Winter 2019

Rating Analyst Degrees of Freedom

Vijay Raghavan

RATING ANALYST DEGREES OF FREEDOM

Vijay Raghavan[*]

I. INTRODUCTION

Research findings in behavioral psychology have fundamentally transformed the way we approach and solve legal problems.[1] In recent years, behavioral psychology has lost some of its luster due to the inability of researchers to replicate some of the central findings in the field.[2] The so-called "replication crisis" raises concerns about the reliability of individual psychological studies[3] and general criticism of the reliance on null hypothesis significance testing ("NHST") in experimental psychology.[4] The replication crisis is largely viewed as a problem limited to experimental social science, and the existing literature on the crisis focuses on how to fix problems in these fields. Yet the crisis, particularly the methodological criticisms of NHST, can help us both frame problems in other markets that rely on statistical modeling as well as understand the limits of existing legal solutions. This article applies these insights to one such market that is widely recognized as broken but where there is little consensus on exactly what the problem is and how to fix it: the credit rating agency market.

In the aftermath of the financial crisis of 2008, academics and lawmakers broadly agreed that credit rating agencies were partially responsible for the

[*] Assistant Attorney General, Illinois Attorney General's Office. The views expressed in this article are my own and do not reflect the official opinion of the Illinois Attorney General's Office. I wish to thank Som Bhattacharya, Kaitlin Caruso, Gregory Jones, Uppili Krishnamachari, Adam Levitin, Frank Partnoy, and Jonathan Reischl for valuable feedback and comments on earlier drafts. I also wish to thank Jonathan Brown, Ashley Crisafulli, and the editors of the *UMKC Law Review*. All errors are my own.

[1] *See, e.g.*, Cass R. Sunstein & Richard H. Thaler, *Libertarian Paternalism Is Not an Oxymoron*, 70 U. CHI. L. REV. 1159 (2003) (an early paper from Sunstein and Thaler arguing for a new paradigm in rulemaking based on insights from behavioral psychology); Donald C. Langevoort, *Taming the Animal Spirits of the Stock Market: A Behavioral Approach to Securities Regulation*, 97 NW. U. L. REV. 135 (2002) (applying insights from behavioral finance and economics to the regulation of financial markets after Enron); Stephanos Bibas, *Plea Bargaining Outside the Shadow of Trial*, 117 HARV. L. REV. 2463 (2004) (using insights from behavioral psychology to challenge conventional wisdom about motives underlying pre-trial plea bargaining).

[2] *See* Kristin Firth, David A. Hoffman & Tess Wilkinson-Ryan, *Law and Psychology Grows Up, Goes Online, and Replicates*, 15 J. EMPIRICAL LEGAL STUD. 320, 334-53 (2018) [hereinafter *Law and Psychology Grows Up*] (describing the replication crisis as a source of skepticism about the value of psychological research in legal analysis).

[3] *See id.*

[4] *See* Joseph Simmons, Leif Nelson & Uri Simonsohn, *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allow Presenting Anything as Significant*, 22 PSYCHOL. SCI. 1359 (2011) [hereinafter Simmons et al., *False-Positive Psychology*]; *see also* Andrew Gelman & Erik Loken, *The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No "Fishing Expedition" or "P-hacking" and the Research Hypothesis was Posited Ahead of Time* (Nov. 14, 2013) (unpublished) (available online at http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf) [hereinafter Gelman & Loken, *Garden of Forking Paths*].

collapse of global financial markets.[5] Legal scholars, however, disagreed about why the rating agencies failed to predict the crisis. Competing explanations included: insufficient oversight of the rating agencies by the Securities and Exchange Commission ("SEC");[6] too much regulation, which led to rent-seeking and excessive reliance on credit ratings;[7] and a compensation model that incentivized rating agencies to rate securities consistent with issuer expectations.[8] Though distinct, these descriptive accounts are unified in their understanding of the problem in the ratings market as external to the ratings process. In contrast, this article suggests the problems in the ratings industry are best understood as *internal* to the credit ratings process.

Modern credit rating agencies rely on sophisticated statistical models to analyze and rate debt offerings.[9] Analysts at rating agencies face complex design decisions in developing these models, including how much data to collect and which statistical techniques to use. In managing these decisions, this article suggests that rating analysts may be driven by their own subjective bias to build

[5] The story is detailed at great length in the sources cited below, but to briefly summarize: Standard & Poor's, Moody's, and Fitch systematically underestimated the risk that mortgages backing or referenced by residential mortgage backed securities ("RMBS") and collateralized debt obligations ("CDO") would default and rated these securities as relatively risk free. Rating agencies' assessment of RMBS and CDOs as safe investments fueled demand for these securities, which, in turn, allowed mortgage lenders to finance excessively risky mortgages that would ultimately default in high numbers, precipitating an international credit crisis. *See, e.g.*, FIN. CRISIS INQUIRY COMM'N, THE FINANCIAL CRISIS INQUIRY REPORT 146 (2011) ("[t]he machine churning out CDOs would not have worked without the stamp of approval given to these deals by the three leading rating agencies: Moody's, S&P, and Fitch"); Jeffrey Manns, *Downgrading Rating Agency Reform*, 81 GEO. WASH. L. REV. 749, 754 n.16 (2013) [hereinafter Manns, *Downgrading*] (collecting empirical studies documenting the failures of the rating agencies)

[6] *See, e.g.*, Lynn Bai, *The Performance Disclosures of Credit Rating Agencies: Are They Effective Reputational Sanctions?*, 7 N.Y.U. J. L. & BUS. 47, 97–98 (2010); Milosz Gudzowski, Note, *Mortgage Credit Ratings and the Financial Crisis: The Need for a State-Run Mortgage Security Credit Rating Agency*, 2010 COLUM. BUS. L. REV. 245, 264–71 (2010).

[7] *See, e.g.*, Mark Flannery, Joel Houston & Frank Partnoy, *Credit Default Swap Spreads as Viable Substitutes for Credit Ratings*, 158 U. PA. L. REV. 2085, 2086–89 (2010); Jonathan R. Macey, *The Politicization of American Corporate Governance*, 1 VA. L. & BUS. REV. 10, 21–24 (2006); Frank Partnoy, *Overdependence on Credit Ratings was a Primary Cause of the Crisis*, *in* THE PANIC OF 2008: CAUSES, CONSEQUENCES, AND IMPLICATIONS FOR REFORM 116, 116-31 (Lawrence Mitchell et al. eds., 2010).

[8] Yair Listokin & Benjamin Taibleson, *If You Misrate, Then You Lose: Improving Credit Rating Accuracy Through Incentive Compensation*, 27 YALE J. REG. 91, 94–95 (2010); Jeffrey Manns, *Rating Risk After the Subprime Mortgage Crisis: A User Fee Approach for Rating Agency Accountability*, 87 N.C. L. REV. 1011, 1015–19 (2009).

[9] *See, e.g.*, Ingo Fender & John Kiff, *CDO Rating Methodology: Some Thoughts on Model Risk and its Implications* 3-8 (Bank for Int'l Settlements, Working Paper No. 163, 2004), http://www.bis.org/publ/work163.pdf (describing differences between Moody's binomial expansion technique and monte-carlo simulation to rate CDOs); Frank Partnoy, *How and Why Credit Rating Agencies Are Not Like Other Gatekeepers*, *in* FINANCIAL GATEKEEPERS: CAN THEY PROTECT INVESTORS? 59, 76-78 (Yasuyuki Fuchita & Robert Litan eds., 2006) [hereinafter Partnoy, *Gatekeepers*] (describing Standard & Poor's use of monte-carlo simulations to establish the default thresholds for a proposed pool of assets in a CDO to achieve particular ratings).

models that confirm their *a priori* beliefs about default risk. To make this case, this article relies on the descriptive framework created by critics of NHST specifically in the field of behavioral psychology.

Null hypothesis significance testing is a statistical test commonly used in experimental science to determine if a causal relationship exists between two variables and is central to many psychological studies. Critics of NHST contend that experimental psychology placed too much faith in NHST and failed to adequately constrain the risk of error NHST poses.[10] These critics have developed a framework to understand problems with NHST that rests on two important insights about the risk of error in statistical modeling more generally: (1) "researcher degrees of freedom," including flexibility in data selection and analysis, can lead researchers to underestimate the existence of false positives; and (2) a researcher's "implicit bias" towards finding statistical significance may lead the researcher to pursue a path that appears data-driven but is, in fact, biased towards finding false positives even without bad faith or malicious intent.[11] This article's central claims are that the criticisms of NHST in behavioral psychology are equally applicable to the statistical models rating agencies use to rate debt offerings and also that rating agency errors are best understood as the product of rating analysts' degrees of freedom (i.e., flexibility in data selection and model choice in model development) and subjective bias.

To be sure, this article's central claims should not be taken to mean that rating agency errors are *in fact* the product of subjective bias. Rather, this article contends that rating agency errors are best understood as the product of subjective bias because errors in the ratings process can be rationalized ex post in a manner that makes it difficult for regulators, law enforcement, or private litigants to determine if the errors were the product of bad faith or honest mistake. The credit rating agency reforms in Dodd-Frank[12] rely on regulators' ability to distinguish between bad faith and honest mistake to curb abuse in the ratings market. This article attempts to show this reliance is misplaced and, in the absence of evidence that rating agencies are unambiguously operating in bad faith, regulators will not be able to detect and discourage objective mistakes in the ratings process.

This article concludes by suggesting some reforms the SEC can make within the current framework of Dodd-Frank that would result in less biased and more accurate credit ratings. There is an emerging consensus that the rating agency reforms in Dodd-Frank are insufficient to address problems with the ratings market.[13] This article agrees with this consensus but suggests that the

[10] *See* Simmons et al., *False-Positive Psychology*, *supra* note 4, at 1359.

[11] *See id.*; Gelman & Loken, *Garden of Forking Paths*, *supra* note 4, at 2.

[12] *See* Dodd-Frank Wall Street Reform and Consumer Protection Act, Pub. L. No. 111-203, 124 Stat. 1376 (2010).

