Drivers of Supervisory Capital Add-ons:

Signal versus Noise in Internal Ratings-Based Models

Andreas Beyer
European Central Bank
Kaiserstraße 29
60311 Frankfurt am Main
Germany
andreas.beyer@ecb.europa.eu

Arndt-Gerrit Kund
European Stability Mechanism
6a Circuit de la Foire Internationale
1347 Luxembourg
Luxembourg
a.kund@esm.europa.eu

First version: August 18, 2022
This version: October 5, 2024

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily correspond to those of the ECB, the SSM and the ESM.

Acknowledgments: We wish to thank Tancrède Polge for his outstanding analytical data support. We are grateful to Andrea Enria, Frank Elderson, Klaus Düllmann, Robert Lauter, Mario Quagliariello, Tobias Beck, participants of the 3rd Workshop on Banks and Financial Markets at the Paris School of Business and to participants and our discussant Elizaveta Sizova at the 2024 ECB Banking Supervision Research Conference for their valuable comments. All remaining errors are our own.

Abstract

Internal models are employed to estimate capital requirements in a risk-sensitive way. However, these models are not flawless. Generally, the usage of models suffers from imperfections, such as oversimplifications or wrong assumptions. As a result, risks may be underestimated, which is in particular troublesome, where it is used to assess solvency. In this paper we address an important gap in the literature with regard to such model risk. We find that the number of deficiencies in a model is unimportant (noise), but rather their severity matters (signal), when calculating the counterfactual RWA impact of supervisory capital add-ons. In fact, we can further differentiate the noise from the signal and point to a subset of non-compliances with the CRR that are the material drivers of model risk. Our results help improve the supervision of banks with IRB models by proposing means to make more efficient use of scarce supervisory resources. The findings are robust to possible reverse causality and distortions from outlier treatment.

JEL Classification: G21, G28, G29
Keywords: Microprudential Regulation, Margin of Conservatism
Banking Supervision, Model Risk

1 Introduction

Statistical models have become de facto ubiquitous in today's financial system. The recent transition to the expected credit loss accounting of IFRS 9 has only increased this prevalence of internal models in banks. In light of this ever growing importance, it must be ensured that the applied models are "[...] consistent, reliable and valid [...]" (see Rogers (2021) and CRE36 of Basel III). While models make risks quantifiable and hence manageable, they are no panacea. Derman points out as early as in 1996 that the mere usage of statistical models does not exclusively help manage risks. Instead, their usage also introduces a novel kind of risk: model risk. Generally speaking, model risk refers to the potential loss an institution may incur due to a decision that was made on the erroneous output of an internal model (cf. CRD Art. 3(1)(11)). More specifically, such losses could arise from the development and implementation (e.g. incorrect computer codes applied or wrong calibration of model parameters), but also from the application of models outside their intended scope (Barrieu and Scandolo, 2015). In light thereof, numerous lines of defense are employed in order to prevent model risk to a reasonable extent. Internally, banks validate their models through the internal validation function (IVF), while externally auditors and supervisors scrutinize new and existent models alike.

Despite theses efforts, the materialization of model risk prevails. While it may be perceived as a hypothetical issue in the public's eye, this view is possibly biased by the fact that banks under-report the materialization of model risk. Even if this risk is reported, its true impact is likely to be concealed by the fact that its realization rather impacts banks' profitability, and only in the utmost instances their viability (Aggarwal et al., 2015). Not surprisingly, a literature or web search that is based on public information for losses stemming from erroneous IRB-models only yields very few results including with e.g. the work of Aggarwal et al. (2015) being such an exception. However, from supervisory data we know that the actual number is larger by orders of magnitude. Nevertheless, understanding the conjunction of a higher prevalence of deficiencies in models on the one hand and their comparatively material impacts with regard to risk-weighted assets (RWA) and hence solvency ratios on the other hand has so far not received sufficiently attention in the academic literature. We fill this gap by using hand-collected data from Internal Model Inspections (IMIs) carried out by the European Central Bank (ECB) and/or the National Competent Authorities (NCAs) on its behalf. Our data set spans banks under the supervision of the Single Supervisory Mechanism (SSM) between 2014 and 2020. During this period, we identify 267 IMIs, which have culminated in a so called "limitation" (i.e. the supervisor imposing higher RWA until a non-compliance/deficiency in an internal model is remediated). The aggregated additional Common Equity Tier 1 (CET1) capital that banks have to hold in response thereto amounts to a double-digit billion Euro figure, which greatly exceeds the equivalent impact of supervisory sanctions imposed by the ECB and Pillar 2 Requirements (P2R) through

its Supervisory Review and Evaluation Process (SREP), respectively. This stark contrast in order of magnitude illustrates the importance of the topic at hand.

We find that the number of deficiencies (henceforth "findings") found in the course of IMIs is – in statistical terms – not a significant driver of the RWA impact of limitations. Instead, it is the severity of the findings that almost exclusively explains those limitations and hence the implied capital add-ons. More specifically, we can link these high severity findings and non-compliances to particular sections of the Capital Requirements Regulation (CRR). Those limitations can then also be interpreted as proxies for the unidentified model risk by the bank. Identifying the drivers of model risk therefore enables us to separate signals from noises in order to gauge the impact of model risk.

Looking back, in 2016 the ECB (2021) launched a "targeted review of internal models" (TRIM) to address inter alia, concerns regarding the complexity of internal models. The TRIM exercise was designed to ensure a level playing field by investigating the adequacy of the largest internal models in banks and harmonizing supervisory practises along this way. In our analysis, we find evidence that limitations from TRIM missions tend to have – on average – higher RWA impacts, suggesting the harmonisation of supervisory practises at a stricter but more level playing field. Our policy implications are twofold. First, our results show that banking supervision could become more risk-based as evidenced by high-severity findings being a main driver of limitations. Second, we show possible room for improvement with regard to banking supervisors' risk-sensitivity, when calibrating limitations. These findings are robust to possible reverse causality and distortions from outliers.

Our results are important for numerous stakeholders: Firstly, for banking supervisors as they might employ the drivers of model risk as early warning signals for shifting the focus of scarce supervisory resources on a particular subset of deficiencies. At the same time a targeted model inspection that is based on those early warning signals may help to facilitate a risk-based approach in the follow-up of imposed limitations after the findings of deficiencies. In doing so and recognizing that certain limitations entail a more severe risk than others, a standard "one size fits all" approach can be substituted in favour of a more tailor-made approach and thus contribute to more efficient banking supervision. Secondly, investors may benefit from more targeted supervision because banks possess private information on the quality of their models, which can have profound implications on capital ratios and hence investors' decisions whether or not to invest, and thereby exercise a disciplining effect for under-capitalized banks.

