the authors and are not necessarily those of the Federal Reserve Board or System.

Human discretion and the problem of noise

Reliance on human discretion pervasive in institutional design

- Judges, scientific reviewers, patent examiners, HR committees…

Humans have a unique ability to process soft information

- Real but hard-to-quantify information that is open to interpretation

But, human discretion can lead to *noisy* judgments (Kahneman, Sibony, and Sunstein 2021)

- Noise: disagreement across decision-makers considering the same information
  - ➢ Assigned to a different decision-maker, same case receives an entirely different decision
- Noise → real distortions, uncertainty and volatility

This paper

We study bank examiners assessing the safety and soundness of US banks

Why do trained professionals disagree?

- Final decision as weighted sum of component level assessments
- Main sources of disagreement:
    1. Discretion applied to all components, even relatively objective ones
    2. High weight on the most subjective components: ~50% weight on Management
    3. Heterogeneity across examiners in how they weight components

Real effects due to noise (& benefits of soft information)

Healthy banks exposed to a 4.2% probability per exam of being rated unsatisfactory due to examiner assignment

- Majority of changes in bank ratings are due to changes in examiner assignment rather than changes in bank quality

Noise impacts bank behavior

- “Exogenous” unit increase in ratings due to examiner discretion causes a persistent 0.27 std dev increase in bank capitalization and a 1.08 std dev reduction in loan growth
- Anticipatory bank responses

Discretion can be beneficial for making predictions through usage of soft information

- However, moderate limits on discretion can translate into more informative predictions

Broader implications: costly noise

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)

- Random assignment only guarantees fairness
- The existence of strong instruments can be very bad for welfare, become outcomes are randomly assigned instead of well-matched to case characteristics

Setting and framework

Background

Prudential supervision of US banks: On-site exams

- A composite CAMELS rating: 1 (safe) to 5 (failing)

CAMELS rating: safety and soundness

Capital adequacy

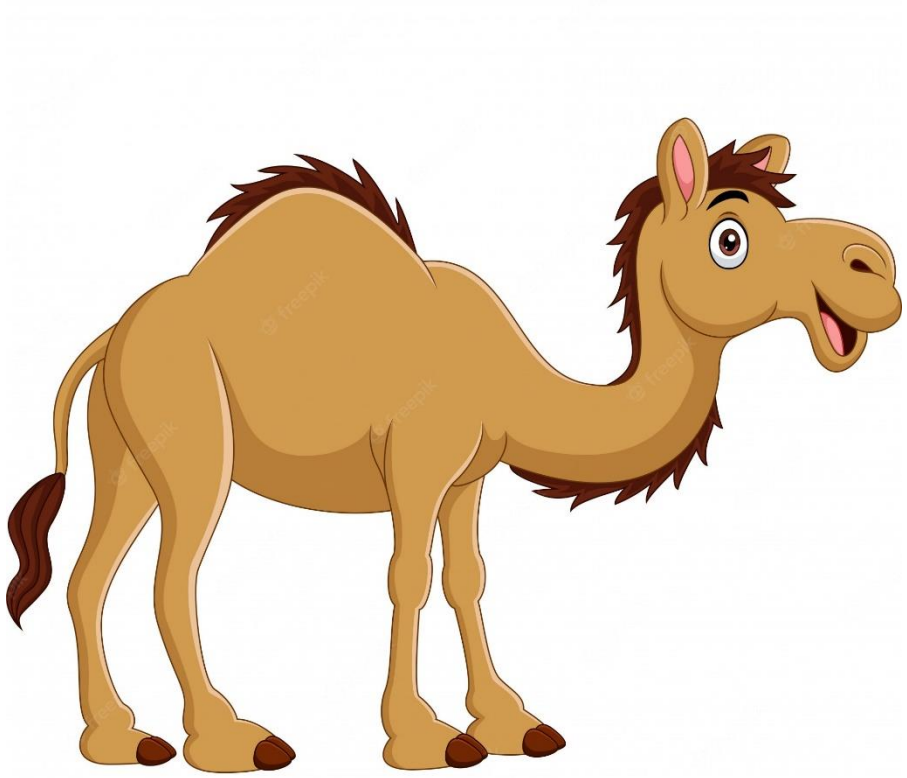
Asset quality

Management

Earnings

Liquidity

Sensitivity to market risk



CAMELS

Composite CAMELS rating has important consequences for banks

- Licensing, M&A, branching
- FDIC insurance premiums, discount window access
- Restructuring decisions, government funding (e.g. TARP)