[13] *See* Frank Partnoy, *What's (Still) Wrong with Credit Ratings*, 92 WASH. L. REV. 1407 (2017), [hereinafter Partnoy, *Wrong*]; Manns, *Downgrading*, *supra* note 5, at 750; Claire Hill, *Limits of Dodd-Frank's Rating Agency Reform*, 15 CHAP. L. REV. 133, 133 (2011) [hereinafter Hill, *Limits*];

problems with Dodd-Frank are ones of implementation, not design. This article suggests two reforms to rating agency regulation, which can be achieved within the current framework of Dodd-Frank through interpretive guidance or formal rulemaking: (1) liberalizing the application process to become a nationally recognized statistical rating organization ("NRSRO");[14] and (2) making rating agencies compete to preserve their NRSRO status. Specifically, the SEC should propose rules that define integrity and accuracy relative to the performance of other rating agencies and revoke an agency's license for a class of securities if its performance is poor. This article explains how the SEC can effectuate these reforms and addresses complications with each.

This article proceeds in four parts: Part II describes the replication crisis in behavioral psychology and the critiques of the use of NHST by behavioral psychologists. Part III applies the insights of NHST critics to the context of credit ratings by walking the reader through the creation of a simple predictive model using machine learning techniques and public data. The goal in Part III is to show the reader how rating analysts can manipulate flexibility in data collection and model choice to build predictive models that underestimate default risk. Part IV uses the model developed in Part III as the subject of a hypothetical regulatory investigation. Part IV attempts to show that the reforms in Dodd-Frank, as currently applied, are insufficient to prevent rating analysts from making objective and consequential errors in the ratings process. Finally, Part V suggests modest reforms within the current framework of Dodd-Frank to minimize the effects of subjective bias in model design.

II. RESEARCHER DEGREES OF FREEDOM AND THE REPLICATION CRISIS IN BEHAVIORAL PSYCHOLOGY

The field of behavioral psychology has been plagued in recent years by the so-called "replication crisis": the failure of many foundational results in the field to replicate under different experimental settings.[15] Fallout from the crisis

Carrie Guo, Note, *Credit Rating Agency Reform: A Review of Dodd-Frank Section 933(B)'s Effect (Or Lack Thereof) Since Enactment*, 1 COLUM B.L. REV. 184, 187 (2016) [hereinafter Guo, *Credit Rating Agency Reform*].

[14] "NRSRO" is the formal designation federal law provides for agencies licensed to issues ratings by the SEC.

[15] Though there have been many articles on the crisis, perhaps the most notable is the Open Science Collaboration study reporting the results of attempting to replicate 100 published experimental findings and finding only a small percentage of statistically significant results were replicable. The Open Science Collaboration study was the subject of fierce debate between the studies' authors and prominent behavioral psychologists such as Daniel Gilbert. *See generally* Open Science Collaboration, *Estimating the Reproducibility of Psychological Science,* 349 SCIENCE aac4716-1 (2015). For the full exchange, *see generally* Daniel Gilbert et al., *Comment on* "Estimating the Reproducibility of Psychological Science", 351 SCIENCE 1037-a (2016); Christopher Anderson et al., *Response to* Comment on "Estimating the Reproducibility of Psychological Science", 351 SCIENCE 1037b (2016); Daniel Gilbert et al., *A Response to the Reply to Our Technical Comment*

has led to deep skepticism among some academics of novel experimental results in the field;[16] skepticism that is based on both the weak replicability of past results and concern that flexibility in experimental design can lead to engineered and incorrect results. This second sentiment is perhaps best captured by two recent articles: Joseph Simmons, Leif Nelson and Uri Simonsohn's 2011 paper *False-Positive Psychology*;[17] and Andrew Gelman and Erik Loken's 2013 paper *The Garden of Forking Paths*.[18]

In *False-Positive Psychology*, Simmons, Nelson, and Simonsohn explain how flexibility in data collection and analysis in NHST can lead to the incorrect rejection of the null hypothesis (i.e., reporting false-positives as true positives). NHST is an inferential statistical test commonly used in experimental research to test a hypothesis (e.g., that a carbohydrate-free diet causes weight loss). The basic NHST procedure is to formulate a null hypothesis (carbohydrate-free diet has no effect on weight loss), define a test statistic (average weight loss), observe the distribution of the test statistic across a population sample, and determine the probability of observing a result as or more extreme than the observed test statistic (commonly referred to as the "p-value"). If the p-value is less than a predefined threshold for significance (typically 5%), then the null hypothesis can be rejected. Rejection of the null hypothesis does not mean that the alternative hypothesis is true, but rather the likelihood of the alternative hypothesis being false is low (i.e., it is unlikely the test yielded a false positive or type II error). However, as Simmons, Nelson, and Simonsohn argue, "despite the nominal endorsement of a maximum false-positive rate of 5% (i.e., $p \leq .05$), current standards for disclosing details of data collection and analyses [in psychological research] make false positives vastly more likely."[19] The authors identify the source of high false positive rates as *researcher degrees of freedom*:

> In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?

on "Estimating the Reproducibility of Psychological Science" (unpublished) (available online at http://gking.harvard.edu/files/gking/files/gkpw_response_to_osc_rebutal.pdf).

[16] *See, e.g.*, STATISTICAL MODELING, CAUSAL INFERENCE, AND SOCIAL SCIENCE BLOG, http://www.andrewgelman.com (last visited Sept. 14, 2018). For an extended post on the evolution of the crisis, see Andrew Gelman, *What Has Happened Down Here is the Winds Have Changed*, STATISTICAL MODELING, CAUSAL INFERENCE, AND SOCIAL SCIENCE BLOG (Sept. 21, 2016, 9:03 AM), http://andrewgelman.com/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/. Few have chronicled the replication crisis and its implications better than Andrew Gelman, whose work is generally available on his blog.

[17] *See* Simmons et al., *False-Positive Psychology*, *supra* note 4, at 1359-66.

[18] *See* Gelman & Loken, *Garden of Forking Paths*, *supra* note 4.

[19] Simmons et al., *False-Positive Psychology*, *supra* note 4, at 1359.

> It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields "statistical significance," and to then report only what "worked." The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.[20]

Simmons, Nelson, and Simonsohn argue that exploiting researcher degrees of freedom may not be the product of "malicious intent, but rather the result of two factors: (a) ambiguity in how best to make these decisions, and (b) the researcher's desire to find a statistically significant result."[21] To demonstrate how flexibility in experimental design can result in high false positive rates, the authors tested the following four researcher degrees of freedom on 15,000 simulated samples randomly drawn from a normal distribution: "(a) choosing among dependent variables, (b) choosing sample size, (c) using covariates, and (d) reporting subsets of experimental conditions."[22] The authors tested each of these conditions on each sample for statistical significance with various p-values. The results showed that combining all four researcher degrees of freedom results in rejecting the null hypothesis in 61% of cases ($p < .05$), which results in a false-positive rate of at least 61%.[23] With such high false positive rates, "[a] researcher is more likely to falsely detect a significant effect by just using these four common researcher degrees of freedom."[24] Simmons, Nelson, and Simonsohn then exploit these researcher degrees of freedom in a mock study of their own and find significant results that are likely false (certain songs make listeners feel younger) and definitely false (certain songs *make* listeners younger).[25]

In the *Garden of Forking Paths*, Gelman and Loken expand on Simmons, Nelson, and Simonsohn's work and introduce the following taxonomy of test procedures:

> 1) Simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$.
>
> 2) Classical test pre-chosen from a set of possible tests: thus, $T(y;\varphi)$, with preregistered φ. For example, φ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.

[20] *Id.*
[21] *Id.*
[22] *Id.* at 1360.
[23] *See id.* at 1361.
[24] *Id.*
[25] *See id.* at 1360.

> 3) Researcher degrees of freedom without "fishing": computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y;\varphi(y))$, where the function $\varphi(\cdot)$ is observed in the observed case.
>
> 4) "Fishing": computing $T(y;\varphi_j)$ for $j=1,\ldots,J$: that is, performing J tests and then reporting the best result given the data, thus $T\ (y;\ \varphi^{best}(y))$.[26]

Per Gelman and Loken, researchers may still exploit researcher degrees of freedom even if they are not explicitly fishing for statistical significance.[27] To illustrate, Gelman and Loken examine Alec Beall and Jessica Tracy's 2013 paper *Women Are More Likely to Wear Red or Pink at Peak Fertility*.[28] Gelman and Loken contend that the authors of the study exploited several researcher degrees of freedom to arrive at their conclusion.

First, in selecting the sample, the authors included a sizeable percentage of women who did not fit their selection criteria (the criteria specified that women should be younger than forty and more than five days away from the onset of menses) and excluded women who lacked confidence in answering how close they were to their menstrual cycle.[29] Second, multiple hypotheses could have supported the authors' finding of a statistically significant pattern:

> [T]he authors found a statistically significant pattern after combining red and pink, but had they found it only for red, or only for pink, this would have fit their theories too. In their words: "The theory we were testing is based on the idea that red and shades of red (such as the pinkish swellings seen in ovulating chimpanzees, or the pinkish skin tone observed in attractive and healthy human faces) are associated with sexual interest and attractiveness." Had their data popped out with a statistically significant difference on pink and not on red, that would have been news too. And suppose that white and gray had come up as the more frequent colors? One could easily argue that more bland colors serve to highlight the pink colors of a (European-colored) face.[30]

Finally, the authors' definition of "peak fertility" as days six through fourteen of the menstrual cycle seemed to be contingent on the data because different sources prescribe different ranges and there does not appear to be an accepted standard range in the literature.[31] Gelman and Loken argue though the authors may not be fishing, they start with a strong subjective hypothesis and implicitly choose routes in the "garden of forking paths" of experimental design

[26] Gelman & Loken, *Garden of Forking Paths*, *supra* note 4, at 2.
[27] *See id.*
[28] *See generally* Alec Beall & Jessica Tracy, *Women Are More Likely to Wear Red or Pink at Peak Fertility*, 24 PSYCHOL. SCI. 1837, 1837-41 (2013).
[29] *See* Gelman & Loken, *Garden of Forking Paths*, *supra* note 7, at 8.
[30] *Id.* at 9.
[31] *See id.*

that reinforce their *a priori* beliefs and steer them towards statistical significance:

> When we say an analysis was subject to multiple comparisons or "researcher degrees of freedom," this does not require that the people who did the analysis were actively trying out different tests in a search for statistical significance. Rather, they can be doing an analysis which at each step is contingent on the data. The researcher degrees of freedom do not feel like degrees of freedom because, conditional on the data, each choice appears to be deterministic. But if we average over all possible data that could have occurred, we need to look at the entire garden of forking paths and recognize how each path can lead to statistical significance in its own way.[32]

III. RATING ANALYST DEGREES OF FREEDOM

Although concerns about researcher degrees of freedom in behavioral psychology seem far removed from the business of credit ratings, there are striking parallels between the work of rating analysts and behavioral psychologists. Rating agencies use statistical models to rate debt instruments. As the complexity of debt instruments has grown in the past few decades, the models rating agencies use to rate these instruments have become increasingly sophisticated. In developing quantitative models and using these models to rate debt instruments, rating analysts, like behavioral psychologists, face challenging design decisions. These design decisions do not necessarily have correct solutions and resolving each provides rating analysts with degrees of freedom in the model development process. Rating analysts can exploit this freedom, intentionally or not, to build models that significantly underestimate default risk.

