FROM SCORES TO SIGNALS:

Understanding Risk Classification in a Multi-Score Mortgage Market

PROSPERITYNOW

About Prosperity Now:

Since 1979, Prosperity Now has been a trusted leader in strengthening financial security, expanding access to capital, and ensuring economic stability for businesses, families, and communities. We work across sectors to develop practical, scalable solutions that create lasting change. Through innovation, strategic investment, and collaboration, we build the infrastructure needed to sustain small business growth, housing opportunities, and financial well-being in an evolving economic landscape.
Learn more at: **www.prosperitynow.org.**

Table of Contents

Executive Summary

This report examines how different credit scoring models perform in practice using loan-level data from the Government-Sponsored Enterprises (GSEs) Fannie Mae and Freddie Mac, covering the period from 2013 through 2023, including the COVID-19 stress period.[1]

The analysis focuses on how Fair Isaac Corporation (FICO) Classic Score and VantageScore 4.0 models rank borrower risk, how that risk is calibrated across score ranges, and how both behave under real-world stress conditions. The objective is to assess how credit scoring functions within the broader mortgage system and what model behavior implies for underwriting, servicing, pricing, and overall market stability.

Across the full population, both models perform a core function reliably: they rank-order borrower risk. However, the findings show that differences between models emerge not in whether risk is identified, but in how that risk is classified, segmented, and translated into decisions within the mortgage system.[2]

Within the current system, risk is consistently rank-ordered, but differences in how risk is calibrated and priced suggest that observed outcomes are not always proportionally reflected in score distribution or loan pricing. These findings also reflect structural characteristics of the current single-score system, including compressed pricing, reliance on a single primary risk signal, and limited alignment between observed risk and how borrowers are segmented across score ranges.

This analysis evaluates how each model classifies and differentiates risk within the same population, and how those differences align with observed loan performance under consistent historical conditions. It does not simulate how outcomes would differ under alternative underwriting or pricing frameworks. Instead, it evaluates how each model performs when applied consistently to the same population under observed historical conditions.
Credit scores do not operate in isolation. They are used to make decisions about eligibility, pricing, servicing, and capital allocation. The findings that follow should therefore be interpreted not only as differences in model performance, but as differences in how risk may be represented and acted upon within the mortgage system.

Key Findings

The findings below are ordered by their practical consequences for the mortgage market, from most to least consequential.

1. **Greater Risk Differentiation Under Stress Conditions**

   During the COVID-19 period, both models maintained consistent rank ordering, but VantageScore 4.0 produced a wider separation between higher- and lower-risk borrowers under stress conditions. At Freddie Mac, VantageScore 4.0 generated a decile-level default rate spre ad of 19.6x compared to 13.9x under FICO Classic Score, approximately 41 percent wider. This difference is most visible at the upper end of the distribution, where borrowers assigned to the lowest risk deciles by VantageScore 4.0 demonstrate lower post-forbearance default rates than comparable segments under FICO Classic Score.

What this means:

Greater risk separation under stress is not a technical nicety — it has direct financial consequences. When a scoring model more accurately distinguishes between a borrower who will recover from temporary hardship and one who will eventually default, it enables better-calibrated capital reserves, more effective loss mitigation targeting, and more accurate investor expectations about pool performance.

In a stressed environment, even a modest improvement in discrimination at the top of the risk distribution can translate into substantially lower realized losses across a large portfolio.

2. **Forbearance Masked, but Did Not Eliminate, Underlying Risk**

   During the COVID-19 period, widespread forbearance programs suppressed observed defaults across both datasets. Using the conditional forbearance-to-default metric (CondFbDef%), which measures the share of borrowers who entered forbearance and subsequently defaulted, VantageScore 4.0 more clearly differentiated borrowers within lower-risk segments. At Fannie Mae, borrowers in the top decile under VantageScore 4.0 exhibited a CondFBDef rate of 0.15 percent, compared to 0.22 percent under FICO Classic Score, a 32 percent difference. These results indicate that borrowers assigned to top-tier segments by VantageScore 4.0 exhibited lower realized default rates once forbearance protections ended, even when raw delinquency metrics were suppressed.

What this means:

Servicers, investors, and GSEs rely on risk scores not just at origination but throughout a loan's life, particularly during crises. A score that more accurately identifies which borrowers will recover from hardship versus transition into default enables more targeted loss mitigation, reduces unnecessary intervention costs, and produces better outcomes for borrowers. The gap observed here is not academic: it translates directly into expected losses, capital planning, and the effectiveness of forbearance programs as policy tools.

3. **The Two Models Classify the Same Borrowers Differently**

   A subset of loans — approximately one percent of the total population — received materially different risk classifications from the two models. These disagreement cases isolate the scenarios where model choice has the most direct practical consequence.

In the larger and more consequential direction of disagreement, VantageScore 4.0 flagged borrowers as higher risk (below 620 or 620–659) while FICO Classic Score placed those same borrowers in prime bands (740 and above). Across 111,726 Fannie Mae loans in this group, the realized default rate was 5.97 percent — more than double the 2.81 percent baseline for the FICO 740+ population.

The reverse direction — where VantageScore 4.0 assigned higher scores and FICO Classic Score lower scores — produced more ambiguous results, with default outcomes near the average for the relevant segment. This asymmetry is significant: disagreement was not evenly distributed, and the consequences were not symmetric.

In both datasets, loan pricing followed the FICO Classic Score classification in disagreement cases, even where VantageScore 4.0 had identified elevated risk that subsequently materialized.

What this means:

The one-sided nature of the disagreement pattern reduces concerns about symmetric routing or selection effects under a multi-score framework.

The risk is not that either score could be selectively used to route borrowers in whichever direction is convenient; it is that one score identifies a genuine risk signal that the other does not.

For originators, investors, and taxpayers who ultimately bear credit losses, the accuracy of that signal matters.

4. Differences in Risk Calibration Across the Score Distribution

Both models consistently rank-order borrowers at risk, where higher scores correspond to lower default rates, lower scores to higher default rates, with no exceptions across either dataset. Where they diverge is in how that risk translates into specific score levels, particularly at the extremes of the distribution.

VantageScore 4.0 assigns lower scores to higher-risk borrowers and applies a higher threshold for top-tier classification than FICO Classic Score. At the lower end of the distribution, where VantageScore 4.0 averaged 557 versus 643 under FICO Classic Score, borrowers in these segments exhibit realized default rates in the range of approximately 17 percent to 21 percent. At the upper end, VantageScore 4.0 assigns fewer borrowers to the highest score bands, requiring scores approximately 16 to 25 points higher than FICO Classic Score to reach the same tier, resulting in a more narrowly defined low-risk segment.

What this means:

Differences in calibration influence how borrowers fall relative to eligibility thresholds, how pricing tiers are assigned, and how risk is distributed across portfolios. For originators, the same borrower may be classified differently depending on the scoring model used, even when relative risk ranking is consistent. For investors, loan pools assembled under one scoring framework may carry different underlying risk than pools assembled under another, even if aggregate metrics appear similar. For borrowers, model choice can affect whether they receive a loan and at what price, without any change in their actual creditworthiness.

5. **Pricing Has Not Kept Pace with Risk Differences in the Current Single-Score System**

   Observed interest rate differences across score segments are substantially more narrow compared to the variation in default risk, with compression ratios ranging from approximately 11.7:1 to 15.2:1 across both datasets and scoring models. When risk varies roughly twelve to fifteen times more steeply across the score distribution than pricing does, borrowers at materially different levels of default risk may receive interest rates that are relatively close together.

   This is a structural feature of the current system, which was designed around a single primary risk signal and a fixed pricing grid. FICO Classic Score-based pricing shows higher compression than VantageScore 4.0-based comparisons, suggesting that the existing LLPA structure is calibrated to a score distribution that may not fully capture the risk gradations VantageScore 4.0 identifies.

   It is important to note that this analysis reflects pricing within the existing framework — loans originated and priced under FICO-based underwriting and current LLPA grids. Adding VantageScore 4.0 to the system without corresponding adjustments to pricing frameworks would not, by itself, resolve this compression. The LLPA grid would need to be evaluated and potentially recalibrated to reflect the risk signals a second score provides.

What this means:

For originators and investors, compressed pricing means that the economic return on loans does not vary proportionally with the actual risk being taken. Taxpayers, who bear residual credit risk through the GSE guarantee structure, face a system in which some higher-risk loans are effectively subsidized by their more creditworthy peers. The introduction of VantageScore 4.0 creates an opportunity to improve pricing alignment, but only if pricing frameworks evolve in parallel with scoring.

6. **Consistent Rank-Ordering of Risk**

   Across the full population and all time periods analyzed, both VantageScore 4.0 and FICO Classic Score produce consistent, monotonic relationships between score and default rates — higher-scored borrowers default less, lower-scored borrowers default more, with no exceptions across deciles or quintiles in either dataset.

What this means:

Both models reliably differentiate relative borrower risk across the population. The primary differences observed in this analysis relate to calibration, segmentation, and behavior under stress rather than the ability to rank-order borrowers. This establishes a baseline: both models reliably identify relative risk, and the meaningful differences arise in how that risk is calibrated and applied.

What These Findings Mean for the Mortgage Market

Taken together, these findings describe not just differences between two credit scoring models, but differences in how risk is represented within a system that has historically relied on a single primary score. As the market transitions to a multi-score framework, these differences introduce new considerations for how risk is interpreted and applied across the mortgage system. The introduction of an additional score does not replace existing risk signals, but introduces scenarios in which multiple classifications must be interpreted and reconciled within a single decision framework.

The implications of this shift vary across participants:

For Originators

The presence of two scoring approaches introduces additional complexity at the point of underwriting and pricing decisions. Borrowers may meet eligibility thresholds under VantageScore 4.0 but fall outside comparable thresholds under FICO Classic Score, or vice versa. In disagreement cases, where the same borrower is assigned to different score bands, the choice of which score informs eligibility and pricing decisions directly affects loan approval outcomes and pricing consistency.

In this dataset, pricing reflects FICO Classic Score classifications under the current system, even in cases where VantageScore 4.0 assigns a lower score band and observed performance outcomes are higher risk. This highlights how current decision frameworks may not fully incorporate differences in risk signals across models and underscore the importance of clear standards for resolving conflicting classifications in a multi-score environment.

For Investors

The composition of loan pools evaluated using a single score may differ from how those same loans would be classified under a multi-score framework. Disagreement cases represent borrowers for whom risk classification differs across models, and in this dataset, these cases are associated with materially different observed performance outcomes relative to baseline expectations for higher score bands.

Investors relying solely on FICO Classic Score distributions to assess pool quality may not fully capture variation in underlying risk across these segments. As multiple scores are incorporated into market practice, greater transparency around score distributions and cross-model disagreement patterns may become increasingly important for assessing credit risk and pricing mortgage-backed securities and credit risk transfer transactions.

For Borrowers

The introduction of a second score may affect how borrowers are evaluated across lenders and channels. Some borrowers may qualify under one scoring approach but not another, while others may be assigned to different pricing tiers depending on the model applied.

These differences do not reflect changes in underlying borrower behavior, but rather differences in how risk is classified and segmented across models. The overall effect on access to credit and loan terms will depend on how lenders and market participants incorporate multiple scores into underwriting and pricing decisions, as well as how consistently those decisions are applied.

For Taxpayers and the Broader System

The GSEs operate within a framework that ultimately allocates credit risk across lenders, investors, and, under certain conditions, taxpayers. The findings on risk-pricing compression indicate that, within the current system, pricing does not move proportionally with observed risk across score segments.

In disagreement cases, where models assign different classifications to the same borrower, pricing reflects the FICO Classic Score signal under current practice, even when observed performance outcomes differ across segments. This suggests that how risk is classified and priced within the system has implications for how risk is distributed and absorbed across market participants.
As a multi-score framework is implemented, the alignment between risk signals, pricing structures, and capital allocation will be an important consideration for maintaining system stability and transparency.