"Unsatisfactory" rating: composite CAMELS of 3 or above

- Banks with unsatisfactory ratings generally increase capitalization and reduce lending

The composite rating … is not an arithmetic average of the individual component ratings. Rather, some components are weighed more heavily than others based on examiner judgment of risk. -- St. Louis Fed

Quasi-random examiner assignment

Our sample covers banks that are subject to regular examiner rotation

- Banks are examined by state and federal agencies in alternate years, and rotated across lead examiners within an agency-region
- **Random assignment assumption**: True bank quality (both hard and soft info) is uncorrelated with examiner identity within a region-time period
    - Empirically, observable measures of bank quality are uncorrelated with examiner assignment

We exclude very large banks from our sample because they are not subject to examiner rotation

- But examiner discretion could also impact larger national banks

Banks in our sample have average assets of 2 billion US$

- Includes large regional banks such as Silicon Valley Bank

Framework

Each case has optimal outcome Z^*

Let $Z = Z^* + b + e$ be the outcome determined by the human decision-maker

- b is bias: extent to which population of decision-makers is too harsh or lenient
- e is additional error by individual decision makers, with $E[e] = 0$ and $Var[e] = \sigma^2$
- σ^2 is noise: extent to which decision-makers disagree with one another

Bank fundamentals include observable hard info X and soft info s, with $\mathrm{E[Z]} = \mathrm{E[Z|X]} + s$

Discretion $d = \mathrm{Z} - \mathrm{E[Z|X]} = s + e$

- Pro: Discretion allows for use of soft information
- Con: Discretion adds individual noise

Random assignment: decision makers see the same soft information s in expectation

- $E[d|decision\ maker\ i] = E[e|decision\ maker\ i]$

Decisions as weighted sum of component issues

Final decision Z modeled as a weighted some of decisions over component issues C_l

$$Z = \sum_{l=1}^{k} w_l C_l$$

$C_1, \ldots, C_k$ and $w_1, \ldots, w_k$ are jointly independent

Disagreement is the cross-sectional variance in final decisions of examiners who review the same case

$$Var(Z) = \sum_{l=1}^{k} E[w_l^2]\ Var(C_l) + \sum_{l=1}^{k} Var(w_l)\, E[C_l]^2$$

Disagreement increases if:

1. Component issues (even objective ones) treated as subjective: $Var(C_l) > 0$
2. Greater weight attached to more subjective issues
3. Greater disagreement in weights $Var(w_l)$ across examiners, especially with subjective issues

Measuring discretion and noise

Measuring discretion

Discretion $d = \mathrm{Z} - \mathrm{E}[\mathrm{Z}|\mathrm{X}] = s + e$

Measure d as the residual rating after conditioning on observable bank hard information

$$Rating_{ijrt} = \alpha X_{jt} + \delta_j + \gamma_{rt} + \varepsilon_{ijrt}$$

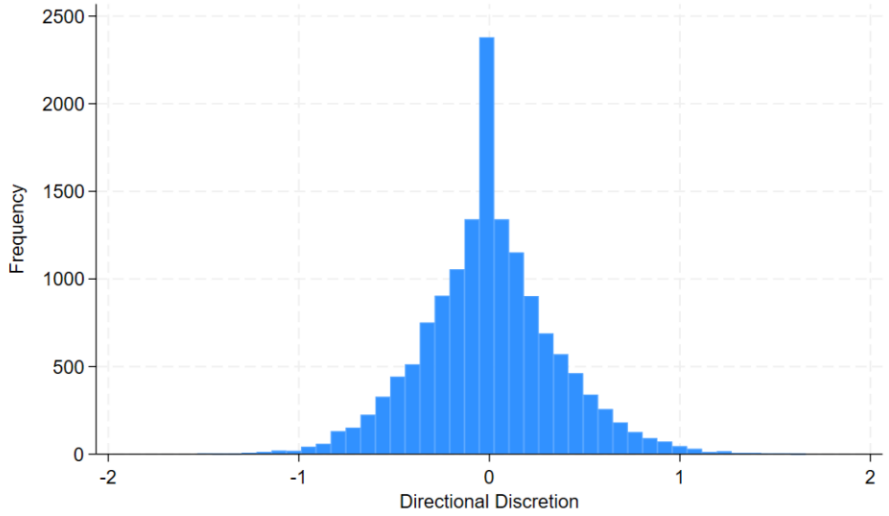
- Examiner *i*, bank *j,* region (state) *r*, year-quarter *t*