In the legal literature on credit ratings, rating agency methodologies are typically discussed at an abstract level and few scholars engage with the actual models rating agencies use or the dynamics of building predictive models.[33] Part III seeks to provide the reader with a more detailed understanding of how rating analysts build predictive models and how the choices analysts make in the model development process can impact the accuracy of a predictive model. This part provides background on the ratings process, walks the reader through the creation of a simple model using machine learning techniques and public data, and then summarizes evidence from post-financial crisis litigation and empirical work about subjective bias in rating agency models.

A. Background on the Ratings Process and the Use of Quantitative Methods to Rate Debt

Rating agencies, such as Standard & Poor's and Moody's, assign letter

[32] *Id.* at 10-11.

[33] A notable exception is Frank Partnoy, whose recent work is described in detail *infra* Part III.

grades to debt instruments. These letter grades correspond to the rating agency's belief about the risk of a particular debt instrument, with higher letter grades corresponding to lower risk and lower letter grades corresponding to higher risk. The specific meaning of "risk," however, varies with the rating agency. For example, Standard & Poor's ratings correspond to the probability that a particular debt instrument will default and not pay out in full;[34] whereas Moody's ratings correspond to the loss an investor can expect with respect to a particular debt instrument given its default probability.[35] Thus, a "AAA" rating from Standard & Poor's with respect to a series of corporate bonds indicates that the corporation issuing the bonds is likely to meet its financial commitments;[36] while an "Aaa" rating from Moody's means the expected loss of the same corporate bonds is very small.[37]

Rating agencies primarily rely on quantitative methods to assign ratings to debt instruments. The complexity of the methods used typically mirrors the complexity of the debt instrument that is being rated. For example, the techniques rating agencies use to rate traditional corporate bonds are relatively simple.[38] By contrast, more complex instruments, such as asset-backed securities ("ABS") or CDOs, require more sophisticated techniques. ABS are securities issued by a trust that are backed by the stream of payments from a pool of financial assets such as residential mortgage loans.[39] ABS issuances are commonly offered in tiers (or tranches) where each tranche carries a different rating and yield. Similarly, CDOs are tiered securities issued by trusts and collateralized by debt securities such as ABS tranches.

Unlike a corporate bond, in analyzing the default risk of complex instruments like CDOs or ABS, rating agencies consider more than just the default risk of a single entity. Rating agencies consider, at a minimum, the default risk of the individual assets in a CDO or ABS, the correlation risk between assets in the pool, and how default and correlation risk affect the likelihood that a CDO or ABS will be able to meet financial commitments to investors in each tranche. Two examples of the kinds of complex models rating agencies used to rate these assets are Moody's Binomial Expansion Technique ("BET") and Standard & Poor's CDO Evaluator—both used to rate CDOs.

Moody's pioneered an early way to assess the risk of CDOs twenty years ago with its BET model. The BET model turned a diverse pool of correlated

[34] *See S&P Global Ratings Definitions*, STANDARD & POOR'S FIN. SERVS. (Apr. 19, 2018) http://www.standardandpoors.com/en_US/web/guest/article/-/view/sourceId/504352.

[35] *See* MOODY'S INVESTORS SERV., RATING SYMBOLS AND DEFINITIONS 5, 13 (2018), http://www.moodys.com/researchdocumentcontentpage.aspx?docid=PBC_79004 [hereinafter MOODY'S RATING DEFINITIONS].

[36] *See S&P Global Ratings Definitions*, *supra* note 34.

[37] *See* MOODY'S RATINGS DEFINITIONS, *supra* note 35.

[38] *See* Partnoy, *Wrong*, *supra* note 13, at 28-36 (detailing Standard & Poor's current corporate methodology).

[39] Where ABS are collateralized by residential mortgages, the securitizations are commonly referred to as RMBS.

CDO assets into a hypothetical pool of uncorrelated assets with a simple binomial default distribution.[40] The BET model then calculated the total expected loss for all potential default scenarios in a securitization and checked this value against empirically derived expected loss rates to determine the rating for a CDO tranche.[41]

Moody's BET model was relatively simple but somewhat inflexible.[42] Several years after Moody's BET model hit the market, Standard & Poor's developed its CDO Evaluator, which was a more complex and modern approach that rated CDOs using monte-carlo simulation.[43] Monte-carlo simulation allows one to repeatedly simulate the performance of a random event (flipping a coin, for example, or making monthly payments on a mortgage) to predict the likelihood of a certain outcome (obtaining tails or the likelihood that a mortgage defaults at a certain time after origination).[44] Standard & Poor's CDO Evaluator used monte-carlo simulation along with empirically derived data on correlation risk and default probabilities of rated assets to rate CDOs.[45]

The BET model and CDO Evaluator are distinct approaches to modeling default risk of CDOs but share features common to all predictive models: each relies on *data* (expected loss rates in the case of the BET model and default and correlation tables in the case of CDO Evaluator) and a *predictive method* (binomial distribution in the case of the BET model and monte-carlo simulation in the case of CDO Evaluator) to make predictions. In choosing a predictive method and making assumptions about expected loss or correlation risk based on data, rating analysts—like behavioral psychologists—face complex decisions including: How should they model the default risk of individual assets pooled together in a securitization? How should they model the default correlation between these assets? How should they model pool default risk based on the default and correlation risk of assets in the pool? How much data should they collect to build these models?

The next section walks the reader through two of these design decisions—data selection and model choice—in the context of a hypothetical model and shows how each design decision can ultimately bias the predictions of a model.

B. Rating Analyst Degrees of Freedom in the Context of a Hypothetical ABS Model

A central claim in this article is that degrees of freedom and subjective

[40] A bell-shaped curve for a binary outcome.

[41] *See* ARTURO CIFUENTES & GERARD O'CONNOR, MOODY'S STRUCTURED FIN. SPECIAL REPORT, THE BINOMIAL EXPANSION METHOD APPLIED TO CBO/CLO ANALYSIS, at 2-4 (1996).

[42] *See* Fender & Kiff, *supra* note 9, at 3.

[43] *See id.* at 22.

[44] *See id.* at 3.

[45] *See id.* at 3-8; *see also* Partnoy, *Gatekeepers*, *supra* note 9, at 76-78.

bias in the model development process leads rating analysts to create predictive models that are biased towards finding fewer defaults. To illustrate this point, this section walks the reader through the creation of a simple model to predict the default risk of a single tranche ABS collateralized by unsecured personal loans. The hypothetical ABS model will be created by using classic techniques from machine learning and publicly available loan data.[46] I opt for machine learning methods over other techniques such as monte-carlo simulation for two reasons. First, classic machine learning methods (as opposed to black-box methods) are relatively straightforward and the machine learning libraries this article relies on to build this model are publicly available. Second, while rating agencies used simulation-based models and mathematical models to make up for scant data in the past, with the advent of "big data," rating agencies may turn to machine learning methods designed for larger data sets to build newer models.[47]

The public data this article uses consists of unsecured loans originated and serviced by Lending Club. Lending Club makes unsecured personal loans between $1,000 and $40,000 and makes loan performance and reject data publicly available.[48]

Lending Club securitizes most of its portfolio and thus its data is useful in building a model to rate an ABS of unsecured personal loans. The specific Lending Club dataset used is a common subset of Lending Club's accepted loans used in instructional courses on machine learning.[49]

The basic paradigm in supervised machine learning is to train a predictive model based on a data set in which the target variable is known. Once identified, the data set is split into a training subset—on which the model is built—and a test or holdout subset—on which the model is tested for accuracy. As background, the data set consists of 466,257 loans originated and serviced by Lending Club between 2009 and 2014. The entire data set includes active loans, which are still performing, and inactive loans, which are no longer performing due to either full performance or nonperformance. Nonperforming inactive loans are categorized as "bad loans" because these loans have generally defaulted or been charged off.

The goal is to build a model that will predict the number of defaults in a single tranche ABS of unsecured personal loans. Such a model needs to train on data with completed performance—that is, loans for which the outcome is

[46] The data files I used to build these models are available online at http://static.turi.com/datasets/lending_club/loanStats.csv.

[47] In fact, there is an emerging literature on the use of machine learning methods in credit-risk analysis. *See* Amir Khandani, Adlar Kim & Andrew Lo, *Consumer Credit-Risk Models Via Machine-Learning Algorithms*, 34 J. BANKING & FIN. 2767 (2010).

[48] For the complete set of Lending Club's publicly available data, visit *Lending Club Statistics*, LENDINGCLUB, http://www.lendingclub.com/info/download-data.action (last visited Sept. 14, 2018).