This paper is organized as follows: Section (2) provides the institutional background on the supervision of internal models by the ECB and reviews the current litera-

ture. Given the novelty of our data, we present it in more detail in Section (3), before deriving our research questions in Section (4). We present our results in the subsequent Section (5). Section (6) presents the robustness tests, before Section (7) concludes.

2 Institutional Background and Literature Review

2.1 Institutional background on internal models

Given that the topic under review in this paper represents rather a niche within the existing academic literature, this section will provide some institutional background on internal rating-based (IRB) models in banks under the supervision of the ECB.

Internal Model Inspections (IMIs) are usually triggered and requested by banks for two reasons: (i) either a bank seeks initial approval of a newly developed model or (ii) an existent model has undergone material changes. Following such a request for an IMI, the applying bank will be notified when the investigation will take place, and be asked to provide first information for a kick-off meeting that precedes a more intense on-site phase, where model experts review the model on the bank's premises. This period concludes the inspection phase, after which an Assessment Report (AR) that lists all deficiencies, ranked by severity on a scale from F1 (low) to F4 (very high), is drafted and checked for consistency. In the absence of comments, respectively incorporation thereof, the findings of the report become a legally binding decision by the ECB through approval by the Supervisory Board of the SSM. At this point in time, the inspected bank is obliged to remediate the listed deficiencies within a given period of time that is proportionate to the severity of the finding and the complexity to remediate it. Where high severity findings suggest an underestimation of the actual risk, a limitation may additionally be imposed to address the risk of the identified underestimation and hence to preserve the level playing field. In more specific terms, the limitation could – exemplary – legally oblige the bank to increase its final PD estimates by a given factor to prevent it from gaining an advantage by calculating unduly low RWAs. For our analysis, we look in particular at the RWA-equivalent of these limitations which represent a gap between the model-based estimation of the PD and its hypothetically *true* value. In technical terms this would be the difference between the "true" final PD and the best-estimate PD in Figure (1) below. In order to explore the drivers of model deficiencies in our analysis, we look at the RWA-equivalent of the limitation that is enacted to remediate this undue difference, as our variable of interest.

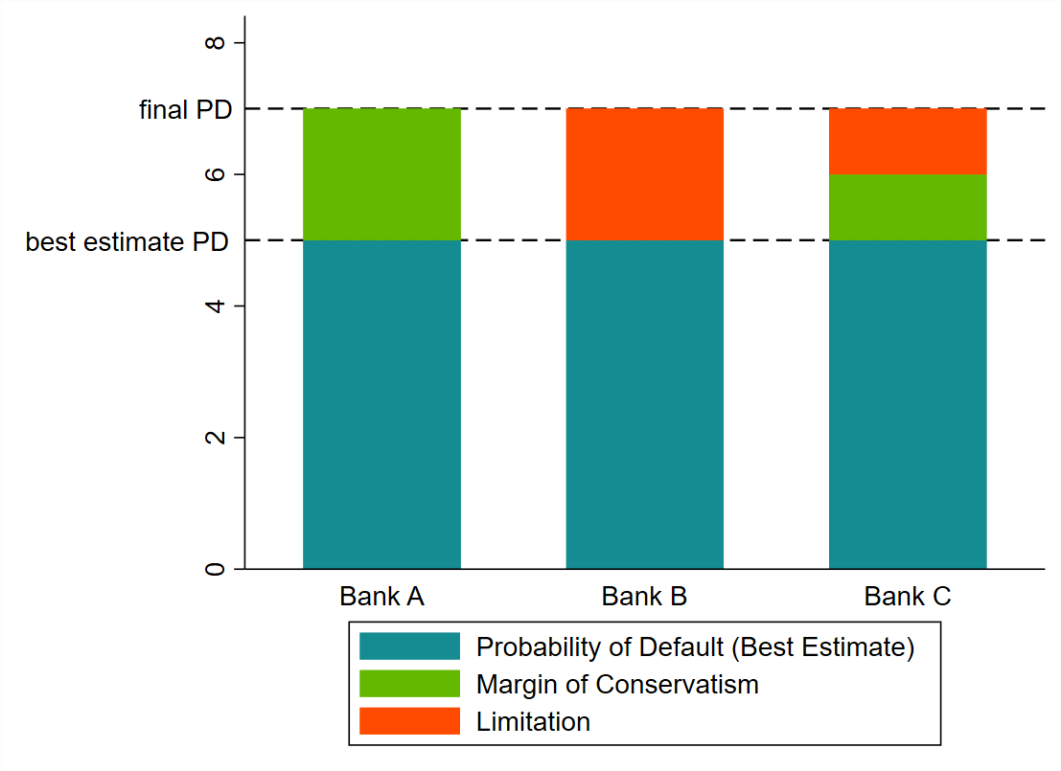
It is worthwhile mentioning that the notion of a "true" PD is associated with the fiction of a "true model". Aggarwal et al. (2015) questions this concept, arguing that differences between reality and a model's output are more reminiscent of an institution's risk appetite. Similarly, Leitner and Yilmaz (2019) suggest that the

bank never reveals this *true* model, but rather opportunistically chooses one that performs well for the given portfolio.[1]

At the same time, it is worth noting that this outcome is not the rule to an inspection. Instead, there are numerous banks that prefer to err on the side of conservatism and faithfully apply Margins of Conservatism (MoC) in order to account for potential inaccuracies in their model in the first place. To illustrate the difference, consider the hypothetical examples of Figure (1). Let us assume that all three banks are asked to estimate independently from each other the PD of an identical obligor, based on the same information set and without any constraints. Bank A estimates the best possible estimate of the PD and supplements it with a MoC to account conservatively for yet unknown potential issues that might occur and thus reaches ex post the *true* PD. While Bank B arrives at the same best estimate PD, it neglects possible deficiencies of the model, and would hence report the final PD as being the equivalent of the best estimate PD, whereby it underestimates the remaining ex post risks that together comprise the *true* entire risk and hence the *true* final PD. In order to prevent this infraction on the level playing field, the inspection team would consequentially step in, and impose a limitation, such that the *true* final PD is reached. Mixtures of this case as illustrated for Bank C can of course occur.

[1] In the applied economic and econometric literature there is often a confusion with respect to the concept of "the true model". True models exist in the context of controlled experiments such as statistical simulations. In a modelling context based on realised observed empirical data, however, all models are "wrong", see e.g. Box (1976). In the context of this paper we therefore rather interpret the "true" final PD as the its best statistical approximation of its unknown Data Generating Process, see e.g. Hendry (1995).

Figure 1: Exemplary differences in estimating the final Probability of Default across stylized banks, in %.