Important Considerations

This analysis reflects observed loan performance within an existing market structure. Loans in the dataset were originated, priced, and managed under established underwriting frameworks and policy conditions, including COVID-era forbearance programs.

The findings should be interpreted as evidence of how scoring models classify and differentiate risk within this environment. They do not represent a forward-looking simulation of how these models would perform under alternative underwriting, pricing, or behavioral conditions.

As a result, conclusions should be interpreted within the context of current system design and data availability.

Introduction

Credit scoring plays a central role in mortgage underwriting, pricing, and secondary market activity. For decades, a single scoring framework has served as the foundation for how borrower risk is assessed, how loans are priced, and how credit risk is distributed across the housing finance system.

That framework is now evolving. Historically, this framework has performed a clear and essential function: it provides a consistent way to rank-order borrower risk across the population. At the same time, it reflects the constraints of a single-score system, where calibration, segmentation, and pricing are anchored to one primary risk signal.

As updated scoring approaches are introduced, the system is moving toward a structure in which multiple models may inform how risk is evaluated. This transition expands not only how scores are calculated, but how risk signals are interpreted and acted upon across the system. Differences in how models classify and calibrate risk may influence borrower eligibility, pricing outcomes, servicing strategies, and investor expectations.[3]

This report provides an independent analysis of credit scoring model performance using loan-level data from Fannie Mae and Freddie Mac, covering the period from 2013 through 2023. The analysis draws upon approximately 44.7 million matched loans across both datasets, representing one of the largest evaluations of scoring model behavior against observed mortgage outcomes. This analysis evaluates VantageScore 4.0 and FICO Classic Score, which are observable within current GSE datasets. Newer scoring models, including FICO 10T, are not yet available with sufficient historical performance data to be included in this analysis.

The analysis focuses on how scoring models:

- Differentiate risk across the borrower population
- Translate risk into score levels, particularly at the tails of the distribution
- Perform under periods of economic stress
- Classify the same borrowers in cases where model outputs diverge

The COVID-19 period provides a unique real-world stress test for evaluating these dynamics. During this time, widespread forbearance programs altered the visibility of borrower distress, suppressing observed defaults and delaying the realization of underlying risk.[4] This creates an environment in which model behavior can be evaluated under both stable conditions and policy-influenced stress.

The objective of this report is to assess how scoring models represent and differentiate risk within this environment, and to translate those findings into system-level considerations. This includes implications for underwriting thresholds, servicing strategies, pricing alignment, and market behavior as the mortgage system adapts to a multi-score framework.

To structure the analysis, the report is organized around four core questions:

1. **Do scoring models maintain consistent risk differentiation under both stable and stress conditions?**
   This includes evaluating rank-ordering performance across deciles and quintiles, and whether that ordering holds during periods of economic disruption.
2. **Where do models produce similar rank-ordering but different calibration outcomes?**
   This focuses on how borrowers are distributed across score ranges, particularly at key thresholds such as sub-620 and higher score bands (e.g., 740+).
3. **What do disagreement cases reveal about how risk is classified?**
   By isolating loans where models assign materially different scores, the analysis examines how those differences translate into observed performance outcomes.
4. **What broader system-level considerations emerge in a multi-score environment?**
   This includes implications for underwriting thresholds, pricing alignment, servicing strategies, and borrower and lender behavior as multiple scoring approaches are introduced.

Market Context and Structural Considerations

The U.S. mortgage market is entering a period of structural transition from a long-standing single-score framework to a multi-score environment.

Historically, mortgage underwriting has relied on a limited set of legacy credit scores, applied through standardized approaches such as tri-merge.[5] In this framework, a single composite score has served as the primary signal for borrower risk, informing eligibility, pricing, and secondary market execution.

Recent policy and market development are changing that structure.

The introduction of VantageScore 4.0 alongside FICO Classic Score creates a dual-score framework in which two potentially different risk signals may be observed for the same borrower at origination. This shift does not simply expand the number of available scores; it introduces the possibility of divergence in how risk is measured and interpreted across the system.
Several structural considerations emerge from this transition:

Score Dispersion
Different scoring models may assign materially different scores to the same borrower, particularly in certain segments of the credit distribution.[6] As shown in the findings that follow, these differences are not uniform across the population and are most pronounced at the tails, where risk classification has the greatest implications for eligibility and pricing.

Threshold and Decision Alignment
In a dual-score environment, the question of which score governs underwriting thresholds, pricing adjustments, or capital treatment becomes operationally significant. A borrower may meet eligibility criteria under one model and fall outside thresholds under another, raising questions about consistency in decision-making and standardization across lenders and channels.

Routing and Selection Effects

The presence of multiple scores introduces the potential for routing behavior. Lenders may determine which score to prioritize based on product, channel, or internal policy, while borrowers may experience different outcomes depending on where and how they apply.

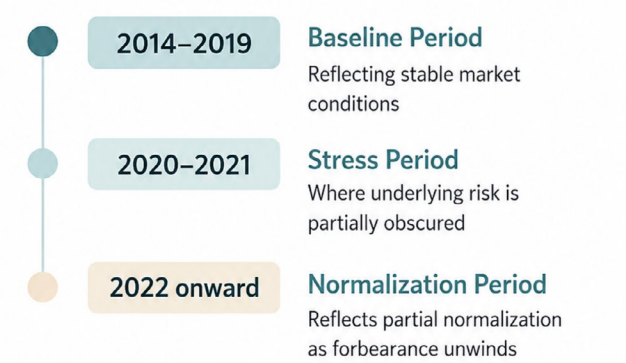
These dynamics are not directly observable in historical data but represent a key consideration for how a multi-score system may function in practice.

Implications for Market Interpretation

Differences in score assignment and calibration have downstream effects on how risk is understood across the mortgage ecosystem. These differences influence how risk is interpreted across the system, including underwriting consistency, pricing alignment, servicing expectations, and investor assessment of loan quality.

The ability of a scoring model to differentiate risk under stable conditions is necessary, but not sufficient for evaluating how that model performs in real-world conditions. The more critical test is how that model behaves when economic conditions deteriorate.
The COVID-19 period provides the only observable stress scenario within this dataset. However, it is also analytically complex. Forbearance programs suppressed observed defaults, limiting the usefulness of raw delinquency measures during 2020–2021.

Within this context:



As a result, evaluating model performance during this period requires a focus on how risk is differentiated and revealed over time, rather than relying solely on contemporaneous default rates.

Understanding how scoring models behave across these conditions is essential as the market adapts to a multi-score framework. The sections that follow examine these dynamics directly, using consistent datasets to isolate differences in risk classification, calibration, and performance under stress.

Data & Methodology

Data Sources and Coverage

Differences in score assignment and calibration have downstream effects on how risk is understood across the mortgage ecosystem. These differences influence how risk is interpreted across the system, including underwriting consistency, pricing alignment, servicing expectations, and investor assessment of loan quality.

This analysis draws on loan-level datasets from Fannie Mae and Freddie Mac, covering the period from 2013 through 2023. Together, these datasets represent approximately 44.7 million matched loans and more than one billion monthly performance observations. Loan performance data were merged with corresponding credit score data using loan-level identifiers, resulting in match rates of approximately 95 percent across both datasets. The remaining unmatched observations are primarily attributable to missing score data in specific reporting periods rather than to systematic data loss.

Key dataset characteristics are summarized in Table 1.

Dataset	Loans	Coverage	Match Rate	Primary Metric
Freddie Mac	19.6M	2013–2022	95.4%	90+ DPD (ever)
Fannie Mae	25.1M	2013–2023	95.0%	Ever 90+ DPD / Default
Combined	44.7M	2013–2023	~95%	CondFbDef (stress)[1]

Table 1: Dataset Overview

For consistency, 'default' is used throughout this report as shorthand for loans that reach 90+ days past due (90+ DPD), unless otherwise specified.
Credit scores are observed at the origination. VantageScore 4.0 is applied using the current method (median-based), and FICO Classic Score reflects the original credit score used in underwriting.

Analytical Framework

The analysis evaluates model behavior across multiple levels of segmentation to capture both broad and granular differences in risk classification.

Score Bands

Borrowers are grouped into standard GSE score bands:

• <620, 620–659, 660–699, 700–739, 740–779, 780+

•Freddie Mac includes additional granularity below 620 (<580, 581–619)

These bands reflect common underwriting and pricing thresholds used in practice.

Deciles and Quintiles

To enable direct comparison across models with different score distributions, the analysis also evaluates performance using:

• **Deciles (D1–D10):** Equal-count groupings, where D1 represents highest-risk borrowers and D10 represents lowest-risk

• **Quintiles (Q1–Q5):** 20 percent buckets, used to assess broader segmentation patterns

Risk differentiation is measured using spread ratios, calculated as the ratio of default rates between the highest-risk and lowest-risk segments (e.g., D1 ÷ D10).

Median vs. Average Score Treatment

Where applicable, VantageScore 4.0 is evaluated using the current (median-based) method, which reflects how scores are derived across credit bureau inputs.

Comparisons using alternative aggregation methods (e.g., average) are included where relevant to assess sensitivity. The choice of method is driven by the specific analytical context and is noted in each case.

Stress Period Segmentation

Model performance is evaluated across distinct market conditions:

- Baseline period: 2014–2019 originations, representing stable credit conditions
- Stress period: 2020–2022 originations, reflecting COVID-era disruption
- Post-COVID period: Limited coverage due to insufficient loan seasoning

Because forbearance programs suppressed observed defaults during the COVID period, stress analysis relies on conditional forbearance-to-default (CondFbDef) as the primary metric for evaluating performance under stress conditions.

Interpretive Context

All scores are evaluated within a consistent historical dataset in which VantageScore 4.0 is applied retrospectively to loans originally underwritten using FICO Classic Score.

As a result, the analysis reflects how different scoring models classify and differentiate risk within the same population, rather than how those models would perform under alternative underwriting or pricing decisions.

All loans in this dataset originated, priced, and managed under legacy underwriting and pricing frameworks based on FICO Classic Score. As a result, pricing observations reflect existing system behavior and do not represent pricing outcomes under a multi-score framework.

Additional considerations related to data limitations, forbearance dynamics, and post-COVID coverage are addressed in the Limitations and Guardrails section.

Findings

Interpreting the Findings

The findings that follow are based on a retrospective analysis of loan-level performance across Fannie Mae and Freddie Mac datasets from 2013–2023. In this analysis, VantageScore 4.0 is applied alongside FICO Classic Score to evaluate how each model classifies and differentiates risk within the same population.

It is important to note that these results do not reflect how loans would have performed under an alternative underwriting or pricing regime. All loans in the dataset were originated, priced, and managed using existing market frameworks, including FICO Classic Score-based underwriting and COVID-era forbearance and loss mitigation policies.

As such, the findings should be interpreted as a comparison of model behavior and risk representation under consistent historical conditions, rather than as a forward-looking assessment of how these models would perform in live market implementation.

Across the analysis, both models demonstrate strong and consistent rank-ordering of borrower risk. The more consequential differences emerge in how risk is calibrated, how borrowers are segmented across score ranges, how models behave under stress conditions, and how outcomes differ when the models assign materially different classifications to the same loan.

These distinctions matter because credit scores are not used only to rank risk. They inform underwriting thresholds, pricing tiers, servicing strategies, investor interpretation, and broader market expectations. Where relevant, additional context on data limitations, stress-period dynamics, and methodological considerations is provided in the sections that follow.

Finding 1:
Rank-Ordering Consistency

Core Insight

Both VantageScore 4.0 and FICO Classic Score demonstrate consistent monotonic relationships between score and default risk across the full borrower population, with no observed risk inversions at the decile or quintile level in either GSE dataset.