[49] Andrew Bruce, *Prediction of Loan Default with Classification Model*, TURI (May 27, 2015), http://turi.com/learn/gallery/notebooks/predict-loan-default.html. For the specific code used for this article, see GITHUB, http://github.com/vamarishnu/Rating-Paper (last visited Sept. 14, 2018).

known—so, first, the data will be split into inactive and active loans. The active loans will serve as the portfolio in the hypothetical single tranche ABS to be rated with the predictive model. Below is a portfolio-level summary of the active loan data:

ACTIVE LOANS
Number of loans = 343,680
Average loan amount = $14,855
Average term = 43 months
Average interest rate = 13%

The inactive loan data will be used to build a model to predict number of defaults in the hypothetical ABS. Below is a summary of the basic characteristics of the inactive loan data:

INACTIVE LOANS
Number of loans = 122,607
Number of defaulted loans = 23,150
Average loan amount = $12,809
Average term = 40 months
Average interest rate = 13%

The next two sections walk the reader through the creation of this hypothetical ABS model. In order to simplify the analysis, assume the following: (1) assets in the pool perform independently and there is no correlation between their default risk; (2) to assign a rating we only need to measure expected defaults in the pool as opposed to more complex outputs such as expected loss and probability of default; and (3) lower expected defaults generally equals a higher rating.[50] These simplifying assumptions reduce rating analyst degrees of freedom and help us explore how analysts can exploit two degrees of freedom as well as data and model selection in rating analysis.

1. Freedom in Data Selection

Following the basic machine learning paradigm, the inactive loan data must first be split into training and test subsets. An initial degree of freedom that rating analysts have in building a predictive model is deciding what data to use to build the model. To illustrate the impact that different underlying data can have

[50] These assumptions are meant to simplify the analysis in this section and not to suggest these design decisions are wise. Indeed, one should definitely not assume assets in the pool perform independently as inaccurate correlation assumptions are widely regarded as one of the fundamental mistakes rating agencies made in rating RMBS and CDOs. *See* Partnoy, *Gatekeepers*, *supra* note 9, at 78; Felix Salmon, *Recipe For Disaster: The Formula That Killed Wall Street*, WIRED (Feb. 23, 2009), http://www.wired.com/2009/02/wp-quant/ (discussing the pervasive use of the Gaussian Copula for estimating default correlations in structured finance prior to the financial crisis).

on predictions, we will iterate through different training/test ratios in order to build the ABS model.[51] To isolate the impact of modifying the training and test subsets, analysis in this section will be limited to a single model logistic regression, commonly used to predict a binary outcome from a set of independent variables. It is common to use 80% of a data set to train a model and 20% of the data set to test the accuracy of the model (an 80/20 ratio). As Figure 1 shows, changing the amount of data the model is trained on by using different splitting ratios can impact the number of predicted defaults:

FIGURE 1: Defaults Predicted for Active Loan Portfolio

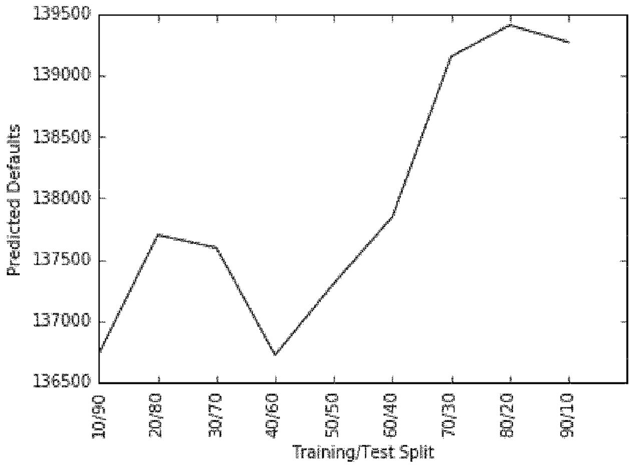


Figure 1 shows the number of defaults predicted for the portfolio of active loans from a logistic regression model trained on the inactive loan data. The x-axis reflects the partitioning ratio between training and test subsets as a percentage of the inactive loan data set. Figure 1 additionally shows that the rather unconventional training/test ratio of 40/60—as opposed to the traditional ratio of 80/20—predicts the fewest number of defaults. The most important question, however, is whether a 40/60 training/test ratio is defensible. Perhaps unsurprisingly, the answer requires us to precisely define what we want the model to accomplish.

There are different metrics we can use to examine the validity of a model that predicts a binary or discrete outcome. These metrics include accuracy (the percentage of defaults accurately predicted in the test data), precision (ratio of true positives over true positive plus false positives), and recall (ratio of true

[51] One problem with this approach is the accuracy of the model will be tested against different test sets, so we will not be making an "apples to apples" comparison in evaluating the accuracy of models built on different data sets. The purpose of this exercise is to illustrate how a rating analyst can exploit freedom in data selection to build inaccurate models. The consequences of poor design choices—such as this one—are explored in some detail *infra* Part IV.

positive over true positives plus false negatives). The accuracy metric gives us a general impression of our model's accuracy, but the accuracy score is agnostic about the directionality of errors. The metrics of precision and recall provide measures of the two kinds of errors available: false positives and false negatives. Because the model is trained to predict defaults, a false positive is inaccurately predicting a loan will default when in actuality the loan did not default. A high precision score means fewer false positives. By contrast, a false negative indicates that we incorrectly predicted a loan will not default and in actuality the loan did default. False negatives are measured by recall. A risk averse rating agency would likely worry much more about false negatives (failing to predict a loan will default) than false positives (inaccurately predicting a loan will default). Table 1 shows the accuracy, precision, and recall scores for various splits of the training and test data and predicted defaults for our hypothetical ABS:

TABLE 1: Active Loan Portfolio Training/Test Split Ratios and Corresponding Default Prediction Rates

Split	Accuracy	Precision	Recall	Predicted Defaults
40/60	0.663726	0.305885	0.609709	136,721
10/90	0.666398	0.307349	0.606571	136,737
50/50	0.661588	0.303113	0.612761	137,304
30/60	0.662802	0.305335	0.614944	137,595
20/80	0.664967	0.306568	0.609735	137,700
60/40	0.660441	0.303024	0.615152	137,846
70/30	0.657206	0.302143	0.616542	139,152
90/10	0.656798	0.298275	0.619214	139,270
80/20	0.660142	0.304697	0.620422	139,408

As Table 1 shows, a 40/60 training/test split predicts few defaults, although it does have a lower recall score than a traditional 80/20 split. On the other hand, the 40/60 split has a higher accuracy score, although the differences in error measured across various training/test splits are relatively small. In essence, what this demonstrates is that a rating analyst could iterate through various training and test combinations to find a split that leads the model to predict fewer defaults.

2. Freedom in Model Selection

Another degree of freedom a rating analyst has in developing a predictive model is their choice of model. As we will see, model choice can have a significant impact on the number of predicted defaults. Here, three models will be examined: (1) logistic regression (from above); (2) a decision tree classifier, which infers decision rules from the training data;[52] and (3) a nearest neighbor classifier, which predicts whether a loan will default by comparing the features of the loan to its closest neighbor (based on mathematical similarity). Table 2 shows the predicted number of defaults for the hypothetical ABS as well as the error metrics for each model trained and tested with a 40/60 split between training and test data from the prior section:

TABLE 2: Comparison of Predicted Defaults and Error Metrics for Various Predictive Models

Model	Accuracy	Precision	Recall	Predicted Defaults
Logistic Regression	0.663726	0.305885	0.609709	136,721
Decision Tree	0.660490	0.304140	0.614082	139,270
Nearest Neighbor	0.724964	0.261549	0.247239	63,885

Running the same data through each of these models indicates that the nearest neighbor classifier predicts the fewest defaults in our hypothetical ABS. The nearest neighbor model also has the virtue of a high accuracy, although it has very poor recall. By contrast, logistic regression predicts more than twice as many defaults than the nearest neighbor model, has much better recall, and thus a much lower false negative risk (which, as a reminder, in this context means a much lower risk of incorrectly identifying a bad loan as a good one).

The examples above demonstrate how two analyst degrees of freedom can impact predicted defaults and the riskiness of a predictive model. It should come as little surprise to readers with some background in quantitative methods that inputs and model choice matter for predicting outcomes. But these are real choices rating analysts face in model design. Indeed, a rating analyst is often confronted with many more degrees of freedom in model design, including correlation risk, cash flow analysis, and recovery expectation.

[52] 1.10 *Decision Trees*, SCIKIT-LEARN, http://scikit-learn.org/stable/modules/tree.html. Note the library I am using for the decision tree is Scikit-learn's decision tree classifier. There is some variance in the output of the decision tree each time it is trained and the results of the decision tree may not mirror the results in Table 2 if the code available on Github, *supra* note 46, is downloaded and independently run.

C. Evidence of Subjective Bias and a Taxonomy of Model Manipulation

The data and model selection scenarios explored here in Part III may seem somewhat contrived, but they track the decisions rating agencies made in designing models to rate RMBS and CDOs prior to the 2008 financial crisis. Empirical studies by John Griffin and Dragon Yongjun Tang suggest one major credit rating agency was biased towards finding fewer defaults in rating pre-crisis CDOs and that the same rating agency made subjective adjustments in certain CDO deals to increase AAA tranche sizes.[53] A separate empirical study by the same authors shows that dual-rated CDOs (i.e., CDOs rated by both Moody's and Standard & Poor's) performed worse than CDOs rated by either Moody's or Standard & Poor's ("S&P"), which suggests that Moody's and S&P, as rating agencies, succumbed to competitive pressure and "relaxed" their criteria in dual-rated deals.[54] Though the authors of these studies suggest business interests may have influenced rating agency decision making, the authors are ultimately agnostic about rating agencies' true motives.

A far more sinister explanation of rating agency motives can be found in the allegations of complaints filed by various law enforcement agencies against the major rating agencies. For example, the Department of Justice ("DOJ") sued S&P[55] in 2013 for violating the Financial Institutions Reform, Recovery, and Enforcement Act of 1989 ("FIRREA").[56] The Department of Justice alleged that S&P falsely represented that they were not motivated by business interests in developing its RMBS and CDO models. In particular, DOJ alleged that: (1) S&P built its RMBS model on a small dataset of "first-lien, fixed rate, prime mortgage loans,"[57] and delayed updating its RMBS to include a larger data set of risky mortgages that more accurately reflected the contents of the securities it was then rating; and (2) in updating the default assumptions for the monte-carlo model S&P then used to rate CDOs, an S&P executive cherry-picked the default probabilities from two competing analytic proposals in order to have competitive ratings in different CDO sectors.[58]

DOJ's Complaint even included allegations that S&P employees attempted to use NHST to expressly "fish" for business-friendly assumptions:

[53] *See* John Griffin & Dragon Yongjun Tang, *Did Credit Rating Agencies Make Unbiased Assumptions on CDOs?*, 101 AM. ECON. REV. 125 (2011); John Griffin & Dragon Yongjun Tang, *Did Subjectivity Play a Role in CDO Credit Ratings?*, 67 J. FIN. 1293, 1295-96 (2012).