2.2 Literature review on internal models

With the advent of Basel II in 2007, a fundamental shift in banking regulation occurred. Since then, banks are allowed to model inter alia the credit risk of their portfolio with IRB models. Subject to compliance with numerous regulatory conditions banks thus do have to abide by the "one size fits all" approach within the Standardised Approach (STA), see Basel Committee on Banking Supervision (2013). The underlying rationale is that as banks have individually a better view of the idiosyncratic risks in their portfolio they can foster systemic stability by measuring risks more accurately through dedicated models at the institutional level. Despite these liberties, banks are not unconstrained in their modelling. Instead, features such as a minimum PD, or more recently the introduction of an output floor are intended to prevent a race to the bottom (Basel Committee on Banking Supervision, 2013).

However, despite this continuous evolution and the implementation of numerous safeguards, the usage of IRB models is "no free lunch". Derman (1996) was one of the first to point out that the introduction of models to measure risk carries a risk of its own: model risk. In this line, Aggarwal et al. (2015) argue that any model will eventually be subject to model risk, as it has to either make simplifications, given the inherent complexity of the modelled relationships; or – in the absence of

historical observations for very high percentiles – appropriate assumptions have to be made. Similarly, Cont (2006) points to the risk of wrongly parametrizing the model and thereby introducing another form of model risk. After all, the critique of Lucas (1976) holds in this context as well: The historical links on which a model was calibrated, might fail "to hold true" at the time of the model's application due to agents' shifts in expectations and therefore a shift in the parameters of the estimated model. As a consequence, the predictions of the estimated model become unreliable. Given these unfavourable properties, numerous attempts have been made by regulators around the globe to reduce model risk. Exemplary, the FED issued in 2011 the "Supervisory Guidance on Model Risk Management", thereby setting minimum standards for, inter alia, the model development and validation. Similarly, the ECB has published its "Guide to Internal Models", which gives granular guidance on topics such as the appropriate counting of defaults. Given that the majority of IRB models is calibrated against the number of observed defaults, this apparently minuscule provision has substantial real world implications. A notable tension lies in the granularity of these provisions: if on the one hand there are too many degrees of freedom, banks may opportunistically manipulate their estimates as suggested by the works of Mariathasan and Merrouche (2014) and Behn et al. (2022). If on the other hand provisions are too strict and banks would price financial instruments unanimously, the resulting lack of divergent views may constitute a financial stability risk through "group think" (Daníelsson et al., 2004). Colliard and Georg (2020) discuss a combination of both cases, where regulation becomes so complex that even a faithful bank fails to apply it correctly, demonstrating the inherent tension on the *right* level of regulation. In this light, BCBS (2013) calls for banking regulation to be generally (i) simple, (ii) comparable, and (iii) risk sensitive.

In particular, the call for models to be simple has been repeated numerous times. Exemplary, Hakenes and Schnabel (2014) argue that complexity in models can be exploited by sophisticated agents. Mariathasan and Merrouche (2014) give credence to suspicion, by showing empirically that banks tend to engage in regulatory arbitrage, where regulatory complexity is high. Barucci and Milani (2018) pin point this concern in particular to the Advanced-IRB approach, where banks are not constrained to only model the PD. Plosser and Santos (2018) demonstrate that this observation does not apply exclusively to European banks, but also holds for U.S. banks, by assessing the riskiness of the same syndicated loan across different banks. Other instances of issues in banks' internal modeling approaches are documented by Vallascas and Hagendorff (2013) who show from a market-point of view a disconnect between the implied portfolio risk and the risk inferred from IRB models. Similarly, Behn et al. (2022) show that the charged interest rate and default rate for an IRB portfolio go up, despite the estimated PD going down according to the bank's internal model. Haldane and Madouros (2012) argue along these lines that not only models should become simpler, but rather bank regulation altogether. More specif-

ically, they call for focusing on simpler metrics such as the Leverage Ratio. Barrieu and Scandolo (2015) stress in this context that the model choice itself is a substantial driver of the final risk measure. This view is also shared by Leitner and Yilmaz (2019) who show that in theory banks may only disclose one of many possible models. The chosen model produces risk-weights which are optimal, relative to its discovery costs and at the same time do not generate a negative signal for the supervisor to scrutinize the model more carefully. Nevertheless, with hindsight, results of many of these papers might be taken with some scepticism, as the data they are based on has since been superseded by yet a new regulation, that is Basel III. At the same time, there is also literature pointing to the benefits of IRB models beyond their increased risk-sensitivity. Exemplary, Bruno et al. (2023) show that more intensive usage of IRB models, and in particular the advanced IRB approach, reduces the opacity of banks due to better data quality and disclosure thereof. Against this background, we set out to investigate the status quo on internal models, and look in particular to drivers of model risk.

3 Our Data Set

Given the novelty of this non-public supervisory data set, we will describe it in more detail in this section. In doing so, two dimensions are highlighted in particular: (i) the data collection and management process that preceded our data analysis, and (ii) an illustration of the final data set.

Our analysis starts from the set of IRB banks supervised by the SSM since its inception in 2014. In order to create a balanced panel, we only look at banks that still continue to exist after the cut-off date in 2020. We deliberately choose 2020 as a cut-off to account for three effects: First, to avoid the presence of data in our sample that is based on IMI's which cannot be fully evaluated as their remediation phase has not been concluded, yet. Second, the introduction of the European Banking Authority's (EBA) Guideline on PD estimation, LGD estimation and the treatment of defaulted exposures (EBA/GL/2017/16) effective as of January 2021 and therefore possibly constituting a structural break. Third, in light of the COVID-19 pandemic numerous regulatory and governmental interventions were initiated and implemented throughout the crisis and could bias our results. For the banks meeting our sampling conditions, we proceed to manually review the so called "Supervisory Board Notes". Each of these Notes constitute an official ECB Decision, which are the final outcome of a corresponding IMI. We collect the RWA-equivalent of the imposed limitations – where available, i.e. not all IMIs conclude with a limitation – from the aforementioned Notes. Notice that we limit the scope of our sample to IRB models for calculating the credit risk RWA, in order to keep our identification strategy as clean as possible and to reduce the effect of different sets of rules for modelling different risk categories. In doing so, we leverage on the unique properties of our supervisory data set that allows us to put a quasi counterfactual price-tag on