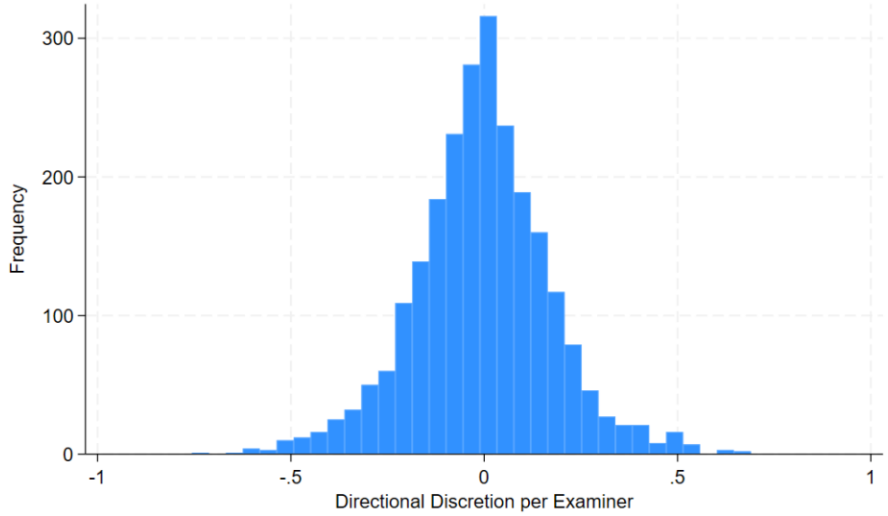
- **Examiner directional discretion** ≡ $\overline{d_i}$: Examiner is consistently more tough or lenient than others
  - Random assignment implies this examiner fixed effect can isolate noise

$$E[d|decision\ maker\ i] = E[e|decision\ maker\ i]$$

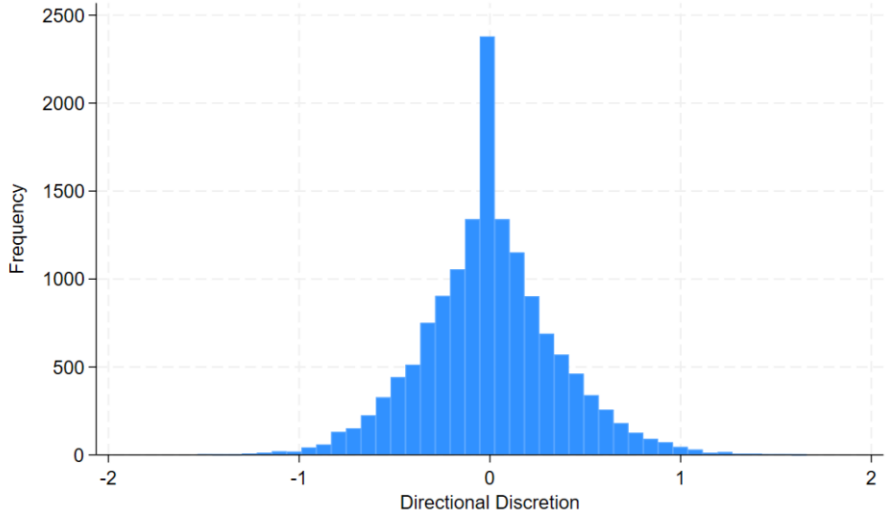
- **Examiner absolute discretion** ≡ $\overline{|d_i|}$: Examiner deviates from predicted rating in either direction
  - An examiner could have zero directional discretion, heavily weights case-specific soft information or gut feelings