[54] John Griffin, Jordan Nickerson & Dragon Yongjun Tang, *Rating Shopping or Catering? An Examination of the Response to Competitive Pressure for CDO Credit Ratings*, 26 REV. FIN. STUD. 2270, 2272-75 (2013).

[55] *See* Complaint at ¶¶ 133-86, United States v. The McGraw-Hill Cos., Inc., No. CV 13-00779 (C.D. Cal. Feb. 4, 2013), 2013 WL 416293 [hereinafter DOJ Complaint].

[56] Financial Institutions Reform, Recovery, and Enforcement Act, Pub. L. No. 101-73, 103 Stat. 183 (1989).

[57] DOJ Complaint, *supra* note 52, at ¶¶ 133 - 57.

[58] *See id.* at ¶¶ 160-79.

> "The Old Way," characterized as a "One Way Street," worked as follows: "To come up with PDs [Probabilities of Default] and asset correlations in [S&P's CDO model], we look at our raw data and come up with a statistical best fit. When this does not meet our business needs, we have to change our parameters ex-post to accommodate." The presentation added: "Does this work [for] our rating business? If it does not, need to tweak PDs."
>
> The "New Way," characterized as a "Two Way Street," worked as follows: S&P "came up with a new methodology emphasizing flexibility. We decide on a number of business friendly [sic] PD matrices first." Then S&P used hypothesis testing to determine whether the business friendly [sic] matrices were "reasonable."[59]

The examples in DOJ's complaint go well beyond the researcher degrees of freedom and fishing for statistical significance found in studies by behavioral psychologists, and suggest S&P misrepresented the basis for its methodological choices. Yet as Gelman and Loken persuasively argue,[60] bad faith or explicit fraud is not necessary for statistical modeling to be biased. One way to frame the choices rating analysts face, and their potential motives, is with Gelman and Loken's taxonomy of test procedures[61] modified to apply to model manipulation by rating analysts:

> 1) Let Y be the output from a predictive model to rate debt. Y can be multivariate and can measure defaults, probability of default, expected loss or any other dependent variable a rating agency seeks to predict.
>
> 2) Let φ represent the features of the predictive model, which includes model type (e.g., parametric or nonparametric regression, monte carlo), input data, independent variables, correlation risk, and error terms.
>
> 3) Rating analyst degrees of freedom without "fishing": $Y \leftarrow \varphi$ where a different φ would have been developed given different data; thus $Y \leftarrow \varphi$ where φ is used in actual model development.
>
> 4) "Fishing": computing $Y \leftarrow \varphi_j$ for $j=1, \ldots, J$: that is, iterating across J different φ and choosing φ_j that minimizes defaults.
>
> 5) "Fraud": pursuing the methods described above in #3 or #4 or simply choosing an arbitrary φ lying about the basis for the decision.

The distinction between the procedures outlined in options four and five above is subtle and turns on a rating agency's justification for its methodological choices. For example, if a rating agency represented the agency iterated across different φ and settled on the φ that minimized predicted defaults based on

[59] *See id.* at ¶¶ 190-91.
[60] *See generally* Gelman & Loken, *Garden of Forking Paths*, *supra* note 4.
[61] *Id.* at 2.

quantitative and qualitative judgments, the rating agency would be guilty of fishing but not necessarily fraud. However, in terms of regulating to protect against the sort of market upheaval that we saw in 2008 and onward, the difference between methods four and five may not even matter. We want rules that discourage a rating agency from abusing analyst degrees of freedom or fishing for models that under-predict defaults even if the agency believes the choices it is making are correct. Part IV examines the extent to which the reforms in Dodd-Frank are effective at preventing the model manipulation discussed in this Part.

IV. DODD-FRANK'S RATING AGENCY REFORMS

The rating agency reforms in Dodd-Frank were designed to change the incentive structure in the ratings industry. Dodd-Frank sought to accomplish this by subjecting rating agencies to increased regulatory oversight, mandating that rating agencies make extensive disclosures about their methodologies, ratings performance, and ratings action, and eliminating certain references to credit ratings in federal law.[62] Part IV examines whether these oversight and disclosure reforms are effective at curbing subjective bias in model design and the extent to which subjective bias is still a problem in the ratings industry. I conclude that while Dodd-Frank imposes substantial costs on rating agencies for pursuing quantitative strategies in bad faith, Dodd-Frank—as presently applied—has little application when rating agencies are not unambiguously operating in bad faith.

A. Effectiveness of Increased Oversight

Dodd-Frank increased oversight over NRSRO's in two ways: (1) by requiring that NRSRO's maintain specific internal policies and modify their internal organization to minimize the impact of business interests on credit rating analysis; and (2) by creating a new division within the SEC, the Office of Credit Ratings ("OCR"), to ensure that NRSRO's maintain appropriate separation between business and credit analysis. Dodd-Frank's internal policy and organizational mandates include the requirement that an NRSRO "establish effective internal control structure governing implementation of and adherence to policies, procedures, and methodologies for determining credit ratings."[63] Such internal controls must include:

- Controls reasonably designed to ensure that newly developed methodology or proposed update to an in-use methodology is subject to an appropriate review process (for example, by persons

[62] *See* Dodd-Frank, Title IX, Subtitle C, §§ 931-939H (codified in scattered parts of the U.S. Code).

[63] 15 U.S.C. § 78o-7(c)(3)(A).

who are independent from the persons that developed the methodology or methodology update) and to management approval prior to the new or updated methodology being employed by the [NRSRO] to determine ratings.[64]

- Controls reasonably designed to ensure that newly developed quantitative models proposed to be incorporated into a credit rating methodology are evaluated and validated prior to being put into use.[65]
- A prohibition against an NRSRO allowing a person who participates in developing or approving procedures or methodologies used for determining the credit rating, including qualitative and quantitative models to participate in the sales or marketing of a product or service of the NRSRO or a product or service of an affiliate of the NRSRO,[66] or to be influenced by sales or marketing considerations.[67]
- The requirement that each NRSRO designate a compliance officer to ensure conflicts rules are adhered to and submit annual reports to the SEC.[68]
- The requirement that each NRSRO have an independent board to establish maintenance and enforcement of policies and procedures related to determining credit ratings.[69]

Dodd-Frank created OCR to regulate NRSRO's and ensure that NRSRO's comply with the above.[70] The OCR's primary enforcement power is its ability to revoke NRSRO status. Pursuant to Dodd-Frank, the OCR can revoke a rating agency's license for a particular class of securities if the rating agency fails "to consistently produce ratings with integrity."[71] In determining whether a rating agency has failed to produce ratings with integrity for a particular class of securities, Dodd-Frank allows OCR to consider the accuracy of an agency's ratings with respect to a class of securities.[72] Moreover, recent SEC regulations expanded OCR revocation authority to allow OCR to revoke a rating agency's license for failing to separate sales activities from marketing activities.[73]

To examine the effectiveness of increased oversight, it is helpful to consider the application of these rules in a hypothetical OCR investigation. Assume two different rating agencies implemented the ABS model described in Part II and that OCR is investigating these agencies for compliance with Dodd-

[64] 17 C.F.R. § 240.17g-8(d)(1)(i).
[65] 17 C.F.R. § 240.17g-8(d)(1)(v).
[66] 17 C.F.R. § 240.17g-5(c)(6), 5(c)(8)(i).
[67] 17 C.F.R. § 240.17g-5(c)(8)(ii).
[68] 15 U.S.C. § 78o-7(j).
[69] 15 U.S.C. § 78o-7(t).
[70] 15 U.S.C. § 78o-7(p)(1)(A)(i)-(iii).
[71] 15 U.S.C. § 78o-7(d)(2)(A).
[72] 15 U.S.C. § 78o-7(d)(2)(B).
[73] 17 C.F.R. § 240.17g-5(g).

Frank. Assume also that the primary evidence OCR is relying on to determine compliance with the rules are the following internal emails[74] from each rating agency:

> RATING AGENCY #1
> *To: executive@ratingagency1.com*
> *From: quant@ratingagency1.com*
> *Subject: Concerned about Training/Test Split and Model Selection*
>
> I am deeply concerned about the direction we're headed with the update to our beta ABS model. Iterating across training/test splits to minimize projected defaults is completely indefensible. Moreover, the accuracy of the model is measured against different test sets, so we are not comparing apples to apples in evaluating the accuracy of models trained on different data. In addition, though the nearest neighbor model has the virtue of being accurate against our test data, nearest neighbors has a tendency to be over-fit to the training data. I really worry about the risk we may be letting into the market. We should not let business interests prevent us from making defensible and accurate predictions.
>
> RATING AGENCY #2
> *To: executive@ratingagency2.com*
> *From: quant@ratingagency2.com*
> *Subject: Progress on our ABS beta*
>
> We are making strong progress on our ABS beta. We have tested it using various training/test splits. Though 80/20 is a common training/test ratio, given the unique data set we are working with, it is not obvious that 80/20 is appropriate. Also, given the non-linearity of the data, we think a nonparametric classifier such as nearest neighbor makes the most sense. We continue to test but believe the ABS beta will prove to be a useful model in our analysis.

The first email reflects a more explicit bias, suggesting Rating Agency 1 violated the terms of Dodd-Frank in developing its ABS model. First, the quantitative analyst at Rating Agency 1 expressly identifies business interests as affecting the agency's judgment ("[w]e should not let business interests prevent us from making defensible and accurate predictions"). Thus, it would likely run afoul of Dodd-Frank's firm prohibition against allowing quantitative analysts to consider business interests, giving the director of OCR a clear basis to revoke the agency's license to rate ABS. Second, the email may hint at a lack of appropriate internal controls, board independence, and appropriate oversight by the agency's designated compliance officer, which would warrant further investigation to

[74] In the applied epistemology of regulatory investigations, few pieces of second order evidence carry more empirical weight than emails.

determine if ratings issued under prior ABS models by the agency lacked integrity sufficient to warrant revocation of the agency's license.