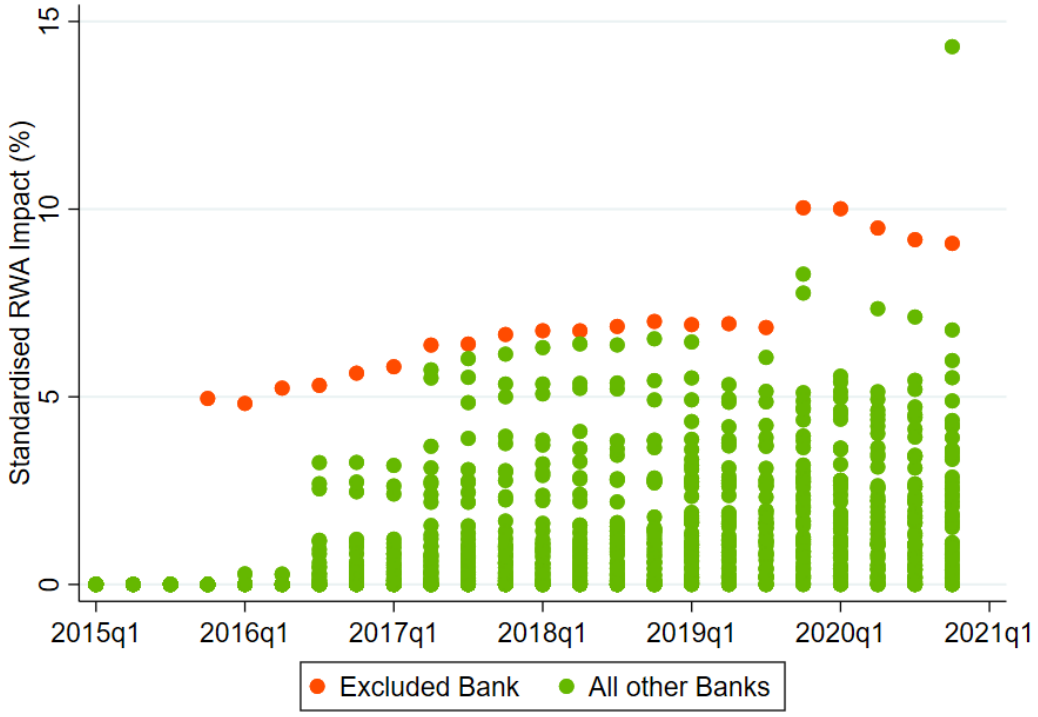
model deficiencies. This metric – standardized by total assets in order to control for larger limitations in banks with larger portfolios – will serve as the dependent variable in all subsequent analyses. One important feature of this analysis is that we isolate from the *overall* RWA impact induced by the model change the *individual* RWA resulting from the imposed *supervisory* limitation. As such, our dependent variable has a strict lower bound at zero, as a limitation cannot have a negative RWA impact. This feature would not hold, if we were to looking just at the aggregate *overall* RWA impact of any model change.

We then proceed by looking at the "Assessment Reports" (ARs), which contain a summary of all reported model deficiencies ("findings"), and, yet, more importantly also provide a reference to any particular non-compliance *vis-à-vis* the regulatory framework, i.e. CRR/CRD. This information along with the attached severity of the finding is available in structured form through the confidential supervisory "Agora" data base. We use this information in all subsequent models as explanatory variables, which are complemented by additional idiosyncratic control variables. A full list of those controls, including comprehensive descriptive statistics can be obtained from Table (1) in the Appendix. To ensure comparability we have chosen our set of explanatory variables to broadly correspond with those employed in other academic studies, such as e.g. Basel Committee on Banking Supervision (2023) and Mariathasan and Merrouche (2014). They find that models are responsible for up to 80 % and 77 % of the overall capital requirements, respectively. The slightly higher number in our sample (i.e. 88 %) is possibly driven by much higher minimum coverage ratios in Germany (92 %), which accounts for roughly one in four banks in our sample. Our figures are consistent with the 50 % minimum coverage ratio to be eligible for using an IRB approach (see ECB Guide to Internal Models General Topics para. 27). As we constrain our analyses on credit risk models, some articles of the CRR will never be infringed in first place, while others are more frequently mentioned. We address this heterogeneity in their distribution by introducing clusters in line with the sections of the CRR. As an illustrative example, articles 138 to 141 would be counted as one section on "Use of the ECAI credit assessments for the determination of risk weights". Lastly, because a bank can have multiple IMIs in a given year, we aggregate the aforementioned information at the bank $\times$ quarter level. As we track a bank through the panel, the reported variables are hence the respective levels of the variable at point t in time, for bank i, unless stated otherwise.

For our analysis we neither winsorize nor truncate our data. Instead, we drop severe outliers as they exclusively relate to one particular bank in our sample. This bank dominates the time-series throughout the panel we analyse, as illustrated in Figure (2). As a result, our final sample consists of 50 IRB banks which we track through 24 quarters. We demonstrate that this choice does not drive our results by running a winsorized model in Section (6.2) below which is dedicated to our

robustness tests.

Figure 2: RWA Impact of the imposed limitation standardised by total assets.



4 Research Question

Given the prevailing importance of using statistical and empirical models in crucial operations of today's financial system, we set out to shed more light on the caveats that are associated with their usage. More specifically, we investigate factors that contribute to our proxy variable for model risk. In doing so, two opposing views exist: One that argues that the number of deficiencies in a model compounds and explains its bad performance. Another argues that instead it is the severity of individual deficiencies which causes material concern.

Research Question 1 *Is the number or the severity of findings the actual driver of the imposed limitation?*

Additionally, we investigate whether early warning signals exist that could alert internal and external stakeholders alike about possible issues with any model under scrutiny. As we can extract the exact non-compliance with the CRR-regulation from our data set, we envisage to pin-point individual pockets of risk that call for greater attention.

Research Question 2 *Can we identify early warning signals that would point to possible drivers of model risk?*

In order to address these questions we apply a standard panel ordinary least squares (OLS) model with fixed effects. In addition to the hand-collected metrics mentioned in Section (3), we complement our regression model with additional metrics. More specifically, we control for different bank business models by calculating the revenue diversification metric (ROID) as shown in Equation (1):

$$\text{ROID}_{i,t} = 1 - \left| \frac{\text{Net Interest Income}_{i,t} - \text{Net Fee and Commission Income}_{i,t}}{\text{Net Interest Income}_{i,t} + \text{Net Fee and Commission Income}_{i,t}} \right| \quad (1)$$

In this notation, t refers to the time, while i represents the individual bank. Additionally, we proxy the riskiness of a bank. To this end, we measure the capital in excess of the minimum Common Equity Tier 1 (CET1) requirements as capital headroom (henceforth Headroom), and its cost of risk (COR), that is:

$$COR_{i,t} = \frac{Impairments_{i,t}}{RWA_{i,t}} \quad (2)$$

where *Impairments* refers to the quarterly flow of loan loss provisions, and RWA is the stock of total risk-weighted assets. In addition we control for the profitability of a bank by employing Return on Assets (ROA), which is net income divided by total assets. We chose ROA over Return on Equity, in order to prevent any systematic biases that may be induced by structurally higher leverage ratios for some banks in the sample. In addition, we control for the size of the bank with the natural logarithm of total assets, in order to account for inherent skewness from banks that enter the sample due to either their relative importance in smaller Euro Area member states and/or being significantly connected with entities abroad.