Decile spreads are comparable across models, with modest differences in magnitude (11.5× vs. 10.1× at Freddie Mac; 14.6× vs. 13.8× at Fannie Mae), while quintile spreads are closely aligned. These results indicate that both models perform similarly in their ability to rank-order relative borrower risk.

Supporting Evidence

Figure 1.1 presents decile-level default rate ladders for both models across each GSE dataset. Each bar represents the observed default rate for borrower cohorts ranked from highest risk (D1) to lowest risk (D10).

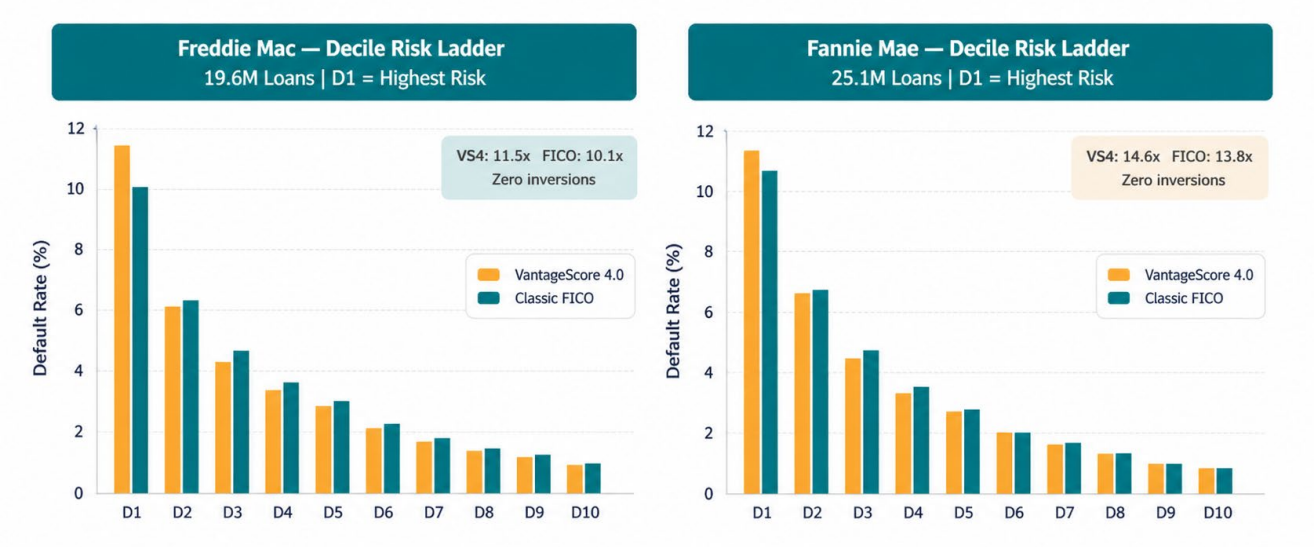


Figure 1.1: Decile Risk Ladders — VantageScore 4.0 vs. FICO Classic Score (Freddie Mac & Fannie Mae). Both models produce perfectly monotonic risk ladders with no inversions. VantageScore 4.0's decile spread is marginally wider at both GSEs; quintile spreads converge. Source: Tables B.1, B.2 (Appendix B).

The following table summarizes key discriminatory power metrics across both GSEs:

Metric	Freddie Mac (19.6M Loans)	Fannie Mae (25.1M Loans)
VantageScore 4.0 Decile Spread (D1÷D10)	11.5× (90DPD)	14.6× (Default)
FICO Decile Spread (D1÷D10)	10.1× (90DPD)	13.8× (Default)
VantageScore 4.0 Quintile Spread (Q1÷Q5)	7.5× (90DPD)	9.3× (Default)
FICO Quintile Spread (Q1÷Q5)	7.6× (90DPD)	9.8× (Default)
Risk Inversions	Zero	Zero
Top Decile Score Floor — FICO	802+	806+
Top Decile Score Floor — VANTAGESCORE 4.0	827+	830+

Source: Tables B.3 (Appendix B)

Interpretation

The similarity in decile and quintile patterns across both datasets indicates that VantageScore 4.0 and FICO Classic Score produce broadly comparable risk ladders at the population level. While VantageScore 4.0 exhibits a marginally wider spread at the decile level, this difference is not observed at broader segmentation levels, suggesting that both models provide similar differentiation for standard underwriting and portfolio applications.

The absence of risk inversion across all segments establishes a clear baseline: both models function as reliable ordinal risk-ranking tools, preserving the expected relationship between score and observed default outcomes. The analytical significance of this finding lies confirming that relative risk ordering is consistent across models, allowing subsequent analysis to focus on how risk is distributed and classified rather than whether it is identified.

Stress-Period Addendum — Rank-Ordering Stability Under COVID Conditions

This rank-ordering consistency persists under stress conditions. During the COVID-19 period (2020-2022), both models maintained monotonic relationships between score and default outcomes across all deciles, despite the presence of widespread forbearance policy interventions.

At Freddie Mac, across approximately 9.58 million stress-period originations, default rates decline consistently from D1 to D10 under both models, with no observed inversions. At Fannie Mae, a similar pattern is observed when using conditional forbearance-to-default (CondFbDef) rates, with both models showing consistent declines from the highest-risk to lowest-risk deciles.

These results indicate that the ordinal ranking property that underpins underwriting, pricing, and servicing applications remains stable even under stressed and policy-influenced conditions.

KEY TAKEAWAY

Rank-ordering performance is strong and consistent for both models. Neither VantageScore 4.0 nor FICO misorders borrower risk at any decile or quintile level in either GSE dataset. This establishes a stable analytical foundation for the report and highlights that differences between models arise not from their ability to rank relative risk, but from how that risk is calibrated, segmented, and expressed across score ranges and market conditions.

Finding 2: Conservative Calibration at the Tails

While both models align in rank-ordering borrower risk, they differ in how that risk is translated into score levels, particularly at the lower and upper ends of the distribution. For the same borrowers, VantageScore 4.0 assigns lower scores in higher-risk segments, with observed gaps of 53 to 86 points relative to FICO Classic Score below the 620 threshold. These segments exhibit realized default rates in the range of 17 to 21 percent.

At the upper end of the distribution, VantageScore 4.0 applies a higher threshold for top-tier classification, with score differences of approximately 16 to 25 points relative to FICO Classic Score. These differences reflect variation in calibration rather than differences in ordinal ranking.

Score Band	Avg VantageScore 4.0	Avg FICO	Gap (pts)	Realized Default Rate
<580 — Freddie Mac	557	643	−86 pts	21.41% (90DPD)
<620 — Fannie Mae	598	675	−77 pts	17.01% (Default)
620–659 — Freddie Mac	605	658	−53 pts	16.63% (90DPD)
780+ — Freddie Mac	809	779	+30 pts	1.44% (90DPD)
780+ — Fannie Mae	~810	~779	+16–25 pts	~0.77%

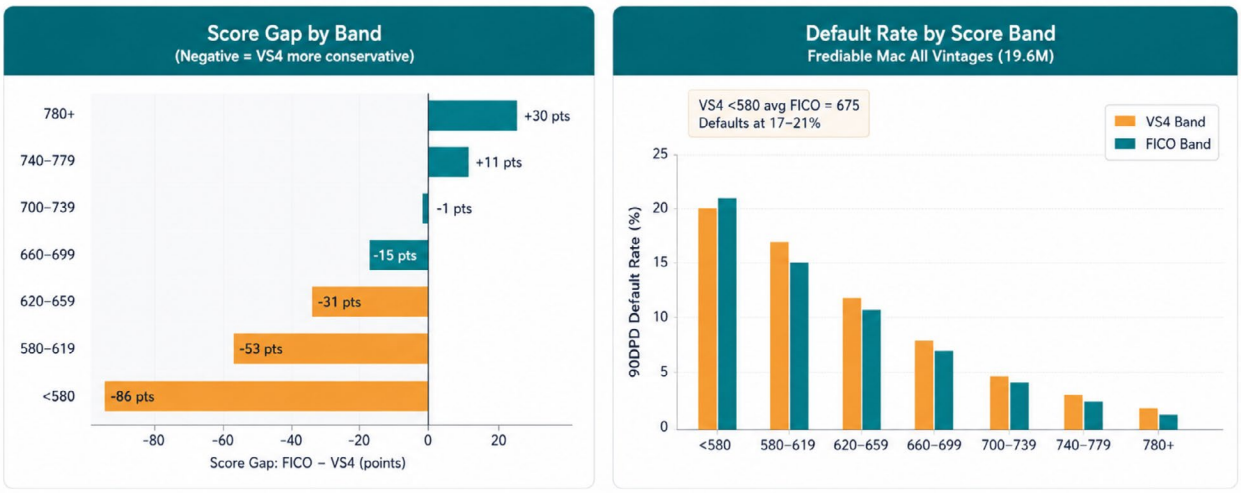
Table 2.1: Score Differences in High-Risk Segments with 17–21% Observed Default Rates (Source: Appendix C)

Threshold	VantageScore 4.0 Score Floor	FICO Score Floor
Freddie Mac — Top Decile (D10)	827+	802+
Fannie Mae — Top Decile (D10)	830+	806+
Freddie Mac — Top Quintile (Q5)	812+	793+
Fannie Mae — Top Quintile (Q5)	814+	798+

Supporting Evidence

At Freddie Mac, the <580 segment shows the largest divergence, with average VantageScore 4.0 scores of 557 compared to 643 under FICO Classic Score, an 86-point difference for the same borrowers. This cohort exhibits a 21.41 percent 90-day delinquency rate. In the 580–619 segment, the gap narrows to 53 points, with a corresponding default rate of 16.63 percent. At the upper end of the distribution, borrowers classified in the 780+ range under VantageScore 4.0 average 809 compared to 779 under FICO Classic Score, with a default rate of 1.44 percent.

Fannie Mae data shows a similar pattern. Borrowers in the <620 segment average 598 under VantageScore 4.0 and 675 under FICO Classic Score, a 77-point difference, with a realized default rate of 17.01 percent.



Interpretation

These differences indicate that the models distribute borrowers differently across score ranges, particularly at the tails of the distribution. Lower scores assigned to higher-risk segments and higher thresholds for top-tier classification reflect differences in how each model translates underlying risk into score levels.

These calibration differences are most visible in segments where default risk is concentrated, as well as in how borrowers are distributed across lower and higher score bands. As a result, borrowers with similar underlying risk profiles may fall into different score bands depending on the model used.

In a multi-score environment, these differences have practical implications for how eligibility thresholds are applied and how borrowers are segmented across risk tiers. The same borrower may be classified differently depending on the scoring framework used, not because relative risk ordering changes, but because score thresholds correspond to different portions ovf the risk distribution.

KEY TAKEAWAYS

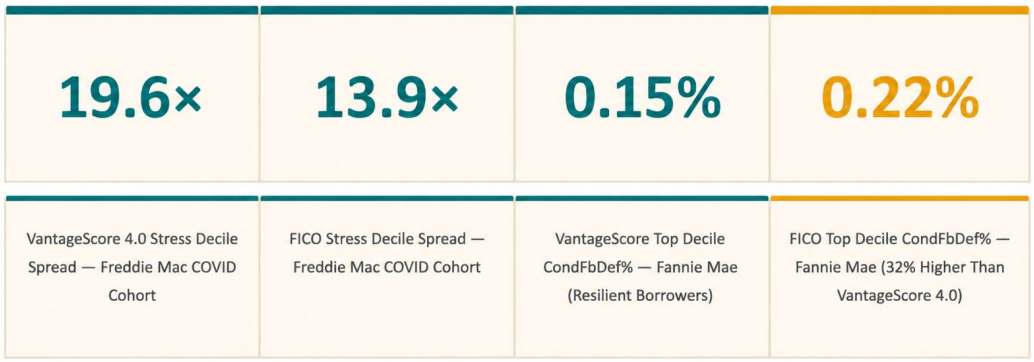
Calibration differences between scoring models are reflected in how borrowers are distributed across score ranges, particularly at the lower and upper ends of the distribution. These differences align with observed variation in default rates across segments. In a multi-score environment, **calibration plays a central role** in determining how borrowers are classified relative to eligibility thresholds, how pricing tiers are assigned, and how risk is allocated across portfolios.