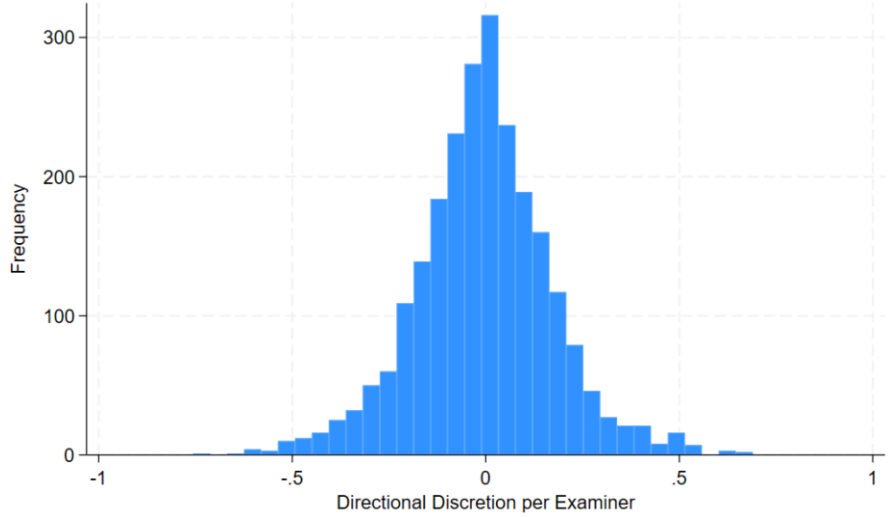
Exam Level





Examiner level





Magnitude of noise

Apply **Empirical Bayes shrinkage** because examiner fixed effects could vary due to finite sample size

Shrinkage-adjusted std dev of examiner mean directional discretion = 0.13

- Some examiners are systematically more harsh than others
- Larger than the federal vs. state regulator gap of 0.08 in Agarwal et al. 2014
- Exists within agency

Consider a healthy bank with CAMELS = 2 absent discretion

- Exposed to a 4.2% prob per exam of being rated unsatisfactory (≥3) and a 5.0% prob of being rated a 1, due to examiner discretion
- Can compare to overall transition probability of moving from a 2 to 3 of 6.7%
  - ➢ Majority of cases in which banks receive a different rating than in the previous year are due to changes in examiner rather than changes in quality

Why do we disagree?
Weights

Disagreement

Disagreement: cross-sectional variance in composite ratings that would arise if different examiners reviewed the same bank

For a small subsample, multiple examiners from different agencies assess the same bank simultaneously, and we can directly observe disagreement

Main sample: leverage the quasi-random rotation of examiners across banks within a region

- Examiners are assigned to banks with similar expected values of hard and soft information within a region and time → systematic differences in rating behavior reveal disagreement

Why do we disagree? Weights

Examiners have discretion in how to weight components to form the composite rating

$$R = w_C C + w_A A + w_M M + w_E E + w_L L + w_S S$$

1. Component issues (even relatively objective ones) are treated as subjective
2. High weight on the most subjective components
3. Heterogeneity in weights across examiners, especially if applied to subjective components