We regress the aforementioned variables on the ratio of the limitation standardized by total assets (henceforth labeled "Limitation"), in order to account for banks with large IRB models that would hence, ceteris paribus, naturally also incur larger limitations. Our full specification is thus:

$$\begin{aligned} Limitation_{i,t} = {} & \beta_1 F1_{i,t-1} + \beta_2 F2_{i,t-1} + \beta_3 F3_{i,t-1} + \beta_4 F4_{i,t-1} + \beta_5 ROA_{i,t-1} \\ & + \beta_6 Size_{i,t-1} + \beta_7 Headroom_{i,t-1} + \beta_8 COR_{i,t-1} + \beta_9 ROID_{i,t-1} \\ & + \alpha_i + \epsilon_{i,t} \end{aligned} \quad (3)$$

We lag our regressors by one period, in order to address potential concerns with regard to simultaneity. Furthermore, we employ bank-fixed effects to account for additional unmeasured heterogeneity, denoted as α_i, while $\epsilon_{i,t}$ is the error term. We do not employ a time-fixed effect, as a subsequent specification of our model will exploit the variance across time. All standard errors are clustered on the bank level.

5 Results

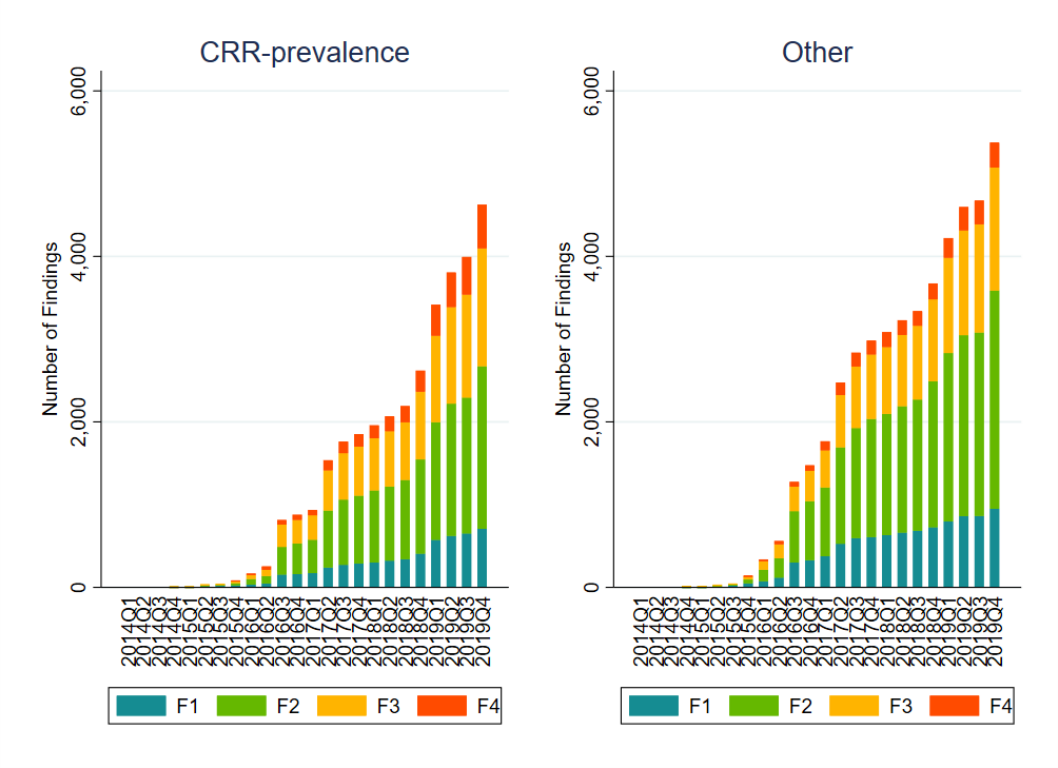
Following or first Research Question above, we begin our analysis by verifying whether there exists a link between the number of findings – clustered by their severity – and the magnitude of the limitation standardized by total assets. To this end, we find in the first column of Table (2) that only in the instance of the highest severity, i.e. "F4", the number of findings is a driver for the severity of the resulting limitation. As this effect is statistically and economically significant we proceed with our analysis and introduce additional regressors to improve the robustness of our model.

Table (2) about here

We conjecture that this observation could also be due to the fact that lower severity findings do not matter in isolation but only in conjunction with higher severity findings. To this end, we interact the counts of low severity findings, i.e. F1 and F2, with the high severities F3 and F4, respectively. We find in the second column that this interaction does not change our results. Instead, F4 remain in statistical terms the exclusive driver of the severity of a limitation. Having established this link, we further extend our model to include bank-level controls. More specifically, it is conceivable that more resilient banks behave differently from riskier banks, which may also be reflected by the modelling approach of these banks. Recall that Blum (2008) shows a stronger incentive for weakly capitalized banks to move towards the IRB approach. Similarly, due to more available resources, larger banks may e.g. be in a more favourable position to model risks along higher quality standards, be it through more available staff or data to calibrate the model. Against this background, we show in column three of Table (2) that our established link between F4 findings and the impact of limitations remains robust at the 95 % confidence level. At the same time, better capitalized banks appear to receive larger limitations as judged by the coefficient of Headroom. A similar story is told from the other end, where riskier banks as judged by COR appear to receive smaller limitations. This observation could be interpreted in line with Agarwal et al. (2024), who show that there is noise in supervisory decision due to human discretion. From a methodological perspective this might not necessarily be harmful, as it allows to incorporate soft information into the decision. In fact, a case can be made that such discretion is necessary, in order to prevent banks from opportunistic behaviour to gamble a possible strictly mechanistic and hence predictable scoring-mechanism. In any case, we scrutinize this observation in more detail in Section (6.1) of the robustness tests, as another way of framing it could arise from reverse causality: banks are only better capitalized because they under-estimate the *true* risk, and hence need bigger limitations to correct for this. Ultimately, our proxy for the business model suggests a statistically significant impact, where more diversified banks receive more severe limitations. However, this effect is economically not meaningful.

Having established the transmission channel that findings of the highest severity are drivers of the impact of the limitation, we proceed to investigate our second research question, that is: is there a possible early warning signal for high severity limitations? We investigate this question using our established empirical model, which we extend by another dimension that is unique to our data: from the findings, we can track the actual non-compliance with the corresponding article of the CRR that led to the finding. Following the clustering technique described in Section (3), we introduce a dummy for each section of the CRR. In untabulated results, we include these dummies in our regression, and find that some of them are particularly meaningful for our model. More specifically, there is a subset of four sections in the CRR, which are less prevalent than all other sections, but at the same time leads to higher severity findings as illustrated in Figure (3) below.

Figure 3: Visualisation of subset of high severity drivers.



While it is obvious to the eye that in Figure (3) "CRR prevalent" articles are less frequent in absolute terms, their relative share of high severity findings, i.e. F3 and F4, is larger, compared to the remainder of the sample. This graphical evidence is further supported by a test for equality of the two groups, illustrated in Table (3). Therefrom, the assumption of both groups being equal is strongly rejected.