Finding 3: Risk Differentiation Under Stress Conditions

Core Insight

COVID-19 forbearance programs suppressed observed defaults during 2020–2021, limiting the usefulness of raw delinquency metrics during this period. To account for this, CondFbDef% — the share of borrowers who entered forbearance and subsequently defaulted — provides a more informative measure of underlying credit performance.

Using this measure, differences in risk separation between models become more pronounced under stress conditions. At Freddie Mac, the decile spread expands from 11.5× to 19.6× for VantageScore 4.0 and from 10.1× to 13.9× for FICO Classic Score. Differences are most visible at the upper end of the distribution, where the composition of the lowest-risk segment varies across models.



Supporting Evidence

At Freddie Mac, across 9.58 million loans originated during the 2020–2022 period, the D1÷D10 90DPD spread reaches 19.6× under VantageScore 4.0 compared to 13.9× under FICO Classic Score. At the top decile (D10), observed 90DPD rates are 0.36 percent under VantageScore 4.0 and 0.46 percent under FICO Classic Score.

At Fannie Mae, among borrowers who entered forbearance, the D10 CondFbDef rate is 0.15 percent under VantageScore 4.0 compared to 0.22 percent under FICO Classic Score.

At the lower end of the distribution, results are closely aligned. The bottom quintile shows nearly identical outcomes (VantageScore 4.0: 10.05 percent; FICO Classic Score: 10.04 percent), indicating similar identification of the highest-risk borrowers.

These results reflect the importance of using a forbearance-adjusted framework. During 2020–2021, CARES Act protections prevented many distressed borrowers from reporting as delinquent, delaying the realization of default risk. CondFbDef% provides a clearer view of how that risk ultimately materialized.

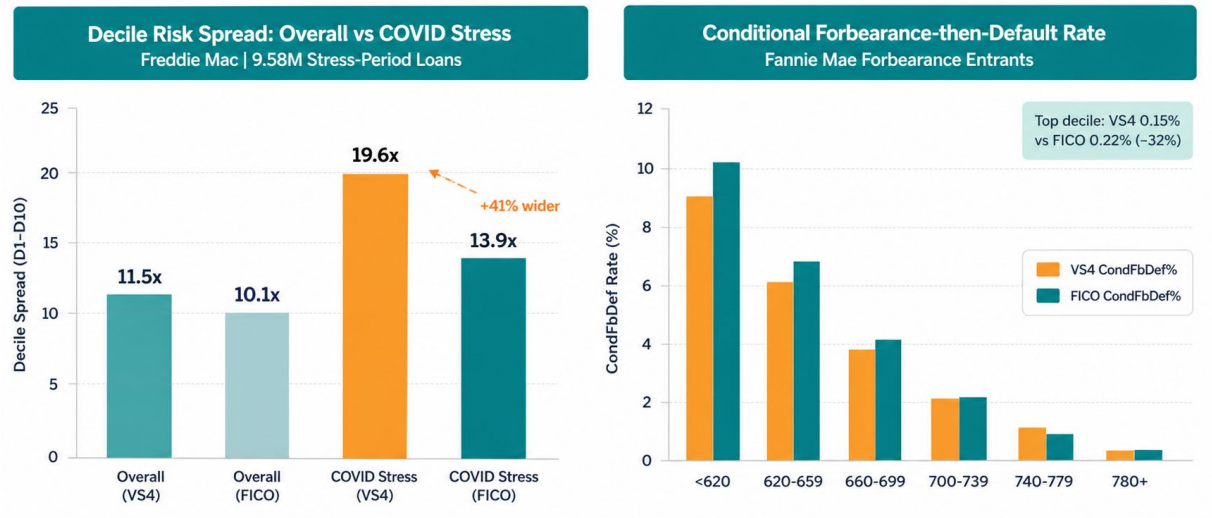


Figure 3.1: Decile spread expands during stress period across models (VantageScore 4.0: 19.6×; FICO Classic Score: 13.9×). VantageScore 4.0 Top Decile CondFbDef is 32% lower than FICO Classic Score at Fannie Mae. Source: Tables 1.5, 1.6, 2.5 (Source: Appendix D)

Interpretation

Stress conditions reveal differences in how models segment borrowers across the risk distribution, particularly at the upper end. While both models continue to rank-order borrowers consistently, the degree of separation between higher- and lower-risk segments varies under these conditions.

The observed differences are concentrated in the lowest-risk segments, where small variations in classification correspond to differences in subsequent performance among borrowers who entered forbearance. At the same time, both models show similar performance in identifying the highest-risk borrowers, suggesting that divergence is not driven by the lower end of the distribution.

This means that, in practice, differences in how models' separate risk under stress may influence expected loss outcomes, the timing and targeting of loss mitigation strategies, and how capital is allocated in stressed environments.

These findings highlight the importance of evaluating model behavior in environments where underlying risk is temporarily obscured by policy interventions. Metrics that account for delayed default realization provide a more complete view of how risk is differentiated during periods of disruption.

KEY TAKEAWAYS

Forbearance programs during the COVID-19 period suppressed observed defaults, delaying rather than eliminating underlying credit risk. As those protections unwound, borrower outcomes reflected differences in risk that were not immediately viasible in raw performance data. Under these conditions, differences in how models segment borrowers became more pronounced, particularly at the upper end of the distribution. At the decile level, this is reflected in a wider spread in risk differentiation during the stress period (19.6× vs. 13.9× at Freddie Mac, approximately 41 percent wider), while results at the lower end of the distribution remain closely aligned. These findings highlight that model evaluation based solely on stable-period performance may not fully capture how risk is differentiated when economic conditions change or when policy interventions affect observed outcomes.

Finding 4: Disagreement Cases Reveal Differences in Risk Classification

Core Insight

A subset of loans receives materially different risk classifications across VantageScore 4.0 and FICO Classic Score, providing a direct view into how differences in calibration translate into observed outcomes.

In 111,726 Fannie Mae loans where VantageScore 4.0 assigned a lower score band (<620 or 620–659) and FICO Classic Score assigned a higher score band (740+), borrowers exhibited a realized default rate of 5.97 percent, compared to a 2.81 percent baseline for the FICO Classic Score 740+ population.

At Freddie Mac, a similar pattern is observed. Borrowers in comparable disagreement cohorts received pricing consistent with their FICO Classic Score classification (e.g. 740+ bands), while realized default rates ranged from 8 to 13 percent.

In the largest Fannie Mae disagreement cohort, VantageScore 4.0 classified borrowers into lower score bands while FICO Classic Score placed the same borrowers in 740+ bands. These borrowers exhibited default rates more than double the baseline for the broader FICO Classic Score 740+ population, indicating that the lower VantageScore 4.0 classification aligned more closely with observed outcomes in this cohort. Differences in classification between VantageScore 4.0 and FICO Classic Score correspond to measurable differences in observed loan performance.

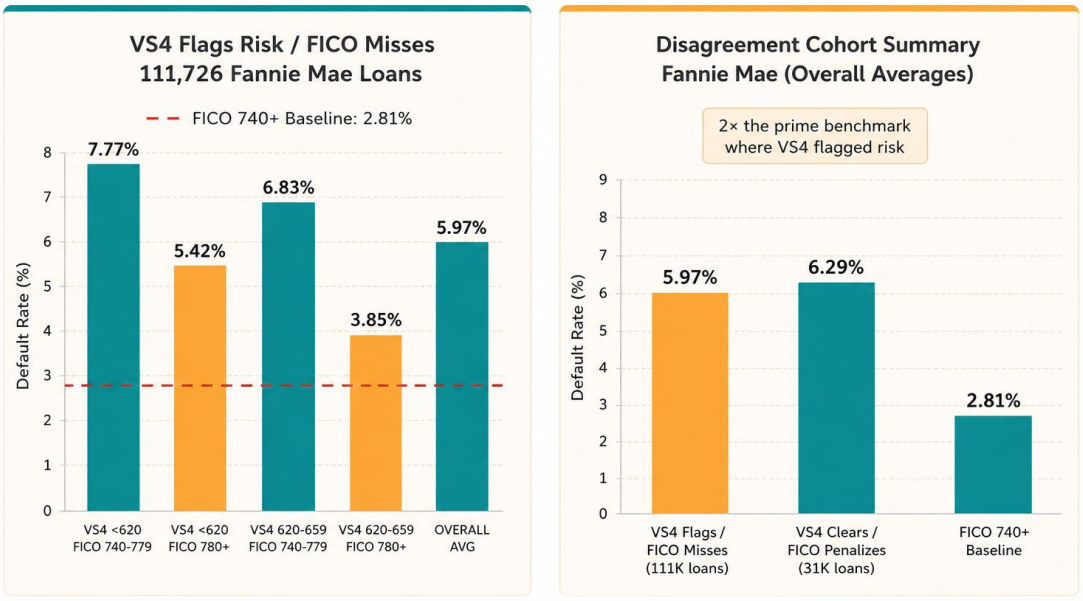


Figure 4.1: Loans with Divergent Score Band Classification Across Models

SUPPORTING EVIDENCE — FANNIE MAE

VantageScore 4.0 Band	FICO Classic Score Band	Loans	Default Rate	CondFbDef%
<620	740–779	14,805	**7.77%**	3.09%
<620	780+	8,913	**5.42%**	2.33%
620–659	740–779	62,913	**6.83%**	2.40%
620–659	780+	25,095	**3.85%**	1.27%
OVERALL	—	**111,726**	**5.97%**	**2.27%**
FICO Classic Score 740+ Benchmark	—	—	**2.81%**	—

Table 4.1: Comparison of Risk Flagging Between VantageScore 4.0 and FICO Classic Score Across 111,726 Fannie Mae Loans

SUPPORTING EVIDENCE — FREDDIE MAC

VantageScore 4.0 Band	FICO Classic Score Band	Loans	Avg VantageScore 4.0	Avg FICO Classic Score	Avg Rate	Default%	LTV	DTI
<620	740–779	7,226	598	755	3.97%	**9.01%**	79.3	48.3
<620	780+	1,587	600	794	3.86%	**4.10%**	75.1	48.1
620–659	740–779	29,976	645	754	3.95%	**7.14%**	78.4	50.2
620–659	780+	4,599	643	794	3.84%	**4.61%**	73.9	49.5
OVERALL	—	**43,388**	—	—	—	**7.07%**	—	—

Table 4.2: Freddie Mac disagreement cohort: VantageScore 4.0 averages 598–645; FICO Classic Score scores the same loans at 754–794 — a 150–195 pt divergence. Overall, the default rate of 7.07% is 2.5–3× the FICO Classic Score 740+ portfolio average. (Source: Appendix E)

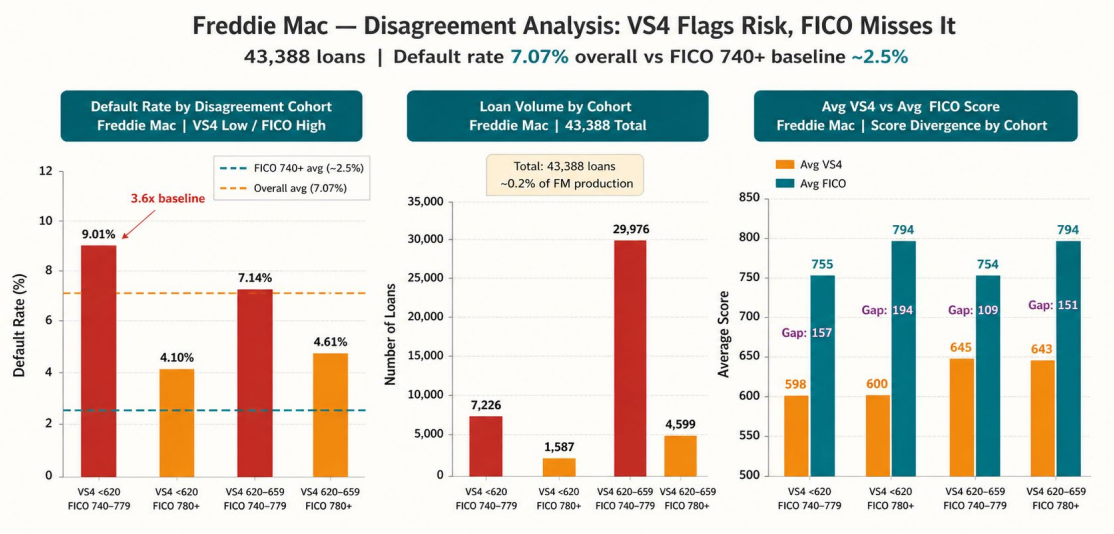


Figure 4.2: A Disagreement Cohorts: Divergent Risk Classification (VantageScore 4.0 vs. FICO Classic Score)

Interpretation

Disagreement cases represent a relatively small share of total volume, approximately 1 percent of loans, but are analytically significant because they isolate conditions where VantageScore 4.0 and FICO Classic Score produce different risk signals for the same borrower.

