

DOI: 10.1111/jmcb.13022

ORLA MCCULLAGH
MARK CUMMINS 
SHEILA KILLIAN

The Fundamental Review of the Trading Book: Implications for Portfolio and Risk Management in the Banking Sector

The Fundamental Review of the Trading Book (FRTB) is the promised overhaul of bankmarket risk regulation. FRTB retains the authorized use of proprietary risk models, however, it introduces two additional criteria: (i) P&L attribution (PLA) tests and (ii) desk-level backtests. We examine empirically whether these additional criteria influence risk management and portfolio management practice, specifically portfolio construction and choice of risk model. We find that the PLA tests demand significant alignment with risk factors, however, the backtests do not incentivize use of superior risk models. This has important implications for the efficacy of the capital-based regulatory system.

JEL codes: G11, G17, G21, G28

Keywords: Basel III, Fundamental Review of the Trading Book, market risk, portfolio management, value-at-risk, P&L attribution tests

THE FAILINGS OF BANK-MARKET RISK management were laid bare in the events of the financial crisis (2007-09). Indeed, during this turbulent time, bank capitalization proved an important factor for resilience. Demirguc-Kunt, Deguchi, and Merrouche (2013) provide evidence that better capitalized banks ex-

ORLA MCCULLAGH is with the Department of Accounting and Finance, Kemmy Business School, University of Limerick (E-mail: orla.mccullagh@ul.ie). MARK CUMMINS is with Financial and Operational Performance Group, Business School, Dublin City University. SHEILA KILLIAN is with Department of Accounting and Finance, Kemmy Business School, University of Limerick.

Received September 11, 2020; and accepted in revised form March 31, 2022.

Journal of Money, Credit and Banking, Vol. 00, No. 0 (January 2023)

© 2023 The Authors. *Journal of Money, Credit and Banking* published by Wiley Periodicals LLC on behalf of Ohio State University.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

perienced higher stock returns during this period, with higher quality capital holdings (Tier 1 capital and tangible common equity) being more relevant. The Fundamental Review of the Trading Book (FRTB) is the regulatory response of the Basel Committee for Banking Supervision (BCBS) to these market risk management failings. The headline change to the market risk regulatory framework under FRTB is the replacement of value-at-risk (VaR) by expected shortfall (ES) for calculating capital requirements, although VaR remains a central metric for backtesting. As described by Gordy and McNeil (2018, p. 3) of the Federal Reserve: “although estimates of ES will be the cornerstone of the risk capital calculation, the risk model approval process will continue to be based on VaR estimates and VaR exceedances.” FRTB will significantly increase regulatory capital requirements, primarily to ensure the stability of the financial system and substantially act as a valve for the availability of credit to the economy. Our paper contributes to the emerging discourse on FRTB through a quantitative analysis of its impact on bank-market risk management and portfolio management practice.

The prerogative for banks to develop their own internal models, including their choice of risk-forecasting models, remains central to FRTB, with the BCBS arguing that this is essential to enable a level playing field between banks in different jurisdictions (BCBS 2019). A key concern with the continued use of internal models across the banking system is the level of variation in the results from banks’ proprietary internal models (Beder 1995, Pritsker 1997, Berkowitz and O’Brien 2002, BCBS 2018). Indeed, VaR became a cynosure for criticism of bank risk taking and undercapitalization in the 2007–09 financial crisis (Nocera 2009, O’Brien and Szerszeń 2017) because of (i) the alarming variability of VaR-implied capital requirements, (ii) the perceived ease with which VaR can be gamed (Danielsson and Zhou 2016, Armstrong and Brigo 2019), and (iii) the issue that VaR is a point estimate that does not quantify losses in the tail. FRTB aims to address these key concerns while ostensibly retaining the role of proprietary internal risk models as a means of enabling level playing field competition and risk-sensitive capital. However, it undergirds their use with additional restrictions.

FRTB introduces two additional criteria for the qualified use of an internal model approach (IMA) in determining market risk regulatory capital. These two criteria are (i) P&L attribution (PLA) tests and (ii) desk-level backtests. These additional criteria have the potential to change the nature of the use of proprietary risk models in the determination of market risk capital. In this context, we formally assess these two criteria and provide new insights of relevance for banking practitioners and regulators.