Table (3) about here

Following this evidence, we proceed to include a dummy that is one, whenever a finding is associated with an article of the "CRR prevalence" group, and zero otherwise. Including this dummy in the specification of our model shown in column four

suggests that the "CRR prevalent" group is both, a statistically and economically meaningful driver of the severity of the limitation.

Ultimately, we add a dummy that is one for missions that were completed after the Targeted Review of Internal Models (TRIM) had been completed. As described above, the TRIM exercise was designed in order to ensure a level playing field by harmonizing supervisory practices. Hence, it serves as a proxy for possible heterogeneity on the supervisory level. By introducing this dummy, we control for these latent differences in model supervision. Looking to column five of Table (2), we find that missions after the completion of the TRIM had – on average – higher limitations. This observation implies a stricter, but more harmonised interpretation of supervisory rules which led to a more level playing field.

Taken together, we establish a link between high severity findings and the magnitude of the imposed limitation. We show that this link is robust, when controlling for bank characteristics such as riskiness, size, profitability, and business model. We can use this observation to extend our model and show that a possible early warning signal for high severity limitations exists: findings associated with non-compliances with regard to (i) the initial application to use an IRB Approach, (ii) the overall use of models, (iii) requirements for its estimation, and (iv) requirements for own estimates of the Loss Given Default (LGD) should get the undivided attention of supervisors and bank managers alike. The common theme of all these categories is their reference to the underlying data used in the modelling process. Indeed, a case could be made that if the input data are subpar, so will the prediction be. Given that the intercept is insignificant in all specifications of our model, we generate evidence against a possible omitted variable bias. Similarly, we dispel concerns with regard to simultaneity by employing lagged values of our independent variables. We will further demonstrate the robustness of our findings in the subsequent section.

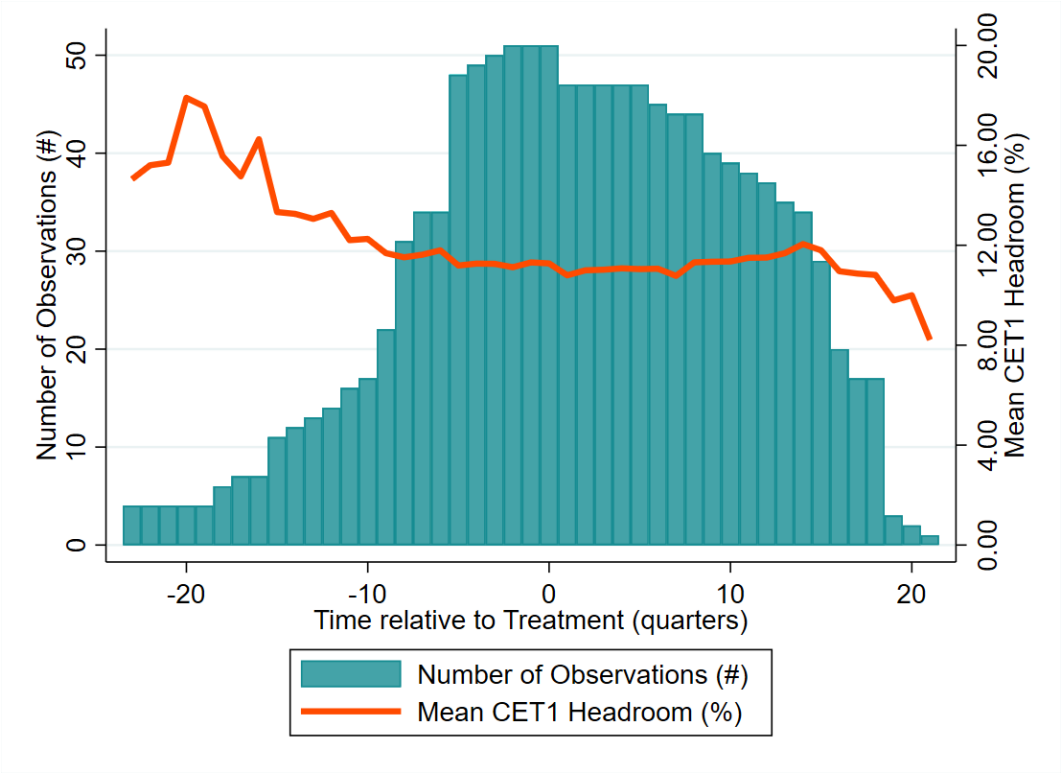
6 Robustness Tests

6.1 Reverse Causality

We treat potential reverse causality with the standard lag of independent variables in our data set. Furthermore, we conduct a quasi event study, using the structure of our data set. In doing so, we exploit the fact that not all banks in our sample had a review of all their IRB models instantaneously upon falling under SSM supervision. Instead, this process was more gradual, hence giving us banks of types A (i.e. no limitation after investigation) and B and C, respectively (full or partial limitation after investigation) – recall Figure (1). Using the IMI as a treatment, type A banks could transition into either B or C types after the inspection. Hence, if there is reverse causality, i.e. better capitalised banks only received higher limitations, because their high level of capital originates from under-estimating their *"true"* risk,

there should be a significant difference in the capital headroom before and after the treatment.

Figure 4: Quasi event study on conducting an IMI as a treatment effect with regard to being a type A bank or not.



As illustrated in Figure (4) there is no visual evidence of reverse causality. If we compare the capital headroom before (i.e. negative value of time) and after (i.e. positive value of time), we find that there is no meaningful difference. Only when expanding the observed time period beyond 10 quarters around the treatment date, any effect becomes visible. However, we argue that the effect and the treatment are too far apart to have a causal relationship. In line with that the decreasing number of banks, for which such a long time-series can be constructed attest that changes in the capital headroom are rather driven by the low number of observations towards the end of the time-series.

Despite this observation, we do not leave it to visual evidence to conduct this robustness test. Using the previously described approach of an IMI as a treatment, we also statistically compare the mean capital headroom of banks and variance thereof before, and after the IMI. In line with the previous observation we find that it can neither be rejected at a statistically meaningful level that the mean capital headroom, nor the variance thereof change significantly through the IMI. The results of this analysis can be obtained from Table (4).

Table (4) about here

We also test the robustness of our results in untabulated results through risk-densities, i.e. the ratio of RWA to total assets, instead of capital headroom. In the presence of the theorized reverse causality, these should change in a statistically significant way, after the IMI by means of a nominator-effect. Using total RWA, credit risk RWA, and IRB RWA as nominators, we find no statistically significant differences, and hence remain confident that our results are not biased by reverse causality. In fact, because an IMI always only covers one model, and hence sub-portfolio, it appears plausible to us that this individual inspection does not suffice to drive capital levels at the entity-level.