Across both datasets, one disagreement cohort — where borrowers are classified into higher score bands under FICO Classic Score but lower bands under VantageScore 4.0 — exhibits consistently elevated default rates relative to the baseline associated with the higher score segment. Borrowers classified differently by VantageScore 4.0 and FICO Classic Score exhibited materially different default outcomes.

The analysis above — where VantageScore 4.0 assigned borrowers to lower score bands while FICO Classic Score placed the same borrowers in higher bands — represents the larger disagreement cohort observed in this dataset. A second disagreement cohort exists and is examined separately to assess whether outcomes are consistent in the opposite direction.

In approximately 31,000 Fannie Mae loans (shown in Appendix Table E.1), VantageScore 4.0 assigned borrowers to higher score bands (e.g., 740+) while FICO Classic Score placed the same borrowers in lower bands (e.g., 620–659). This cohort represents the reverse of the primary disagreement pattern, in which the FICO Classic Score took a more favorable view of the borrower.

Observed performance outcomes for this reverse cohort are more mixed. Conditional forbearance-to-default rates were near parity across classifications (2.29 percent versus 2.27 percent), indicating that neither model classification consistently aligned more closely with realized outcomes for this group. Given the smaller sample size and similarity in observed outcomes, these results do not support a directional conclusion for this segment.

The contrast between the two disagreement cohorts indicates that disagreement is not evenly distributed across the population, and that performance implications differ depending on the direction of classification. In this dataset, the primary disagreement cohort is associated with more pronounced differences between classification and observed outcomes, while the reverse cohort does not exhibit a comparable pattern.

This asymmetry is an important feature of findings. It suggests that differences in classification between VantageScore 4.0 and FICO Classic Score do not have uniform implications across all borrowers, and that the effects of model divergence may be concentrated in specific segments of the population. In a multi-score environment, this has implications for how disagreement cases are interpreted and managed, particularly for borrowers near decision thresholds where classification differences may affect eligibility, pricing, or channel outcomes.

The pricing dimension reflects how these dynamics operate within the current system. In disagreement cases, loan pricing aligns with the score used within the existing underwriting framework in this dataset, the FICO Classic Score — even when alternative classifications assign borrowers to different risk segments. As a result, pricing does not fully reflect variation in observed outcomes across these cohorts when the models produce different classifications. This observation is specific to the current system structure and does not represent pricing behavior under a future multi-score framework.

KEY TAKEAWAYS

Disagreement cases provide a focused view of how differences in model calibration translate into observable outcomes for the same borrower. While these cases represent a small share of total volume, they highlight where model choice has the most direct influence on classification, pricing, and risk exposure. The observed asymmetry across disagreement cohorts suggests that differences in how models segment risk may have uneven implications across the population. In a multi-score environment, these cases are likely to be operationally significant, as they represent the scenarios where conflicting risk signals must be interpreted and resolved within existing decision frameworks.

Finding 5: Risk-Pricing Compression Across Both Models

Core Insight

Differences in interest rates across score segments are substantially smaller than the corresponding differences in observed default risk across both GSE datasets and scoring approaches. Compression ratios range from approximately 11.7:1 to 15.2:1, indicating that variation in pricing across the score distribution is significantly narrower than variation in realized performance.

In practical terms, this means that borrowers with significantly different levels of default risk often receive interest rates that are relatively similar, indicating that pricing does not fully reflect differences in underlying risk.

Differences in compression are observed across scoring approaches, with FICO Classic Score based pricing exhibiting higher compression ratios than VantageScore 4.0-based comparisons in certain segments. In disagreement cases, pricing tends to align with the more favorable score classification. This is a directional, observational finding and does not establish causation or prescribe specific pricing outcomes.

KEY METRICS

12.8:1
VantageScore 4.0 Compression Ratio
Freddie Mac — Risk Gradient ÷ Rate Gradient

–0.74
VantageScore 4.0 Rate Spread
D1 → D10 ppts
(Steeper Gradient)

15.2:1
FICO Classic Score Compression Ratio
Freddie Mac — More Compressed

–0.57
FICO Classic Score Rate Spread
D1 → D10 ppts
(Flatter Gradient)

SUPPORTING EVIDENCE — FREDDIE MAC & FANNIE MAE

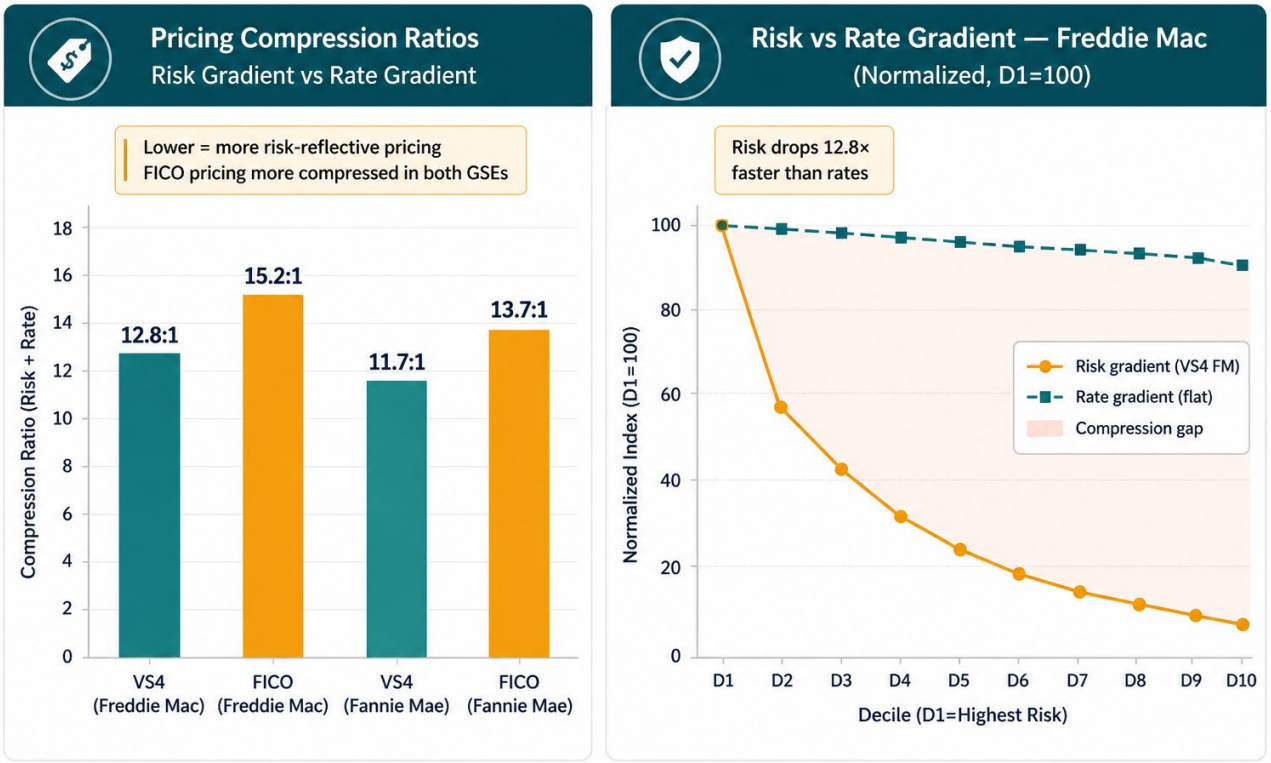


Figure 5.1: Compression ratios by model and GSE (Source: Appendix E)

Metric	Freddie Mac (VantageScore 4.0)	Freddie Mac (FICO Classic Score)	Fannie Mae (VantageScore 4.0)	Fannie Mae (FICO Classic Score)
Rate Spread D1→D10	−0.7355 ppts	−0.5719 ppts	−0.8617 ppts	−0.7201 ppts
Risk Spread D1→D10	9.45 ppts	8.69 ppts	10.06 ppts	9.88 ppts
Compression Ratio	**12.8:1**	**15.2:1**	**11.7:1**	**13.7:1**

Compression Ratio = Risk Spread ÷ Rate Spread. A ratio of 12.8:1 means risk differentiates 12.8× more steeply than pricing. FICO Classic Score-based pricing is more compressed across both datasets. Source: Appendix E

Interpretation

The compression ratios observed across both datasets indicate that pricing does not move in direct proportion to measured risk. Across all model and dataset combinations analyzed, the difference in default rates between higher- and lower-risk borrower segments is an order of magnitude larger than the corresponding difference in interest rates. This pattern appears consistently and is not specific to a single scoring approach.

Differences in compression across scoring frameworks reflect variation in how pricing aligns with each model's segmentation of the score distribution. At Freddie Mac, for example, FICO Classic Score-based pricing shows higher compression relative to VantageScore 4.0-based comparisons (15.2:1 vs. 12.8:1), though both remain substantially compressed relative to observed risk differences.

Disagreement cases provide a more focused view of this dynamic. In these instances, pricing tends to follow the more favorable score classification assigned to the borrower, even when the alternative score places that borrower in a different risk segment. This suggests that pricing may reflect one of multiple available risk signals rather than a unified view of borrower risk.

Important caveat: This analysis is based on observed origination rates and does not control for other factors that influence pricing, including loan-to-value ratios, debt-to-income ratios, product type, or market conditions. As a result, findings should be interpreted as directional rather than causal.
These differences are analytically distinct from rank-ordering performance. While both models agree on relative risk, they do not assign borrowers to equivalent positions within the score distribution. This distinction has direct implications for how risk is interpreted and acted upon in underwriting, pricing, and capital allocation decisions.

KEY TAKEAWAYS

Pricing compression appears to be a structural feature of the current mortgage market, with interest rate variation across score segments substantially narrower than differences in observed default risk.

This pattern is present across both datasets and scoringapproaches. In a multi-score environment, the presence of multiple risk signals introduces an additional dimension, particularly in disagreement cases where pricing may align with one score over another.

How pricing frameworks incorporate and reconcile these signals may influence how risk is distributed across borrowers and portfolios.

Interpretation of Findings

Taken together, the findings point to a consistent pattern: differences between VantageScore 4.0 and FICO Classic Score are less about whether risk is identified, and more about how that risk is classified, calibrated, and expressed across the score distribution.