The email from Rating Agency 2, however, is more ambiguous. The analyst does not seem to be motivated by the agency's business interests but rather the analyst's self-interest in delivering a product that is useful to the business. In this sense, the analyst at Rating Agency 2 is like Beall and Tracy, the researchers behind the menstrual cycle study discussed in Part II. Beall and Tracy did not appear to be fishing for significance nor did they appear to be obviously motivated by bad faith. Rather, as Gelman and Loken point out, Beall and Tracy made a series of choices in data collection and hypothesis formulation which increased the likelihood of finding a statistically significant result.[75] The analyst at Rating Agency 2 may not be explicitly iterating across various training and test splits or choosing nearest neighbors to minimize reported defaults. These choices may appear deterministic to the ratings analyst. However, looking at the entire "garden of forking paths" reveals that these choices may simply be the product of subjective bias and, in effect, increase the likelihood the agency's ABS model underestimates default risk.

Though the email from Rating Agency 2 does not appear to directly implicate Dodd-Frank's separation of sales and marketing activities, it may suggest a failure to have appropriate internal controls, which, in turn, could reflect the rating agency's failure to have a truly independent board. The rules themselves, however, do not provide clear guidance on this front. The regulations implementing Dodd-Frank's internal control provisions require appropriate review of proposed quantitative models and the proper vetting and evaluation of active quantitative models. Using a peculiar training/test ratio and a nonparametric model with a low recall score may suggest, in particular, poor internal controls. On the other hand, as noted in Part III, these data and model selection choices are relatively accurate with respect to overall errors even though they are weak with respect to Type I errors as they have low recall scores. The Dodd-Frank rules themselves do not provide clear guidance on what exactly constitutes appropriate vetting and evaluation in this context.

Moreover, it is not immediately clear that these design decisions would result in ratings with less "integrity." One could imagine in creating a model that predicts few defaults, an analyst motivated by both self-interest while at the same time concerned about making a model that significantly deviates from the agency's prior and extant ratings. As noted in Part III, empirical studies have shown that credit ratings migrate in the same direction in certain situations and are biased towards the same kinds of errors.[76] In this sense, it is hard to see how the director of OCR could conclude Rating Agency 2's model lacks integrity if it is within a reasonable margin of error of the predictions of other agencies.

As I explain in Part V, Dodd-Frank gives the OCR flexibility in determining

[75] Gelman & Loken, *Garden of Forking Paths*, *supra* note 4, at 9.

[76] *See* Griffin, *supra* note 50.

when ratings lack integrity and the OCR could issue interpretive guidance with respect to Dodd-Frank's revocation provisions and internal control provisions to give these provisions more teeth. In the absence of meaningful guidance, the increased oversight mandated by Dodd-Frank may be too weak to prevent the soft but no less dangerous bias in data selection and model choice that the email from Rating Agency 2 suggests.

B. Effectiveness of Mandated Disclosures

Dodd-Frank mandates three important kinds of disclosure by rating agencies: (1) disclosure of credit rating methodologies; (2) disclosures associated with rating actions; and (3) disclosure of ratings performance.[77] Provisions targeting transparency in rating methodology require both qualitative and quantitative disclosures of credit rating methodologies. The qualitative disclosures must include:

- The main assumptions and principles about correlation of default across underlying assets;[78]
- Potential limitations of the credit ratings, and the types of risks excluded from the credit ratings that the rating agency does not comment on;[79]
- Information on the uncertainty of the credit rating, including: (1) information on the reliability, accuracy, and quality of the data relied on in determining the credit rating; and (2) a statement relating to the extent to which data essential to the determination of the credit rating were reliable or limited, including: any limits on the scope of historical data; and any limits in accessibility to certain documents or other types of information.[80]

Additionally, the quantitative component requires that rating agencies disclose information on the sensitivity of the rating to assumptions made by the rating agency, including:

- Five assumptions made in the ratings process that, without accounting for any other factor, would have the greatest impact on a rating if the assumptions were proven false or inaccurate;[81] and
- An analysis, using specific examples, of how each of the five assumptions identified impacts a rating.[82]

[77] *See* 15 U.S.C. § 78o-7(s)(3).
[78] 15 U.S.C. § 78o-7(s)(3)(A)(ii).
[79] 15 U.S.C. § 78o-7(s)(3)(A)(iii).
[80] 15 U.S.C. § 78o-7(s)(3)(A)(iv).
[81] 15 U.S.C. § 78o-7(s)(3)(B)(iii)(I).
[82] 15 U.S.C. § 78o-7(s)(3)(B)(iii)(II).

Provisions concerning disclosure associated with rating actions require that rating agencies disclose extensive information about the methodologies used to rate a security, the quantitative and qualitative assumptions made, and the reasons for the rating.[83] In addition, SEC regulations promulgated under Dodd-Frank require the following with each rating action:

> Attestation: The [NRSRO] must attach to the form a signed statement by a person within the [NRSRO] stating that the person has responsibility for the rating action and, to the best knowledge of the person:
>
> (A) No part of the credit rating was influenced by any business activities;
>
> (B) The credit rating was based solely upon the merits of the obligor, security or money market instrument being rated; and
>
> (C) the credit rating was an independent evaluation of the credit risk of obligor, security or money market instrument.[84]

Though the disclosure provisions are not expressly enforceable, Dodd-Frank's passive reforms may expose rating agencies to unfair and deceptive acts and practices ("UDAP") claims by state law enforcement to the extent a rating agency makes a false or misleading disclosure.[85] State attorneys general relied on such claims to sue NRSROs for alleged misconduct in rating RMBS and CDOs prior to the financial crisis.[86] In lawsuits filed against S&P and Moody's, various state attorneys general alleged that S&P and Moody's represented to the public that they were independent and objective in their analysis of RMBS and CDOs, when in fact they were not.[87] Disclosures of methodologies and the attestation required with each credit rating provide a clearer link between a deceptive statement and misconduct than the general statement of independence

[83] 17 C.F.R. § 240.17g-7.

[84] 17 C.F.R. § 240.17g-7(a)(1)(iii).

[85] *See generally* Mark Totten, *Credit Reform and the States: The Vital Role of Attorneys-General After Dodd-Frank*, 99 IOWA L. REV. 115, 119-21 (2013) (discussing the evolution of consumer protection laws, including the enactment of state UDAP laws).

[86] *See generally In re Standard & Poor's Rating Agency Litig.*, 23 F. Supp. 3d 378, 395 (S.D.N.Y. 2014) (summarizing UDAP claims brought against S&P and Moody's by state attorneys general in the context of discussing federal jurisdiction over consolidated state law claims). The author was involved in one such lawsuit, *Illinois v. The McGraw-Hill Cos., Inc.*, No. 12 CH 02535 (Cir. Ct. Cook Cnty. Jan. 25, 2012), as an assistant attorney general for the Illinois Attorney General. Though this article draws from the author's experience in this litigation and a similar investigation into the practices of Moody's, this article does not disclose any confidential or non-public information gleaned from the author's participation. See Press Release, Illinois Att'y Gen., Madigan, DOJ & States Reach Settlement with Moody's for Misleading Investors in Lead Up to Economic Collapse (Jan. 13, 2017), http://www.illinoisattorneygeneral.gov/pressroom/2017_01/20170113b.html.

[87] *See In re Standard & Poor's Rating Agency Litig.*, 23 F. Supp. 3d at 395.

and objectivity state attorneys general relied on, thus potentially exposing a rating agency to broader UDAP liability than before Dodd-Frank. Moreover, state law enforcement officials have broad pre-discovery investigative powers and would likely look to second-order evidence—such as internal emails—to determine if rating agencies misled the public in disclosures about their methodologies and ratings of securities.

As with Dodd-Frank's oversight reforms, it is useful to examine the effectiveness of mandated disclosures with a hypothetical law enforcement investigation. Using the example from the oversight discussion above, assume a particular state attorney general is investigating the same two rating agencies to determine if these agencies violated state UDAP in creating the ABS model from Part III. Assume as well that the state attorney general is relying on the same emails discussed in the previous section.

Turning to our two emails, we again see that the passive reforms in Dodd-Frank are better suited to curb the explicit fraud suggested in the email from Rating Agency 1 and not the implicit bias contained in the email from Rating Agency 2. The rating analyst's express acknowledgement in the Rating Agency 1 email that business interests affected model development would make it difficult for a rating agency to make forthcoming public disclosures about the qualitative and quantitative assumptions in their business models. Honest disclosures would almost certainly leave the rating agency vulnerable to revocation for failing to appropriately separate sales and marketing interests. Dishonest disclosure, on the other hand, would expose the rating agency to state UDAP liability for making false and misleading public statements. Though the Rating Agency 1 email concerns development of a quantitative model, one could imagine a similar email related to the rating of a security. In such a case, to the extent the rating agency was attesting that its rating was free from the consideration of business interests, the rating agency would be exposed, based on the contents of the email, to clear liability under state UDAP laws.

The Rating Agency 2 email, on the other hand, again proves to be more problematic. As this email appears to be the product of an honest but biased analytic approach, a rating agency could be forthcoming about quantitative and qualitative assumptions in its methodology without obviously running afoul of state UDAP laws. Gelman and Loken's discussion of Beall and Tracy is again useful here.[88] Beall and Tracy did not perceive themselves as fishing for statistical significance and disclosed vulnerabilities of their study. Similarly, a rating agency that discloses its justification for using a curious training/test split and a model that predicts few defaults is not necessarily being deceptive if it is truly the product of a biased but honest intellectual inquiry. Moreover, if the Rating Agency 2 email was related to the rating of a security as opposed to model development, it seems plausible that a rating agency could truthfully attest that the associated rating was not compromised by business interests. Thus,

[88] Gelman & Loken, *Garden of Forking Paths*, *supra* note 4, at 8-9.

while the disclosure requirement may invite cleaner state UDAP claims when a rating agency misleads the public about its methodologies and analysis of individual securities, the passive reforms in Dodd-Frank are similarly ill-equipped to curb soft bias that may result in inaccurate credit ratings.