6.2 Winsorization

As explained in the data section of this paper, we exclude one bank from our sample as it continuously dominates all other observations in the sample (cf. Figure (2)), and hence biases the results. In this subsection, we investigate the counterfactual of not excluding this particular bank, but instead winsorizing our observations at the 1^{st} and 99^{th} percentile. The results thereof are reported in Table (5).

Table (5) about here

For comparison, we display the original values of our regression adjacent to the winsorized results in Columns (1) and (2) of Table (5). We find that our main observation holds: that is that in particular findings of the highest severity drive the standardized RWA impact of the imposed limitations. All control variables demonstrate comparable signs and magnitudes as well as statistical significance, with the exception of COR, which becomes insignificant in the winsorized specification. Given the number of regressors, we deem this variation to not be meaningful, but rather a statistical artefact.

7 Conclusion

Modern finance has grown increasingly complex and thereby risky. In order to manage these risks, a vast number of statistical techniques are applied, which we broadly refer to as "models". While these models may allow to quantify and thus arguably manage risks, they are not flawless in and by themselves. Mistakes in models due to e.g. misspecification or wrong parametrisation may lead to erroneous decisions because of misleading model output. Those models thereby create a new risk of their own, that is model risk.

In our study, we set out to get a better understanding of model risk by analyzing a confidential data set of supervisory inspections of IRB models which are employed by European banks for measuring Pillar 1 credit risk. We narrow down our analysis to this particular set of risk, in order to have an as clean as possible identification

of the actual drivers of this new category of risk, i.e. model risk.

In doing so, we investigate two research questions. The first relates to two opposing schools of thought as to what constitutes a *"bad"* model: Is it either a model with numerous deficiencies, or rather a model with limited numbers of but substantially big concerns? Our analysis shows that the number of issues identified in a model is – statistically speaking – not a significant driver of the counterfactual penalty that the supervisor has to impose in order to maintain a level playing field. Instead the severity of the associated findings drive the limitation. Having established a link between high-severity findings and the underestimation of a risk model, we ask in our second research question whether there exist early warning signals for these issues. Indeed, we find a subset of articles from the CRR, which, despite being less frequent than other articles, are more often associated with higher severities. In this sense, we disentangle the signal from the noise by pointing to particular parts in the risk modeling, which should be focal points for supervisors to check. At the same time, this should not create a false sense of security, as "all models are wrong".

That being said, our results are useful for bankers and supervisors alike. First, they allow to prioritize scarce resources on both sides in order to be particularly mindful on key topics that are risk drivers within the field of risk modeling. Second, and in line with this thought of prioritization, we generate evidence suggesting that supervisors should not occupy themselves with checking every minuscule detail, but instead focus on the material deficiencies in a model. This recommendation is robust to possible reverse causality concerns as well as outlier treatments in the data we analyze.

Future research should try to reinstate our findings in different jurisdictions and different legal settings in order to narrow down the list of early warning signals that point to deficiencies in models that are used to ultimately calculate solvency ratios.

References

Agarwal, S., Morais, B. C., Seru, A., and Shue, K. (2024). Noisy Experts? Discretion in Regulation. Working Paper 32344, National Bureau of Economic Research.

Aggarwal, A., Beck, M. B., Cann, M., Ford, T., Georgescu, D., Morjaria, N., Smith, A., Taylor, Y., Tsanakas, A., Witts, L., and Ye, I. (2015). Model Risk - Daring to open up the Black Box. *British Actuarial Journal*, 21:229–296.

Barrieu, P. and Scandolo, G. (2015). Assessing financial model risk. *European Journal of Operational Research*, pages 546 – 556.

Barucci, E. and Milani, C. (2018). Do European Banks manipulate Risk Weights? *International Review of Financial Analysis*, 59:47–57.

Basel Committee on Banking Supervision (2013). The regulatory framework: balancing risk sensitivity, simplicity and comparability.

Basel Committee on Banking Supervision (2023). Basel III Monitoring Report.

Behn, M., Haselmann, R., and Vig, V. (2022). The Limits of Model-Based Regulation. *Journal of Finance*, 77(3):1635–1684.

Blum, J. M. (2008). Why 'Basel II' may need a leverage ratio restriction. *Journal of Banking & Finance*, 32(8):1699–1707.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Bruno, B., Marino, I., and Giacomo, N. (2023). Internal ratings and bank opacity: Evidence from analysts' forecasts. *Journal of Financial Intermediation*, 56:101062.

Colliard, J.-E. and Georg, C.-P. (2020). Measuring Regulatory Complexity. CEPR Discussion Papers 14377, C.E.P.R. Discussion Papers.

Cont, R. (2006). Model uncertainty and its impact on the pricing of derivative instruments. *Mathematical Finance*, 16.

Daníelsson, J., Shin, H. S., and Zigrand, J.-P. (2004). The Impact of Risk Regulation on Price Dynamics. *Journal of Banking & Finance*, 28(5):1069–1087.

Derman, E. (1996). Model Risk in Quantitative Strategies Research Notes from Goldman Sachs.

European Banking Authority (2017). Guidelines on PD and LGD estimation.

European Central Bank (2021). Targeted Review of Internal Models - Project report. Technical report.

European Central Bank (2024). ECB guide to internal models.

Hakenes, H. and Schnabel, I. (2014). Regulatory Capture by Sophistication. *ERN: Econometric Studies of Government Regulation of Financial Markets (Topic)*.

Haldane, A. and Madouros, V. (2012). The Dog and the Frisbee. *Proceedings - Economic Policy Symposium - Jackson Hole*, pages 109–159.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford University Press.

Leitner, Y. and Yilmaz, B. (2019). Regulating a model. *Journal of Financial Economics*, 131(2):251–268.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46.

Mariathasan, M. and Merrouche, O. (2014). The manipulation of basel risk-weights. *Journal of Financial Intermediation*, 23(3):300–321.

Plosser, M. C. and Santos, J. A. C. (2018). Banks' Incentives and Inconsistent Risk Models. *Review of Financial Studies*, 31(6):2080–2112.

Rogers, C. (2021). Basel III and global cooperation: where do we go from here? – Keynote speech at the The Kangaroo Group virtual debate.

Vallascas, F. and Hagendorff, J. (2013). The Risk Sensitivity of Capital Requirements: Evidence from an International Sample of Large Banks. *Review of Finance*, 17(6):1947–1988.