Both models demonstrate strong rank-ordering performance across all segments. The more consequential differences emerge in calibration, stress-period behavior, and disagreement cases – particularly at the boundaries where borrower eligibility, pricing, and portfolio composition are determined.

Across both datasets, VantageScore 4.0 consistently assigns lower scores to higher-risk borrowers and requires a higher threshold for classification into top-tier classification than FICO Classic Score. At the lower end of the distribution, score differences of 50 to 80 points are observed, with borrowers in these segments exhibiting realized default rates in the range of approximately 17 percent to 21 percent. At the upper end, a higher threshold for top-tier classification results in a more tightly defined low-risk segment.

These differences are analytically distinct from rank-ordering performance. While both models agree on relative risk, they do not assign borrowers to equivalent positions within the score distribution. This distinction has direct implications for how risk is interpreted and acted upon in underwriting, pricing, and capital allocation decisions.

Stress-Period Interpretation

The COVID-19 period provides an important test of model behavior under stressed conditions, but it also introduces unique interpretive challenges.

Observed default rates during 2020–2021 were materially suppressed by widespread forbearance programs, limiting the usefulness of raw delinquency metrics as indicators of underlying credit risk. As a result, stress-period analysis relies on alternative measures, specifically the rate at which borrowers who entered forbearance subsequently defaulted. Using this lens, differences in model behavior become more pronounced. While both models continue to maintain consistent rank ordering, VantageScore 4.0 exhibits a wider separation than FICO Classic Score between higher-risk and lower-risk borrowers during the stress period. This is particularly evident in the upper end of the distribution, where borrowers identified as lowest risk demonstrate lower post-forbearance default rates, indicating wider separation among resilient borrowers.

At the lower end of the distribution, both models perform similarly in identifying borrowers most likely to experience distress. The primary difference emerges in the ability to distinguish which borrowers will ultimately avoid default despite entering forbearance. Taken together, these findings suggest that differences in calibration and segmentation may have greater consequences under adverse conditions, even when models appear broadly similar under normal market environments.

This analysis also reinforces that the apparent decline in defaults during the COVID-19 period reflects delayed rather than eliminated risk. Borrowers identified as higher risk prior to or during the pandemic were more likely to default once forbearance protections ended, indicating that underlying credit risk remained present even when not immediately observable in performance data.

Why Disagreement Cases Matter

Disagreement cases provide a focused lens into how differences in calibration translate into real-world outcomes.
Although these cases represent a relatively small share of total loan volume, they isolate instances where models assign materially different risk classifications to the same borrower. In a multi-score environment, these are the cases where decision-making is most sensitive to model choice.

Across both datasets, disagreement cases show a consistent directional pattern. Borrowers identified as higher risk by VantageScore 4.0 but placed in higher score bands (e.g. 740+) by FICO Classic Score exhibit materially elevated default rates relative to baseline for those higher FICO Classic Score bands.

This suggests that differences in score assignment are not merely theoretical, but correspond to observable differences in realized performance.

At the same time, disagreement outcomes are not perfectly symmetric. In cases where the alternative model assigns more favorable scores, performance outcomes are more mixed, reinforcing that disagreement analysis provides directional insight rather than a categorical determination of model superiority.

The importance of these findings lies not in their share of total volume, but in their implications for system design. As markets move toward environments where multiple scores are available, disagreement cases represent the points at which underwriting, pricing, and servicing decisions must reconcile conflicting signals.

Synthesis of Findings

Overall, the findings highlight that model evaluation cannot rely solely on rank-ordering performance.

While both models effectively identify relative risk, differences in calibration, segmentation, and stress-period sensitivity shape how that risk is translated into decisions and outcomes. These factors become particularly important in environments where models are used not only to rank borrowers, but to determine eligibility thresholds, pricing levels, and servicing strategies.

In this context, understanding how models behave under stress, how they classify borrowers at the margins, and how their signals diverge in key cases is essential to assessing system readiness as scoring approaches evolve.

System-Level Implications

The findings from this analysis point to a set of system-level considerations that extend beyond model performance alone. As the market moves toward a multi-score environment, differences in calibration, stress behavior, and risk signaling will shape how credit risk is evaluated, priced, and managed across the mortgage ecosystem.

Underwriting and Score Transition

As multiple scoring approaches are introduced into the underwriting process, differences in calibration create new complexity around borrower's eligibility and threshold-setting.

While both models broadly agree on relative risk, they do not consistently assign borrowers to equivalent score ranges. In practice, this means that some borrowers currently classified in higher score bands under one model may fall into lower score bands under another.

This raises a fundamental implementation question: in a multi-score environment, how should thresholds be defined and applied? If multiple scores are available but a single eligibility decision must be made, lenders and

market participants will need to determine which score governs approval, pricing, and secondary market execution.

The transition to a bi-merge or multi-score framework does not eliminate this question; it introduces it more directly. Consistency in how thresholds are applied will be important to ensure predictable borrower outcomes and stable underwriting standards.

Servicing and Loss Mitigation

Differences in model behavior under stress conditions have direct implications for servicing strategies, particularly in periods of economic disruption.

The stress-period analysis suggests that models vary in their ability to distinguish between borrowers who are likely to recover from temporary hardship and those who are more likely to transition to default. In practice, this affects how servicers prioritize outreach, allocate resources, and design loss mitigation strategies.

More precise identification of borrower resilience—particularly among those entering forbearance—can support more targeted intervention strategies, improving outcomes for both borrowers and servicers. At scale, even modest improvements in risk differentiation can translate into meaningful operational efficiencies and reduced losses.

These findings highlight that scoring models are not only underwriting tools, but also inputs into downstream servicing decisions, particularly in stress environments.

Pricing, Capital Allocation, and Investor Interpretation

The relationship between credit scores, pricing, and realized risk has implications for how capital is allocated and how portfolios are evaluated.

The observed compression between risk gradients and pricing gradients suggests that differences in borrower risk are not always fully reflected in loan pricing. In a multi-score environment, this dynamic may become more pronounced in cases where different models provide divergent signals for the same borrower.

For lenders, this raises questions about how pricing frameworks should incorporate multiple risk signals. For investors, it introduces additional complexity in interpreting loan pools, as individual loans may carry multiple risk classifications depending on the scoring approach used.

Over time, this may increase the importance of transparency around score distributions, calibration differences, and the composition of loan cohorts. Understanding how different scoring approaches map to realized performance will be important for accurately assessing risk and return.

Market Behavior in a Multi-Score Environment
The introduction of multiple scoring approaches also has the potential to influence borrower and lender behavior.

When more than one score is available, and those scores produce different outcomes, there may be incentives to favor the more favorable classification in underwriting or channel selection. This creates the potential for routing or optimization behavior, particularly among borrowers near key eligibility thresholds.

Disagreement cases provide an early indication of where these dynamics may emerge. While they represent a relatively small share of total volume, they are concentrated among borrowers whose characteristics place them near decision boundaries. In a multi-score environment, these are the borrowers most likely to experience variability in outcomes.

These dynamics introduce new considerations for market design, including consistency, transparency, and the potential need for guardrails to ensure that differences in scoring approaches do not lead to unintended distortions in borrower access or risk distribution.

Synthesis

Taken together, these implications reinforce that the transition to a multi-score environment is not solely a technical change. It is a system-level shift that affects how risk is classified, how decisions are made, and how outcomes are distributed across the market. Understanding not only how models perform, but how their differences interact with underwriting, servicing, pricing, and behavior, will be critical to ensuring that the system functions effectively under both normal and stressed conditions.

Limitations And Guardrails

While the findings presented here offer meaningful insight into how scoring models behave across different conditions, they should be interpreted within the context of several important analytical constraints. These limitations do not undermine the results, but rather define the boundaries of what this analysis can and cannot establish.

Retrospective Application of Scores

This analysis applies VantageScore 4.0 retrospectively to loans that were originated, priced, and underwritten using FICO Classic Score-based frameworks.
As a result, the findings reflect how different scoring approaches evaluate the same historical population, rather than how loans would have performed if VantageScore had been used in live underwriting decisions.
This distinction is important. The results demonstrate differences in risk classification and calibration, but they do not represent a forward-looking simulation of real-time credit decisioning or market behavior under an alternative scoring regime.

Forbearance and Suppressed Default Signals

The COVID-19 period introduces a structural limitation in interpreting performance outcomes.

Widespread forbearance programs prevented borrowers in distress from being reported as delinquent, resulting in artificially suppressed default rates during 2020–2021. As a result, raw delinquency and default measures during this period are not clean indicators of underlying credit risk.

To address this, the analysis relies on conditional forbearance-to-default (CondFbDef) metrics as a more appropriate measure of stress-period performance.
However, even this approach is constrained by limited visibility into the specific type of forbearance or loss of mitigation pathway applied to individual loans.

Limited Post-COVID Performance Window

Loans originated in the most recent periods (2022–2023) have limited performance history, particularly in the Freddie Mac dataset where VantageScore coverage does not extend through the full period.

These gaps are primarily driven by missing VantageScore observations in certain origination windows, rather than methodological exclusion. The final matched dataset represents approximately 95 percent of available observations and is considered robust for population-level analysis.

Absence of Behavioral and Underwriting Context

This analysis is based on loan-level performance data and observed credit scores at origination. It does not capture borrower-level behavioral factors such as income volatility, financial shocks, or lender-specific underwriting decisions.

As a result, the analysis cannot fully isolate the causal mechanisms that drive observed differences in performance. The findings reflect correlations between score classification and outcomes, rather than definitive causal relationships.

Disagreement Cohort Size and Interpretation

Disagreement cases, where scoring models assign materially different risk classifications to the same borrower, represent approximately 1 percent of the overall dataset.

While these cases are analytically important and provide direct insight into differences in model calibration, their relatively small size requires careful interpretation. Directional findings from these cohorts are informative but should not be extrapolated without additional validation in larger or forward-looking samples.

Pricing Analysis is Directional and Observational

The pricing analysis is based on observed origination interest rates and does not control for the full set of factors that influence loan pricing, including loan-to-value ratios, debt-to-income ratios, or broader market conditions at the time of origination.

As a result, the findings related to pricing and risk compression should be interpreted as directional. They highlight potential misalignment between pricing and observed risk, but do not establish causation or prescribe specific pricing adjustments.

Summary of Guardrails

Collectively, these limitations reinforce that:

- The analysis evaluates how scoring models classify and differentiate risk within a historical population
- It does not simulate real-time underwriting, pricing, or borrower behavior in a multi-score environment
- Observed relationships are directional and descriptive, rather than causal

Further analysis, including forward-looking modeling, expanded pricing controls, and borrower-level behavioral data, would be required to fully assess how these dynamics play out in live market conditions.

Areas For Future Research

The analysis presented here establishes a foundation for understanding how scoring models behave under consistent historical conditions. At the same time, it highlights several areas where additional research would deepen insight and support more informed implementation decisions as the market evolves.

Evaluation of FICO 10T as Historical Performance Data Becomes Available

This analysis evaluates VantageScore 4.0 and FICO Classic Score, which are the two models currently observable within GSE historical datasets. FICO 10T is not yet available with sufficient historical performance data to support comparable analysis.

As FICO 10T data becomes available, extending this analysis to include a three-model comparison would provide additional insight into how trended data influences calibration, stress-period performance, and disagreement patterns. This would help assess how a fully implemented multi-score environment may represent and differentiate risk relative to the current system.

Consumer-Level Outcomes and Credit Access

This analysis focuses on loans that originated and observed through performance. It does not capture how differences in score calibration may affect borrower's access to credit at the point of decision.

Further research could examine how borrowers who are reclassified across score bands, particularly those shifted from higher score bands (e.g., 740+) to lower bands, experience different approval outcomes, pricing, and longer-term financial trajectories. Understanding these dynamics will be important for assessing how model differences translate into real-world borrower experiences.