C. Is Subjective Bias Still a Problem in the Ratings Industry?

Dodd-Frank's oversight and disclosure provisions are insufficient to address the problem of subjective bias in model design. An important question, then, is whether the Dodd-Frank reforms aimed at reducing reliance on credit ratings were effective at changing the incentive structure in the ratings industry in such a way as to avoid the problematic influence of subjective biases. OCR's annual examination reports have consistently found problems with NRSRO's adherence to policies, procedures, and methodologies; management of conflicts of interest; and internal supervisory controls.[89] Moreover, recent scholarship suggests these reforms have failed and that subjective bias remains a problem in the ratings industry. For example, Frank Partnoy persuasively argues in a recent paper that rating agencies continue to build seemingly arbitrary models that likely underestimate default risk.[90] Recent empirical work bolsters Partnoy's account by showing that rating agencies continue to use indefensibly low default correlations in rating structured securities.[91]

Partnoy demonstrates plain weaknesses within the rating agencies' current methodologies, although he does not suggest (and it is not obvious from his work) that these weaknesses are the product of bad faith. Take for example Partnoy's criticism of Standard & Poor's updated diversification criteria, which "changed in a way to favor diversified conglomerate firms—and that disfavored undiversified firms."[92] Partnoy notes that Standard & Poor's appears to have taken the issue of diversification seriously after the financial crisis, but contends that Standard & Poor's updated criteria rests on amorphous distinctions and misinterprets the relevant literature on the value of diversification.[93] It is possible that Standard & Poor's analytic justifications for its updated criteria are illusory and Standard & Poor's was motivated by its interest to appease large conglomerates. Yet it is equally plausible that Standard & Poor's updated criteria was the product of well-meaning but misguided analysis. As an example, assume that Standard & Poor's updated its diversification criteria based on its truly held beliefs that diversification is strongly linked with low default risk.

[89] OCR's summary examination reports for each year since 2011 are available on the SEC's website. See *Reports and Studies*, U.S. SEC. & EXCH. COMM'N (Dec. 1, 2016), http://www.sec.gov/ocr/ocr-reports-and-studies.html.
[90] *See* Partnoy, *Wrong*, *supra* note 13, at 1450-68.
[91] *See* John Griffin & Jordan Nickerson, *Debt Correlations in the Wake of the Financial Crisis: What are Appropriate Default Correlation for Structured Products?*, 125 J. FIN. ECON. 454 (2017).
[92] Partnoy, *Wrong*, *supra* note 13, at 1460.
[93] *See id.* at 1459-60.

New research indicates that this view is not necessarily correct with respect to conglomerates. Standard & Poor's, however, fails to appropriately update its views on diversification based on its strong priors. Standard & Poor's new criteria may be misguided, but it is certainly not the product of bad faith. My central argument, however, is that Dodd-Frank, as presently applied, does little to push agencies to eliminate the damaging impacts of subjective bias, and is only useful where second-order evidence strongly implies bad faith. It would be unlikely to protect against, or correctly assign blame after, these misguided positions.

V. SOME MODEST REFORMS

Nearly a decade after the financial crisis, there is an emerging consensus among legal scholars that the legislative reforms in Dodd-Frank are insufficient to address the problems the rating agencies pose.[94] Scholars contend that Dodd-Frank resulted in rules narrowly tailored to address past harms but inadequate to resolve the structural problems that still plague the rating industry.[95] Critics of Dodd-Frank's reforms propose a number of a different legislative changes, including banning the issuer-pays model in favor of a subscription-based model[96] and breaking up the rating agencies into smaller entities.[97]

As a simple solution to the problem of subjective bias in model design, the SEC could be allowed to audit rating agency models for design flaws. Assuming we could draft rules that define what constitutes a design flaw (and assuming the SEC would be capable of administering this standard), this solution would contravene federal law, which prohibits rules "regulat[ing] the substance of credit ratings or the procedures and methodologies by which any nationally

[94] *Id.*; Manns, *Downgrading*, *supra* note 5; Hill, *Limits*, *supra* note 13; Guo, *Credit Rating Agency Reform*, *supra* note 13.

[95] *See* Manns, *Downgrading*, *supra* note 5, at 753 (arguing that while the SEC's increased oversight may have curbed the most egregious abuses of the crisis, "[t]he most important part of [Dodd-Frank] remains the most unresolved: the SEC's mandate to design an alternative rating industry business model to address the conflicts of interest created by debt issuers' selecting and paying their rating agency gatekeepers"); Hill, *Limits*, *supra* note 13, at 133 (arguing that while Dodd-Frank's solution is not necessarily bad, Dodd-Frank failed to address "why rating agencies gave such inflated ratings to subprime securities"); Partnoy, *Wrong*, *supra* note 13, at 1408 (arguing that Dodd-Frank failed to reduce reliance on credit ratings enough to overcome the "stickiness of regulatory licenses").

[96] *See* Hill, *Limits*, *supra* note 13, at 146 ("[t]he solution [to Dodd-Frank's problems] is to get away from an 'issuer pays' model, in which those paying for ratings are the securities' sellers, and return to 'subscriber pays,' in which ratings are paid for by people buying research as to securities' quality"); *but see* Claire Hill, *Why Did The Rating Agencies Do Such A Bad Job Rating Subprime Securities*, 71 U. PITT. L. REV. 585, 586 (2010) (arguing against mono-casual explanations of the rating agencies' failures, including assigning exclusive blame to the issuer pays model, and casting doubt on many proposed solutions to reform the ratings industry).

[97] *See* Manns, *Downgrading*, *supra* note 5, at 801.

recognized statistical rating organization determines credit ratings."[98] Reform proposals such as switching to a subscription-based model or breaking up the rating agencies into smaller units get around this federal prohibition by changing the incentive structure in the ratings industry without explicitly regulating the substance of credit ratings. These proposals, however, suffer from two problems. First, each is hard to implement and would require significant legislative action. Second, while these reforms have the virtue of potentially minimizing the benefits of issuing biased ratings, neither increases the costs to rating agencies of making well-intentioned mistakes. Put differently, it is naive to assume rating agencies will build better and unbiased models absent rules that punish rating agencies for building bad models.

To reform the credit rating industry, we want rules that force rating agencies to internalize the costs of quantitative and analytical mistakes. Some reform proposals attempt to accomplish this by linking rating agencies' performance to pay. One such proposal involves compensating rating agencies with rated debt to be distributed as the debt matures in an amount equal to average discount rate of debt with same rating and maturity.[99] Another similar proposal would require the three major rating agencies to contribute a portion of their revenue to a performance bonus fund to be distributed periodically to the best performing rating agency and used to subsidize a secondary rating market of smaller players.[100]

Although these proposals each have the virtue of punishing rating agencies for underestimating default risk, each would require legislative changes that are unlikely in the current political environment.[101] I offer a simpler solution, which can be achieved within the framework of Dodd-Frank: liberalize the NRSRO application process and revoke regulatory status over particular classes of securities for the poorest performing rating agencies.

A. Decreasing Barriers to Entry

Increasing competition among rating agencies has long been a goal of proponents of rating agency reform. Yet the current market for credit rating remains highly concentrated. Since the enactment of Dodd-Frank in 2010, only

[98] 15 U.S.C. § 78o-7(c)(2).

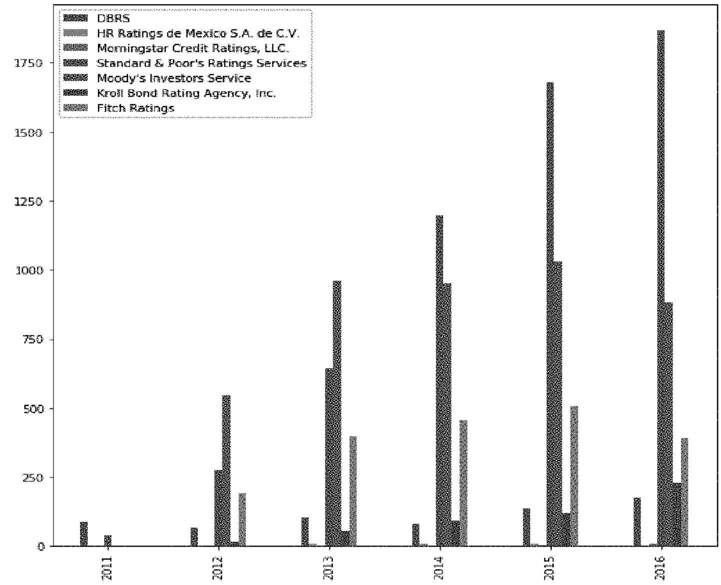
[99] *See* Listokin & Taibleson, *supra* note 8, at 94.

[100] *See* Robert Rhee, *On Duopoly and Compensation Games in the Credit Ratings Industry*, 108 NW. U. L. REV. 85, 89 (2013) (proposing "to establish a mandatory pay-for-performance compensation scheme in which a fixed percentage of accrued revenue is ceded to fund a performance bonus").

[101] One set of proposals I do not take up is the push to eliminate the federal licensing regime entirely. *See, e.g.*, Flannery, Houston & Partnoy, *supra* note 7; Gretchen Morgenson, *Should Free Markets Govern the Bond Rating Agencies?*, N.Y. TIMES, May 6, 2017, at BU1 (arguing against keeping the current federal licensing scheme for rating agencies). Though I am sympathetic to eliminating the federal licensing regime, there seems appetite to get rid of CRARA and, as such, I devote little attention to these proposals in this article.

one new firm has been granted NRSRO status, increasing the total current number of NRSROs to ten. Figure 2 shows the market concentration for ABS issuances from 2011-2016:

FIGURE 2: Asset Backed Security (ABS) Issuances (2011-2016)[102]



The small number of NRSROs and market concentration are notable for two reasons. First, capital and data costs have been low since 2010. Firms can raise equity capital relatively easily and debt financing is remarkably cheap. Not only is capital cheap, but firms can acquire information for free or at low cost from third-party data brokers and can then utilize powerful algorithms to analyze the data at relatively low cost. Low capital and data costs have led to significant growth in many industries that are heavily reliant on large capital infusions and data analysis. For example, during the same period in which the rating agency field only grew eleven percent, there has been significant growth in marketplace lending, payment processing, and other banking technologies. In this environment, one would expect new rating agencies to emerge with novel ways to analyze and rate debt offerings. Yet none have emerged, or at least none have been granted regulatory licenses by the SEC.