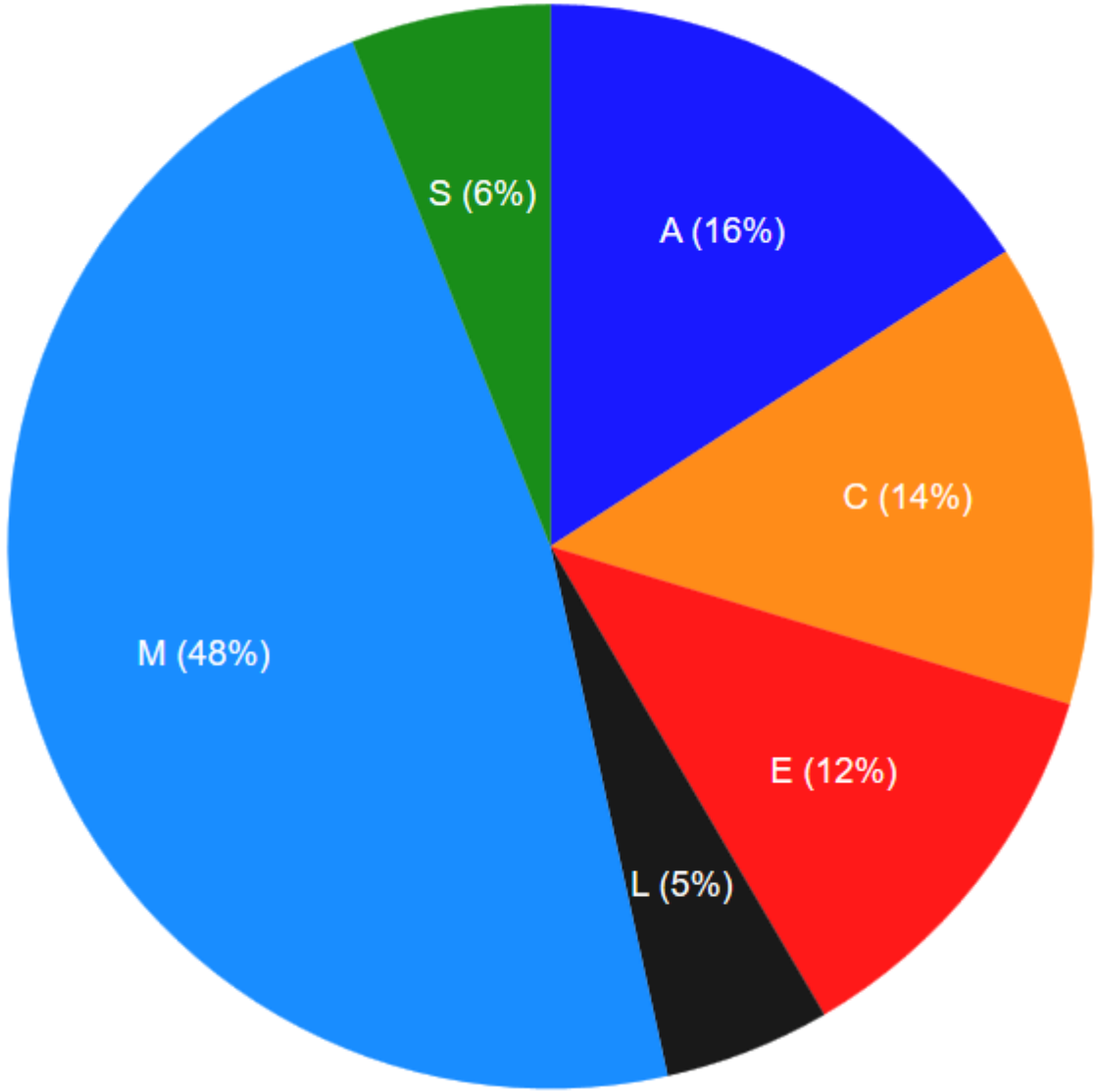
1. All components treated as subjective

Sample of simultaneous exams

- Management is the most subjective (most disagreement)
- Assets is second-most subjective
- But disagreement applies to all component ratings

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

2. High weight assigned to the most subjective issues



High weight on management quality is consistent with psychology research showing that people place too much weight on face-to-face interactions (Levine, Park, and McCornack 1999)

Halo effects and affective spillovers (Thorndike, 1920; Bol, 2011) show impressions in other areas can spillover into judgments in a very subjective area

3. Heterogeneity in weights across *individual* examiners

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

Estimate component weights for each of 415 examiners for whom we observe $\geq$10 exams

Greatest disagreement in weights for M, which is also the component with highest average weight

Real effects of noise

Causal effect of noise on bank behavior

DirecDiscLO$_{i,-jt}$ ≡ examiner's leave-out-mean directional discretion, excluding current exam

Jack-knife instrumental variable strategy

$$Rating_{ijrt} = \alpha\, DirecDiscLO_{i,-jt} + X_{jt} + \delta_j + \gamma_{rt} + \varepsilon_{ijrt}$$

$$BankOutcome_{j,t+1} = \beta^{IV}\, \widehat{Rating}_{i,-jt} + X_{jt} + \delta_j + \gamma_{rt} + \eta_{ijrt}$$