Expanded Pricing Analysis

The pricing findings presented here are directional and based on observed origination rates. A more comprehensive pricing study would incorporate additional loan-level characteristics, including loan-to-value, debt-to-income ratios, and lender-specific pricing frameworks.

Such analysis would allow for a more precise assessment of how risk is priced across score distributions, and whether differences in calibration across scoring models translate into consistent or divergent pricing outcomes over time.

Behavior and Routing in a Multi-Score Environment

The transition to a multi-score framework introduces new dynamics that are not observable in historical data.

When multiple scores are available, borrower and lender behavior may adapt. Lenders may make decisions about which score to rely on for underwriting, pricing, or capital treatment, while borrowers may experience different outcomes depending on channel or product selection.

Tracking these routing and selection dynamics once implementation occurs will be critical to understanding how the system functions in practice and whether it aligns with intended policy outcomes.

Performance Across the Full Credit Lifecycle

While this analysis includes a full performance window for earlier vintages, more recent originations, particularly those from 2022 onward, have limited seasoning. Revisiting this analysis as additional performance data becomes available will provide a more complete view of how scoring models perform across the full credit lifecycle, including during periods of economic normalization following stress events.

Comparison of Alternative Score Methodologies

This analysis applies a single implementation approach to VantageScore 4.0. However, multiple score construction methodologies exist, including variations in how bureau data is combined.

A structured comparison of these methodologies against observed performance outcomes would provide additional clarity on how implementation choices influence risk classification and downstream results.

Looking Ahead

These areas suggest that while model behavior can be evaluated through historical analysis, the full implications of score transition will depend on how models are implemented, interpreted, and acted upon in practice.

As the market moves forward, continued analysis, particularly incorporating real-time underwriting decisions, pricing frameworks, and borrower outcomes, will be essential to fully understanding how these dynamics shape credit access, risk management, and market stability.

Conclusion

This analysis draws on one of the largest matched loan-level datasets available across Fannie Mae and Freddie Mac, covering approximately 44.7 million loans and a period that includes both stable market conditions and the COVID-19 stress event. The findings provide a detailed view of how VantageScore 4.0 and FICO Classic Score classify and differentiate borrower risk within the current mortgage system.

Across the full population, both models consistently rank-order borrower risk, producing clear and monotonic relationships between score and default outcomes. The primary differences observed in this analysis emerge not in the identification of relative risk, but in how that risk is calibrated, how it is segmented across the score distribution, and how it performs under stress conditions.

The COVID-19 period underscores the importance of these distinctions. While policy interventions provided the forbearance programs that temporarily suppressed observable defaults, underlying risk remained and was revealed over time. Under these conditions, differences in risk separation became more visible, particularly in how models distinguish among lower-risk borrowers and identify resilience within the highest score segments.

Taken together, these results suggest that model evaluation cannot rely solely on rank-ordering performance. Calibration, segmentation, and stress-period behavior are central to how risk is translated into underwriting decisions, pricing outcomes, servicing strategies, and investor expectations. These dynamics are particularly relevant as the market transitions from a single-score framework to one in which multiple risk signals must be interpreted and applied.

This analysis reflects performance within the current system, where loans were originated, priced, and managed using FICO Classic Score-based underwriting and pricing frameworks. The observed pricing patterns, disagreement outcomes, and risk segmentation reflect how these VantageScore 4.0 and FICO Classic Score interact within that structure, rather than how a fully implemented multi-score system would function. Newer models, including FICO 10T, are not yet available with sufficient historical GSE data to support comparable analysis and represent an important area for future research.

The implications of these findings vary across participants in the mortgage system. For originators, differences in calibration and disagreement cases introduce complexity in how borrowers are evaluated against eligibility thresholds and how loans are priced. When multiple scores assign different classifications to the same borrower, the choice of which score governs decision-making has direct implications for loan approval, pricing consistency, and operational processes.

For investors, these differences affect how risk is distributed within loan pools and how performance expectations are formed. Disagreement cases, in particular, highlight segments where risk classification varies across models and where observed outcomes may differ from expectations based on a single score. As multiple scores are incorporated into market practice, transparency around score distributions and disagreement patterns may become increasingly important for assessing credit risk.

For borrowers, the introduction of multiple scoring approaches may affect both access to credit and the terms offered across channels. Some borrowers may qualify under one model and not another or may be assigned to different pricing tiers depending on the score used. These differences reflect variations in how risk is classified rather than changes in underlying borrower behavior, and their effects will depend on how consistently and transparently multi-score decisions are implemented across lenders and channels.

For the broader system, including taxpayers who bear residual risk through the GSE structure, the findings on pricing compression and risk segmentation highlight the importance of alignment between risk signals, pricing frameworks, and capital allocation. The introduction of additional risk measures may improve how risk is identified, but outcomes will depend on how those signals are incorporated into underwriting and pricing decisions.

The mortgage market is not simply adding another credit score. It is introducing an additional representation of risk into a system that was designed around a single primary signal. This shift requires adaptation across underwriting thresholds, pricing frameworks, servicing strategies, and investor disclosure practices.

This report does not prescribe a specific implementation approach. Instead, it provides an empirical foundation for understanding how scoring models behave in practice and where differences in that behavior have the greatest practical significance. As implementation evolves, continued analysis will be essential to assess how these dynamics unfold under live market conditions and how multiple risk signals are ultimately integrated into decision-making across the system.

TECHNICAL APPENDIX

Supporting Data Tables

This appendix provides a complete reference to all primary analytical tables underlying the five findings.

APPENDIX A — DATA & METHODOLOGY

A.1 Dataset Overview

Dataset	Loans	Coverage	Match Rate	Primary Metric
Freddie Mac	19.6M	2013–2022	95.4%	90+ DPD (ever)
Fannie Mae	25.1M	2013–2023	95.0%	Ever 90+ DPD / Default
Combined	44.7M	2013–2023	~95%	CondFbDef (stress)

Table A.1: Dataset Overview

A.2 Fannie Mae Data Sources & Coverage

Element	Detail
Performance Data	Fannie Mae Single-Family Loan Performance — 44 quarterly ZIPs, 2013Q1–2023Q4, ~29 GB compressed, 1.37B monthly servicing rows, 26.4M unique loans
VantageScore 4.0 File	FNM_VANTAGESCORE 4.0_HLP.zip — Current Method and Tri-Merge scores, coverage 2013Q2–2023Q1
Total Loans	26,396,019 unique loans in performance dataset
Matched Loans	25,081,784 loans matched to VANTAGESCORE 4.0 score (95.0% match rate)
Unmatched Loans	1,314,235 — entirely attributable to VANTAGESCORE 4.0 file coverage gaps (2013Q1 and 2023Q2–Q4 originations absent from VANTAGESCORE 4.0 source file)
Stress Window	January 2020 – December 2022 (YYYYMM: 202001–202212)
Score Breakpoints	Standard FHFA GSE bands: <620, 620–659, 660–699, 700–739, 740–779, 780+
Analytical Unit	Pre-built loan-level parquet (lifetime flags collapsed — one row per loan)

A.3 Freddie Mac Data Sources & Coverage

Element	Detail
Performance Data	Freddie Mac Single-Family Loan Performance — annual ZIPs, 2013–2023, ~16.6 GB raw, 20.5M unique loans
VantageScore 4.0 File	FRE_VANTAGESCORE 4.0_SFLLD_Historical.txt — Current Method, TriMerge, BiMerge scores, coverage 2013Q1–2022Q4
Total Loans	20,516,977 unique loans in performance dataset
Matched Loans	19,573,258 loans matched (95.4% match rate)
Per-Year Match Rate	99.9% in every year 2013–2022 without exception
Unmatched Loans	~943,719 — 100% are 2023 originations (VANTAGESCORE 4.0 file contains zero F23 loan identifiers, confirmed by direct query)
Score Distribution	Avg VANTAGESCORE 4.0 763.9, Avg FICO CLASSIC SCORE 750.6 across full matched dataset (2013–2022)
Stress Window	January 2020 – December 2022 — matching Fannie Mae definition
Score Breakpoints	Extended lower-end bands: <580, 580–619, 620–659, 660–699, 700–739, 740–779, 780+
Analytical Unit	Origination and servicing data stored separately; outcome metrics derived at analysis time from raw monthly servicing rows

A.4 Primary Outcome Metrics

Metric	Definition	Primary Use
90DPD % (ever default)	Loan ever reached 90+ days past due over its observed life	Primary default measure — overall and stress periods
30DPD %	Loan ever reached 30+ days past due	Early delinquency / payment stress signal
FbDef % (ever)	Loan ever entered forbearance AND subsequently defaulted	Combined stress measure
CondFbDef %	Among loans that entered forbearance, % that subsequently defaulted (FbDef ÷ FbEntry)	Primary stress metric — corrects for forbearance suppression of defaults
FbEntry %	% of loans that entered any forbearance plan	Forbearance uptake signal
Composite Rate %	Ever default OR ever forbearance	Broadest stress exposure measure (Fannie Mae)

APPENDIX B — FINDING 1: RANK-ORDERING PARITY

Table B.1 — Fannie Mae Decile Risk Ladder, VantageScore 4.0 vs. FICO Classic Score

(25.1M Matched Loans)

Dec	VantageScore 4.0 Range	Avg VantageScore 4.0	Default%	FbDef%	FICO Classic Score Range	Avg FICO Classic Score	Default%	FbDef%
D1	383–688	658	10.80%	4.09%	441–690	665	10.65%	4.08%
D2	688–715	703	6.42%	2.25%	690–716	704	6.54%	2.25%
D3	715–736	726	4.58%	1.52%	716–737	727	4.79%	1.55%
D4	736–759	748	3.45%	1.04%	737–754	745	3.59%	1.07%
D5	759–778	769	2.77%	0.74%	754–767	761	2.67%	0.72%
D6	778–791	785	2.08%	0.51%	767–779	773	2.03%	0.49%
D7	791–801	796	1.59%	0.37%	779–789	784	1.56%	0.35%
D8	801–814	808	1.29%	0.28%	789–798	793	1.24%	0.27%
D9	814–830	822	1.12%	0.22%	798–806	802	0.99%	0.21%
D10	830–850	838	0.74%	0.14%	806–850	812	0.77%	0.19%
Spread	D1÷D10	VantageScore 4.0 14.6x	FbDef% (D1÷D10)		FICO Classic Score 13.8x	Zero inversions		

Table 1.1: D1 = lowest scores (highest risk); D10 = highest scores (lowest risk). VantageScore 4.0 risk spread (D1÷D10): 14.6x. FICO Classic Score risk spread: 13.8x.

Zero risk inversions.

Table B.2 — Freddie Mac: Decile Risk Ladder, VantageScore 4.0 vs. FICO Classic Score

(19.6M Matched Loans, 2013–2022)

Dec	VantageScore 4.0 Range	Avg VantageScore 4.0	90DPD%	30DPD%	FICO Classic Score Range	Avg FICO Classic Score	90DPD%	30DPD%
D1	357–689	658	10.35%	26.81%	300–685	659	9.64%	27.01%
D2	689–715	703	5.92%	18.51%	685–709	698	6.07%	18.73%
D3	715–736	726	4.31%	15.00%	709–729	719	4.53%	15.08%
D4	736–759	748	3.35%	12.19%	729–745	737	3.59%	12.83%
D5	759–777	768	2.72%	10.21%	745–760	753	2.80%	10.76%
D6	777–789	783	2.09%	8.61%	760–772	766	2.20%	9.20%
D7	789–799	794	1.68%	7.71%	772–783	778	1.72%	7.86%
D8	799–812	805	1.42%	7.28%	783–793	788	1.38%	6.79%
D9	812–827	819	1.28%	7.29%	793–802	798	1.12%	5.99%
D10	827–850	836	0.90%	6.23%	802–850	810	0.95%	5.59%

VantageScore 4.0 risk spread (D1÷D10 90DPD): 11.5×. FICO Classic Score risk spread: 10.1×.