[102] *See* GitHub, *supra* note 46 (source code for Figure 2 is available on the GitHub page). Figure 2 shows the number of new ABS ratings issued between 2011-2016 for each applicable NRSRO. The data reflected in Figure 2 was culled from public reports NRSROs are required to make available pursuant to 17 C.F.R. § 240.17g-7 as discussed in Part III.B *infra*. The data used in Figure 2 was based on the web scraper developed by Mark Joffe and Frank Partnoy to extract and consolidate this information in a simpler format, which is discussed in their recent 2018 article. See Mark Joffe & Frank Partnoy, *Making Credit Ratings Data Publicly Available* 1-2 (San Diego Legal Studies, Paper No. 18-320, 2018), http://ssrn.com/abstract=3103974.

Second, the low number of rating agencies is not obviously a result of the Credit Rating Agency Reform Act of 2006 ("CRARA"),[103] Dodd-Frank, or their associated regulations. Neither CRARA nor Dodd-Frank require much from NRSRO applicants beyond credit ratings measurement statistics, the applicant's procedure and methodologies for determining ratings, the applicant's capacity to comply with Dodd-Frank and its associated regulations, and written certification. Section 78o-7 does provide that the SEC may deny an application if the applicant "does not have adequate financial and managerial resources to consistently produce credit ratings with integrity" and to "materially comply" with the provisions of Dodd-Frank.[104] However, it is not clear what a failure to consistently produce credit ratings with integrity actually means.

Despite the relatively straightforward application provisions in CRARA and Dodd-Frank, the current NRSRO application process is opaque and firms have experienced substantial difficulty becoming a NRSRO. For example, Egan-Jones applied for NRSRO status in 1998 but only received it in 2007.[105] Similarly, R&R Consulting, a firm founded by former structured finance experts at Moody's, has been trying to become a NRSRO since 2011.[106] R&R Consulting provides ratings to initial issuances and subsequent issues of securities "trading in the second, or resale, market, after they are issued."[107] R&R Consulting suggests that one reason for the delay is the SEC's rigid enforcement of the written certification requirement under CRARA. CRARA requires that aspiring NRSRO provide at least ten written certifications from "qualified institutional buyers,"[108] a term of art in securities regulation that generally refers to any regulated entity (as opposed to an individual) that can invest in securities.[109] Although CRARA requires ten letters from institutional buyers, CRARA does not specify the form of the letters. Anne Rutledge, one of the principals of R&R Consulting, however, suggests that in practice, the SEC is rigid about the content of the letters:

[103] Credit Rating Agency Reform Act of 2006, Pub. L. No. 109-291, 120 Stat. 1327, 1327–39 (codified at 15 U.S.C. § 78o-7 (2006)) (federal legislation creating the current licensing regime for rating agencies).

[104] 15 U.S.C. § 78o-7(a)(2)(C)(ii)(I) (2018).

[105] *See* EMILY MCCLINTOCK EKINS & MARK CALABRIA, CATO INST., POLICY ANALYSIS NO. 704: REGULATION, MARKET STRUCTURE, AND ROLE OF THE CREDIT RATING AGENCIES, at 21 (2012). Moreover, once a rating agency obtains NRSRO status they may struggle to dislodge the heavy market bias in favor of the existing rating agencies. For example, as Figure 2 shows, Kroll's Bond Rating Agency has struggled to significantly penetrate the ABS market despite obtaining NRSRO status in 2008.

[106] *See* Gretchen Morgenson, *On the Waiting List at the Debt-Rating Club*, N.Y. TIMES, Feb. 10, 2013, at BU1.

[107] *See id.*

[108] 15 U.S.C. § 78o-7(a)(2)(C)(i) (2018).

[109] *See* 17 C.F.R. § 230.144A(a) (2018) (defining qualified institutional buyer).

> Proof that you've done business with them is not enough; it says you must have letters. And [the SEC] have a suggested text for the letter. When we changed the text slightly they said it was not in conformity.[110]

The instructions to Form NRSRO available on the SEC's website suggest some additional reasons why the SEC issues so few licenses. In its application, the SEC requires NRSRO applicants to include as exhibits extensive information on the performance of rated debt, including a transition and default matrix for issued ratings,[111] and information about the analytic assumptions and validity of the applicant's procedures and methodologies for issuing ratings. These requirements suggest that the SEC gives a strong preference to applicants with a history of issuing unlicensed ratings and that the SEC makes value judgments about the integrity of an applicant's predictive models. Though the SEC is strictly forbidden from ex post evaluation of an NRSRO's procedures and methodologies, it appears to be using ex ante screening to prune methodologies the SEC deems invalid.[112] The costs of the SEC's screening (significant market concentration and making firms prone to underestimate default risk) may outweigh the benefits of the SEC's subjective rejection of methodologies it deems unreliable.[113] Thus, a relatively straightforward administrative change to decrease barriers to entry would be for the SEC to give little weight to an applicant's ratings history or specific methodologies, and grant NRSRO status to all applicants who can demonstrate a capacity to comply with the terms of CRARA, Dodd-Frank, and its associated regulations.

B. A Market for Regulatory Status

Liberalizing the NRSRO application is, however, insufficient to fix what actually ails the credit rating industry. Indeed, some scholars contend that increased competition may encourage a rapid race to the bottom, as rating agencies vie for issuer business by continually relaxing their ratings criteria. To avoid this, the SEC should use its power under Dodd-Frank to revoke the licenses of poor-performing rating agencies. The OCR can revoke a rating agency's license for a particular class of securities if the rating agency fails "to consistently produce ratings with integrity,"[114] and, in determining whether a rating agency has failed to produce ratings with integrity for a particular class of securities, OCR can consider the accuracy of an agency's ratings with respect to

[110] *See* Morgenson, *supra* note 101.

[111] *See* U.S. SEC. & EXCH. COMM'N, FORM INSTRUCTIONS: APPLICATION FOR REGISTRATION AS A NATIONALLY RECOGNIZED STATISTICAL RATING ORGANIZATION (NRSRO), at 21.

[112] There is some support in the literature that the SEC is, in fact, engaging in *ex ante* screening. See Rhee, *supra* note 95, at 96 (noting that "fearing fly-by-night rating agencies, the SEC has parsimoniously granted the NRSRO status").

[113] *But see id.* at 96, 104 (defending the duopoly in the ratings market and the few licenses the SEC issues).

[114] 15 U.S.C. § 78o-7(d)(2(A).

a class of securities.[115] My proposal is simple: the SEC should propose rules that define integrity and accuracy relative to the performance of other rating agencies and revoke an agency's license for particular class of securities if its performance is poor. There are two obvious complications with this proposal: (1) measuring performance and (2) ensuring bond markets do not collapse. I take each in turn below.

First, I suggest the OCR rank rating agencies at periodic intervals (e.g., quarterly or semi-annually) by the sum of the absolute difference between each rating agency's weighted average expected loss forecast and the weighted average actual loss for all outstanding securities of a particular class (e.g., ABS, CMBS).[116] Rating agencies with the larger sums (i.e., largest difference between weighted average expected loss and weighted average actual loss) are ranked below rating agencies with smaller sums. The OCR can then use a relegation-like system to temporarily revoke the licenses of the lowest ranked rating agency for at least one subsequent period. I suggest expected loss as opposed to rating grade to avoid rating agencies' gaming of the system by strategically downgrading poorly performing securities at intervals when they face the lowest risk of relegation. Similarly, I suggest absolute difference instead of the difference between average actual loss and expected loss to avoid allowing rating agencies to game the system by underrating securities. Finally, I believe the OCR should use weighted averages to account for the magnitude of mistakes (i.e., small deltas for large issuances should be weighed equally or more than large deltas for small issuances).[117]

Second, to avoid bond markets collapsing when one or more large NRSROs are suspended and to ensure that the new rules do not prejudice smaller firms, the OCR should require that all NRSRO's with active licenses for a class of securities provide expected loss forecasts for all issuances, regardless of which company initially rated them. To facilitate this, the SEC must amend Rule 17g-5 to allow the free flow of information between rating agencies. Rule 17g-5 currently provides that an NRSRO that rates a structured finance security and is paid by the issuer must provide sufficient information to other NRSROs such that they can rate the same debt.[118]

There are important qualifications that limit the application of Rule 17g-5. First, it only applies to NRSROs paid by issuers and does not apply to NRSROs who are compensated via subscription or any other method. Second,

[115] 15 U.S.C. § 78o-7(d)(2)(B)(i).

[116] Put more technically, my proposal calls for treating expected loss forecasts like a multivariable regression problem and ranking rating agencies based residual sum of squares to actual expected loss. This is not to suggest that we use sum of squares to the exclusion of other measures, but simply to offer one way to measure performance.

[117] To be sure, I am not wed to measuring accuracy with default performance and open to other measures scholars have suggested such as bond or swap spreads. *See, e.g.*, Flannery, Houston & Partnoy, *supra* note 7.

[118] *See* 17 C.F.R. § 240.17g-5(a)(3).

requesting NRSROs can only access the information if they have rated at least ten percent of the structured securities for which they requested information in the prior calendar year. Third, requesting NRSROs are limited to ten requests per year. The SEC should amend Rule 17g-5 to apply to all securities and allow NRSROs access to this information without the limitations above.

To be sure, the solutions above do not resolve all implementation issues. There are still open and important questions, including: how securities of the same class but with different maturities are treated; whether a rating agency has any recourse to fight suspension; what happens if all rating agencies do a poor job; and whether the SEC should wait to propose these rules until there are enough NRSROs to avoid significant disruptions to bond markets from the temporary suspension of one or more large NRSROs. I do not mean to suggest that any of these proposals is necessary or adequate to improve rating performance; instead, I offer them to illustrate how the SEC might build a system that would incentivize NRSRO's to compete over accurate ratings as opposed to fees.

VI. CONCLUSION

The explicit goal of rating agency reform was originally to curb perceived abuses by rating agencies prior to the financial crisis. Implicitly, however, the goal of rating agency reform was to create incentives for rating agencies to issue reliable ratings that help steer the efficient allocation of capital in debt markets. This article's central contention is that it is difficult to achieve better outcomes in rating analysis without first confronting subjective bias in model design. Liberalizing the NRSRO application process and strengthening the NRSRO revocation standards are changes the SEC can implement through formal rulemaking and informal guidance, which may help mute the soft and dangerous bias that likely continues to afflict the credit ratings market.