Quasi-random assignment means $DirecDiscLO_{i,-jt}$ is uncorrelated with bank quality (hard and soft), so $\widehat{Rating}_{i,-jt}$ captures noise e — the systematic variation in directional discretion across examiners

- β^{IV} measures causal effect of higher rating due to examiner noise

Impact of (exogenous) change in ratings, 2nd stage

Panel A: Full sample				
	(1)	(2)	(3)	(4)
Future bank outcome	Next rating	Troubled Bank	Tier1 ratio - 1yr	Loan growth - 1yr
Pred composite rating	-0.458*	-0.121**	0.680*	-11.922***
	(0.274)	(0.054)	(0.395)	(4.172)
Observations	12,802	12,802	13,040	13,033
Bank controls	Yes	Yes	Yes	Yes
Bank FE	Yes	Yes	Yes	Yes
Location-quarter FE	Yes	Yes	Yes	Yes

Exogenous unit in ratings is associated with bank taking conservative actions:

- 0.46 point decrease in the next CAMELS rating
- 0.27 std dev increase in bank capitalization
- 1.08 std dev decline in loan growth

Anticipatory response

Anticipating the uncertainty caused by examiner discretion, banks may take conservative actions ex ante

- Bloom (2009): Firms reduce investment and hiring due to macroeconomic uncertainty
- Banks in states where examiners exercise a high degree of absolute discretion or high variation in directional discretion have a greater anticipatory response

	(1) Tier1 ratio	(2) Tier1 ratio	(3) Loan growth	(4) Loan growth
State absolute discretion	0.82*		-4.76**	
	(0.43)		(1.88)	
State sd direction discretion		0.63*		-3.20**
		(0.35)		(1.54)
Observations	14,237	14,228	14,067	14,067
R-squared	0.47	0.46	0.12	0.11
Quarter FE	Yes	Yes	Yes	Yes
Lagged Bank Controls	Yes	Yes	Yes	Yes

Is there an upside to discretion and can we make it better?

Does discretion lead to "better" ratings that are more predictive of bank health?

$$d = s + e$$

- Is there a way to make human ratings more predictive and less noisy?

Measurement challenge: Ratings both affect and predict future bank outcomes

- **Before**: IV measured causal effect of noise e, instrumented by leave-out-mean
- **Now**: assess predictive power of the $d = s + e = Rating - RatingPred$

Do ratings predict outcomes that are unlikely to be affected by ratings

- Performance of loans made prior to the exam
- Exploit the fact that the **effect** and **predictive power** of ratings go in opposite directions: higher rating should predict higher bank risk but cause lower risk

Does ***more*** discretion equal better predictions?

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

- Soft information can help examiners forecast bank health
- But high absolute discretion examiners introduce more noise without making better predictions

Comparing actual ratings to machine-driven and constrained ratings

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

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

Composite rating: actual rating

Predicted rating: predicted rating from a regression of actual ratings on bank observables

- Parallel to the naïve machine algorithm in Dawes et al. 1989

Reweighted rating: weights chosen to provide the best estimate of bank's composite rating in the next year

- Weight on M = 29% instead of 50%

Equal weighted rating: all components weighted equally

Constraints on weights increase predictive power of ratings while reducing noise

Improving decision-making

Human discretion can be useful

- Contrary to the more extreme conclusions of the algorithm aversion literature, professional bank examiners outperform simple linear regressions

But putting bounds on human discretion could be useful for predictive power

- Some introduce more noise without additional predictive power
- Constraints on weights lead to more accurate predictions with less noise

Weights are an important cause of disagreement

- Reaching agreement on most issues will fail to lead to consensus on the final verdict if we disagree in how to weight issues and/or heavily weight a highly subjective issue

Conclusion

Bank examiners exhibit high degrees of directional and absolute discretion

Disagreement arises from differences in weights over components, the heavy weighting of the a highly-subjective issue, and treating even relatively objective issues as subjective

Discretion by regulators impacts bank capitalization and credit supply

Discretion can be valuable, but higher discretion does not translate into obviously more accurate predictions of future bank observables

- Placing moderate limits on human discretion can translate into better predictions
- Modern regulation: trade-off between expert discretion and systemic noise