VantageScore 4.0 produces a wider risk ladder at the overall level.

Table B.3 — Decile Spread Summary — Both GSEs

Metric	Freddie Mac — VantageScore 4.0	Freddie Mac — FICO Classic Score	Fannie Mae — VantageScore 4.0	Fannie Mae — FICO Classic Score
Decile Spread (D1÷D10, 90DPD / Default)	11.5x	10.1x	14.6x	13.8x
Quintile Spread (Q1÷Q5)	7.5x	7.6x	9.3x	9.8x
Risk Inversions Across Deciles	Zero	Zero	Zero	Zero
Bottom Decile (D1) Default Rate	10.35%	9.64%	10.80%	10.65%
Top Decile (D10) Default Rate	0.90%	0.95%	0.74%	0.77%
Top Decile Score Floor	827+	802+	830+	806+
VantageScore 4.0 High Score Threshold Premium (vs FICO Classic Score)	+25 pts	—	+24 pts	—

Table 1.1: D1 = lowest scores (highest risk); D10 = highest scores (lowest risk).
VantageScore 4.0 risk spread (D1÷D10): 14.6x. FICO Classic Score risk spread: 13.8x. Zero risk inversions.

APPENDIX C — FINDING 2: CONSERVATIVE CALIBRATION

Table C.1 — Freddie Mac: Risk Profile by VantageScore 4.0 / FICO Classic Score Band (All Vintages, 19.6M Loans)

Band	Loans (Count)	Avg VantageScore 4.0	Avg FICO Classic Score	90DPD% (VantageScore 4.0)	90DPD% (FICO Classic Score)	30DPD% (VantageScore 4.0)	30DPD% (FICO Classic Score)
<580	37,560	557	643	21.41%	21.67%	44.88%	51.03%
580–619	146,699	605	658	16.63%	15.44%	37.92%	40.72%
620–659	645,176	644	675	11.68%	11.29%	29.22%	30.65%
660–699	1,813,253	683	698	7.75%	7.37%	22.12%	21.86%
700–739	3,506,727	720	721	4.69%	4.54%	15.85%	15.10%
740–779	4,034,785	761	750	2.91%	2.45%	10.84%	9.82%
780+	9,389,058	809	779	1.44%	1.19%	7.35%	6.26%

VantageScore 4.0 <580 avg = 557; FICO Classic Score scores the same loans at 643 — an 86-point gap confirmed by 21.41% 90DPD.
Freddie Mac uses extended lower-end bands, revealing additional variation within the lowest score ranges (e.g., below 620).

Table C.2 — Fannie Mae: Risk Profile by FICO Classic Score Band (All Vintages, 26.4M Loans)

Score Band	Loans (Count)	% of Total	Default Rate%	30+ DPD%	Forbearance%	Composite% (Default OR Forbearance)	FbDef% (Forbearance then Default)
<620	4,875	0.0%	11.32%	32.59%	94.91%	95.51%	6.97%
620–659	877,262	3.3%	13.00%	32.91%	79.12%	79.93%	5.08%
660–699	2,536,753	9.6%	8.57%	23.87%	80.35%	80.80%	3.08%
700–739	4,791,782	18.2%	5.15%	16.28%	82.27%	82.48%	1.66%
740–779	7,714,351	29.2%	2.61%	10.42%	83.72%	83.82%	0.69%
780+	10,455,271	39.6%	1.11%	6.35%	84.73%	84.77%	0.24%

Composite rate = ever default OR forbearance. FbDef% = forbearance then default.

APPENDIX D — FINDING 3: STRESS-PERIOD PERFORMANCE

Table D.1 — Freddie Mac: Decile Risk Ladder During COVID Stress (2020–2022 Originations)

Dec	VantageScore 4.0 Range	Avg VantageScore 4.0	90DPD%	60DPD%	FICO Classic Score Range	Avg FICO Classic Score	90DPD%	60DPD%
D1	383–695	665	7.06%	9.91%	300–688	665	6.40%	9.19%
D2	695–721	709	3.72%	5.40%	688–713	701	3.73%	5.38%
D3	721–743	732	2.52%	3.72%	713–732	723	2.66%	3.88%
D4	743–766	755	1.84%	2.71%	732–749	741	2.03%	2.99%
D5	766–781	774	1.43%	2.12%	749–762	756	1.51%	2.29%
D6	781–792	787	1.03%	1.58%	762–774	768	1.15%	1.73%
D7	792–801	797	0.82%	1.28%	774–784	779	0.86%	1.32%
D8	801–814	808	0.67%	1.08%	784–794	789	0.68%	1.07%
D9	814–830	822	0.57%	0.97%	794–803	799	0.54%	0.88%
D10	830–850	837	0.36%	0.67%	803–850	810	0.46%	0.78%

VantageScore 4.0 stress spread 19.6× (D1+D10 90DPD) vs. FICO Classic Score 13.9× — ~41% wider risk separation under COVID stress. Zero rank inversions.

Table D.2 — Fannie Mae: Decile Risk Ladder During COVID Stress (Forbearance Entrants)

Dec	VantageScore 4.0 Range	Avg VantageScore 4.0	Def% Stress	CondFbDef% (FbDef + FbEntry)	FICO Classic Score Range	Avg FICO Classic Score	Def% Stress	CondFbDef% (FbDef + FbEntry)
D1	383–690	659	12.70%	5.11%	441–691	666	12.50%	5.10%
D2	690–717	704	7.40%	2.73%	691–717	705	7.58%	2.72%
D3	717–738	727	5.24%	1.81%	717–738	728	5.49%	1.84%
D4	738–761	750	3.96%	1.23%	738–755	747	4.06%	1.25%
D5	761–780	771	3.15%	0.88%	755–768	762	3.02%	0.84%
D6	780–792	787	2.36%	0.59%	768–780	774	2.31%	0.57%
D7	792–804	798	1.79%	0.43%	780–789	785	1.78%	0.41%
D8	804–815	809	1.46%	0.33%	789–798	794	1.41%	0.31%
D9	815–831	824	1.25%	0.24%	798–807	802	1.12%	0.25%
D10	831–850	839	0.79%	0.15%	807–850	813	0.85%	0.22%

VantageScore 4.0 D10 **CondFbDef: 0.15%** vs. FICO Classic Score 0.22% — 32% lower at the safest tier. Perfect monotonic ordering throughout the stress period.

Table D.3 — Freddie Mac: Score Band Performance, Stress Period (2020–2022 Originations)

Band	Loans (Count)	Avg VantageScore 4.0	Avg FICO Classic Score	90DPD% (VantageScore 4.0)	90DPD% (FICO Classic Score)	60DPD% (VantageScore 4.0)	60DPD% (FICO Classic Score)
<580	12,542	558	669	16.40%	15.38%	21.73%	15.38%
580–619	51,611	605	668	12.91%	10.07%	17.43%	14.60%
620–659	258,074	644	677	8.72%	8.14%	12.14%	11.70%
660–699	780,425	683	698	5.50%	4.97%	7.85%	7.12%
700–739	1,615,527	720	720	3.09%	2.81%	4.52%	4.08%
740–779	1,936,932	761	749	1.71%	1.37%	2.52%	2.05%
780+	4,929,383	809	779	0.71%	0.60%	1.14%	0.96%

VantageScore 4.0 shows stronger performance across all score bands during the stress period.
Largest 90DPD advantage occurs in the lowest band (<580): **16.40%** vs. **15.38%** (1.02 pp).

APPENDIX E — FINDINGS 4 & 5: DISAGREEMENT ANALYSIS & PRICING

Table E.1 — VantageScore 4.0 Clears Borrowers That FICO Classic Score Does Not: Fannie Mae (31,016 Loans)

VantageScore 4.0 Band	FICO Classic Score Band	Loans	Default Rate	CondFbDef%
740–779	620–659	23,328	7.18%	2.81%
780+	620–659	7,688	5.41%	1.77%
OVERALL	**—**	**31,016**	**6.29%**	**2.29%**

The VantageScore 4.0-cleared group (31K loans) is ~4× smaller than the VantageScore 4.0-flagged group (111K loans).
CondFbDef near-parity (2.29% vs. 2.27%).
Caution warranted, not a directional verdict.

Table E.2 — Freddie Mac Disagreement Cohorts: Pricing vs. Risk Divergence

Cohort	Loans	Avg VantageScore 4.0	Avg FICO Classic Score	Avg Rate	Yr-Adj Rate	90DPD%	LTV/DTI
VantageScore 4.0 <620 / FICO Classic Score 700–739	22,461	**596**	716	4.12%	+0.194	13.36%	79.1/59.3
VantageScore 4.0 <620 / FICO Classic Score ≥ 740+	8,813	**598**	762	3.95%	+0.053	8.12%	78.5/48.3
VantageScore 4.0 ≥ 740+ / FICO Classic Score <660	34,066	767	647	3.87%	+0.136	5.65%	67.1/137.0
VantageScore 4.0 ≥ 780+ / FICO Classic Score <620 (*)	171	803	598	4.05%	+0.092	7.02%	66.5/643.9
Agreement Zone (Baseline)	9,818,075	777	767	3.68%	−0.035	2.70%	71.6/67.7

VantageScore 4.0 780+ / FICO Classic Score <620: 171 loans, DTI = 643.9 — treat as data anomaly.
Agreement zone (baseline) = both scores assign the same band; pricing/risk benchmark. **Yr-Adj** = year-adjusted rate vs. annual average.

Table E.3 — Compression Ratio Analysis: Freddie Mac vs. Fannie Mae

Gradient Metric	Freddie Mac	Fannie Mae (Reference)
VantageScore 4.0 Rate Spread D1→D10	−0.7355 ppts	−0.8617 ppts
FICO Classic Score Rate Spread D1→D10	−0.5719 ppts	−0.7201 ppts
VantageScore 4.0 Risk Spread D1→D10	9.45 ppts (90DPD)	10.06 ppts (Default)
FICO Classic Score Risk Spread D1→D10	8.69 ppts (90DPD)	9.88 ppts (Default)
Compression Ratio — VantageScore 4.0	**12.8:1**	**11.7:1**
Compression Ratio — FICO Classic Score	**15.2:1**	**13.7:1**

Compression ratio = Risk Spread ÷ Rate Spread. Higher ratio = more compressed pricing.
FICO Classic Score is more compressed than VantageScore 4.0 in both datasets.
FRE metric: 90DPD; FNM metric: ever default.

References

1. "Single Family Loan-Level Dataset," www.freddiemac.com, n.d., https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset. See also "Fannie Mae Single-Family Loan Performance Data | Fannie Mae," capitalmarkets.fanniemae.com, n.d., https://capitalmarkets.fanniemae.com/credit-risk-transfer/single-family-credit-risk-transfer/fannie-mae-single-family-loan-performance-data.

2. "Credit Scores | FEDERAL HOUSING FINANCE AGENCY," FHFA.gov, July 11, 2024, https://www.fhfa.gov/policy/credit-scores.

3. *Ibid.*

4. Amit Seru, "On Target: Debt Forbearance Policies Help Curb Pandemic Financial Woes | Stanford Institute for Economic Policy Research (SIEPR)," siepr.stanford.edu, April 2021, https://siepr.stanford.edu/publications/policy-brief/target-debt-forbearance-policies-help-curb-pandemic-financial-woes.

5. "Credit Scores | FEDERAL HOUSING FINANCE AGENCY," FHFA.gov, July 11, 2024, https://www.fhfa.gov/policy/credit-scores.

6. Ricardo Nunez Magana, "Cracking the Tape: What You Need to Know about VantageScore 4.0," Milliman.com, August 24, 2024, https://www.milliman.com/en/insight/cracking-the-tape-vantage-score-4.