8 Appendix

Table 1: Descriptive Statistics of used Variables

	$Obs.$	$Min.$	$Q_{0.25}$	$Mean$	$Median$	$Q_{0.75}$	$Max.$	σ
Limitation (%)	1,180	0.0000	0.0000	0.0099	0.0020	0.0140	0.1433	0.0152
F1 (#)	1,200	0.0000	0.0000	15.6175	8.0000	24.0000	121.0000	19.7479
F2 (#)	1,200	0.0000	0.0000	39.7808	22.0000	60.5000	272.0000	48.2814
F3 (#)	1,200	0.0000	0.0000	24.3858	10.0000	32.0000	274.0000	38.1216
F4 (#)	1,200	0.0000	0.0000	6.5717	2.0000	8.0000	84.0000	12.0921
ROA (%)	1,180	-0.0507	0.0041	0.0092	0.0087	0.0135	0.0497	0.0094
Size (ln)	1,180	22.5844	24.8145	25.7818	25.7247	26.9714	28.5260	1.3344
Headroom (%)	1,200	0.0200	0.0873	0.1158	0.1037	0.1290	0.3262	0.0437
COR (%)	1,180	-0.0395	0.0003	0.0031	0.0013	0.0039	0.663	0.0060
ROID ($\epsilon[-1;1]$)	1,180	-0.9925	0.3399	0.5088	0.5384	0.7472	0.9995	0.3369
CRR-prevalent ($\epsilon[0;1]$)	1,200	0.0000	0.0000	0.7208	1.0000	1.0000	1.0000	0.4488
TRIM ($\epsilon[0;1]$)	1,200	0.0000	0.0000	0.3333	0.0000	1.0000	1.0000	0.4716

Note: The table above gives an overview of the used variables and their descriptive statistics. The scale of the variable is given in parenthesis behind the name. Size is the natural logarithm of total assets. ROID is calculated in line with Equation (1) and COR according to Equation (2). CRR-prevalent refers to the subset of articles in the CRR that were identified as particularly important for the relative height of the imposed limitation. TRIM is a dummy that is one for all IMIs after the conclusion of the TRIM exercise.

Table 2: Regression on the EUR impact of a limitation over total assets

	(1)	(2)	(3)	(4)	(5)
F1 (#)	0.0002	0.0002	0.0002	0.0001	0.0001
F2 (#)	0.0001	0.0001	0.0001	0.0000	-0.0000
F3 (#)	-0.0001	-0.0001	-0.0001	-0.0000	-0.0000
F4 (#)	0.0005**	0.0006*	0.0006*	0.0005*	0.0005*
low severity × F3 (#)		-0.0000	-0.0000	-0.0000	-0.0000
low severity × F4 (#)		-0.0000	-0.0000	-0.0000	-0.0000
ROA (%)			-0.0451	-0.0716	-0.0206
Size (ln)			-0.0026	-0.0025	-0.0029
COR (%)			-0.2213*	-0.2191*	-0.2069*
Headroom (%)			0.1339**	0.1051**	0.1004**
ROID ($\epsilon[0;1]$)			0.0004*	0.0002	0.0002
CRR-prevalent ($\epsilon[0;1]$)				0.0071*	0.0067*
TRIM ($\epsilon[0;1]$)					0.0044**
Intercept	-0.0020	-0.0030	0.0507	0.0483	0.0582
Bank-fixed effects	Yes	Yes	Yes	Yes	Yes
N	1,180	1,180	1,130	1,130	1,130
R^2	0.3337	0.3486	0.3927	0.4199	0.4334

Note: The table above demonstrates the step-wise extension of our model, which establishes in the first column a significant link between the severity of a limitation and the associated findings. As shown in the second column, this link is not severed by the coinciding of many lower severity findings. From there we proceed to include bank-level controls to prove the resilience of this transmission channel, even when accounting for idiosyncrasies. Drawing from further analysis, we demonstrate in the fourth specification that a particular subset of articles in the CRR are a statistically significant driver of the severity of the limitation, although they are generally less frequent in our data set. Ultimately, missions after the TRIM exercise have – on average – higher capital requirements, suggesting that supervisory practices were harmonized at a higher common standard. Because of this, we do not include time-fixed effects, as a cross sectional analysis has shown that they only become significant after said completion of the TRIM exercise. Significance is denoted at the 5 % (*), 1 % (**), and 0.1 % (***) level.

Table 3: Test for equality using the mean.

	Average Findings		
	CRR-prevalence	Other	p-value
F1	351.0556	574.0000	0.0000
F2	1,001.0560	1,465.7220	0.0000
F3	757.6944	839.0833	0.0001
F4	260.2778	175.4444	0.0002

Note: The table above displays the mean number of findings for the two groups per severity. We test for the equality of the mean of both groups and find that they are different at a statistically highly significant level. Another interesting observation made from this table is a possible excess of F2 findings, which are more frequent that two of the three remaining categories combined.

Table 4: Comparison of Mean and Variance between types of banks in the sample

	Before	After	Difference / Ratio	Probability
Mean	11.5564	11.4632	0.0932	0.7103
Variance	0.1681	0.1863	0.9023	0.1248

Note: The table above details a comparison of the mean and variance of two sets of banks in our sample (recall Figure (1)). Banks of type A (Before) are those that are compliant with regulation and do not have a limitation in place at time t. Types B and C, however, have – to varying degrees - limitations in place in order to enforce the level playing field (After). We use this structural difference to investigate possible reverse causality. More specifically, it could be the case that well capitalized banks do not receive larger limitations, but rather that banks are well capitalized because they have not (yet) received a limitation. To this end, we investigate whether banks of type A and B/C are structurally different, by comparing the mean and variance between both group, relative to them moving from one to the other. From this analysis, we cannot reject at a statistically meaningful confidence level that the mean and variance of the two groups are different.

Table 5: Regression on the EUR impact of a limitation over total assets

	(1)	(2)
F1 (#)	0.0001	0.0001
F2 (#)	0.0000	-0.0000
F3 (#)	-0.0000	-0.0000
F4 (#)	0.0005*	0.0004*
low severity × F3 (#)	0.0000	0.0000
low severity × F4 (#)	-0.0000	-0.0000
ROA (%)	-0.0345	-0.0327
Size (ln)	-0.0053	-0.0083
COR (%)	-0.2106*	-0.1094
Headroom (%)	0.1158*	0.1011**
ROID ($\epsilon[0;1]$)	0.0052	0.0049
CRR-prevalent ($\epsilon[0;1]$)	0.0064*	0.0070*
TRIM ($\epsilon[0;1]$)	0.0044**	0.0044**
Intercept	0.1128	0.1859
Bank-fixed effects	Yes	Yes
N	1,130	1,154
R^2	0.4349	0.4638

Note: The table above depicts the original regression in Column (1) and contrasts it to the winsorized model in Column (2). Our variable of interest, the number of findings with severity F4 remains statistically significant in this specification, as do the control variables, with the exception of the Cost of Risk variable. The importance of being well-capitalised is in statistical terms more significant in the second specification. At the same time, its economic impact only differs marginally. Significance is denoted at the 5 % (*), 1 % (**), and 0.1 % (***) level.