The 2018 McKinsey report on FRTB (Azoulay et al. 2018b) depicts practitioners’ views of FRTB as requiring merely a change of statistical metric, namely, a move from VaR to ES. Contemporary financial media, in their evaluation of FRTB, have focused much of their attention on the PLA tests (Nield 2017), which have been revised since the original FRTB publication¹ (Mahfoudhi 2018). However, as impact

1. The original specifications of the PLA test incorporated the joint test of mean and variance of the RTPL and the HPL: (BCBS 2016)

analyses are published (EBA 2019), the seismic shift implied by FRTB is being recognized. According to the European Banking Authority, the expected average impact of FRTB on Pillar 1 market risk capital is an increase of 81%, with an interquartile range of 32% to 140% (EBA 2019). This excludes the application of the output floor.²

Despite this context, there are a limited number of studies examining FRTB to date. This is likely due to its recent development and postponed implementation. In March 2020, the Basel Committee's oversight body, the Group of Central Bank Governors and Heads of Supervision (GHOS), approved a revised implementation schedule (now January 2023) to facilitate banks and supervisors in the management of financial stability issues arising from the impact of COVID-19 on the banking system (ICMA 2020). Soobraty, Stern, and Cheng (2020) find that regulators view this postponement as a pragmatic respite, though assert that banks must move forward with their implementation plans. They further argue that the feasibility of banks' adoption of IMA hinges on their preparedness, including quantitative analysis and backtesting. This reinforces the relevance, timeliness, and prescience of our quantitative impact study of the FRTB framework.

One of the first FRTB studies is that of Thompson, Luo, and Fergusson (2016, 2017)³ who examine the theoretical failure rate of the PLA tests (based on the January 2016 specification). Subsequently, Farag (2017) argues that alignment between front office pricing models and middle office risk models will be key to passing the PLA tests. Correspondingly, Farag (2018) examines anomalies and asymmetries in the proposed FRTB SA and IMA framework that could be problematic upon implementation. Additionally, Pederzoli and Torricelli (2019) provide a useful illustration of the impact of both FRTB's standardized approach (SA) and IMA on a stylized portfolio. The calculative requirements for SA are radically overhauled in FRTB, establishing greater risk sensitivity in the capital calculations. FRTB requires that IMA and SA are ran in parallel, which presents significant challenges to banks' information technology systems. Indeed, Li and Xing (2018) examine different computational approaches to align internal capital allocation to FRTB regulatory capital calculation. In the context of the postponed implementation of FRTB due to the COVID-19 pandemic, Lazar and Zhang (2020) quantitatively analyze the impact of market reaction to the COVID-19 pandemic under a stylized interpretation of the FRTB-prescribed ES measure. They find that the measure generally overestimates the risk, driving increases in regulatory capital, and that assets with longer liquidity horizons display disproportionately high capital requirements.

2. The output floor limits the benefits achieved under the IMA relative to the SA (SA). BCBS argue that the output floor will strengthen the principle of the level playing field between SA and IMA banks, improve the comparability of disclosures, and enhance the credibility of capital calculations (BCBS 2017). Capital requirements will be calculated as the higher of: (i) capital calculated using the IMA (where the bank has approval for their use) and (ii) 72.5% of the capital requirements calculated under the standardized (or simplified standardized, where appropriate) approach.

3. Thompson et al. (2016) working paper was used for immediacy and subsequently augmented by the 2017 published paper.

We add to this emerging literature with our quantitative impact study, building on an established literature examining the role of VaR within the Basel II capital regulation framework (e.g., the studies by Borio 2003, Angelidis and Degiannakis 2009, Rossignolo, Fethi, and Shaban 2012, Burchi 2013). The comprehensive nature of the FRTB framework for the calculation of market risk capital addresses many of the key concerns raised by industry and commentators. Given that proprietary risk models remain central to the FRTB IMA framework subject to additional IMA criteria, an investigation of their potential impact is imperative. Our research questions pertaining to FRTB are therefore as follows:

- (i) What is the impact of the additional IMA criteria on portfolio management practice?
- (ii) What is the impact of the additional IMA criteria on risk management practice, and do they incentivize the use of superior risk models?

For IMA banks, internal risk models will remain as a conduit between portfolio and risk management and capital requirements. Therefore, these additional criteria have the potential to influence both risk management and portfolio management practice. There have not been any prior studies that have specifically examined the impact of FRTB in this way. We address this gap in the literature.

In respect of our first research question, we design an empirical study to examine how stylized market capitalization–based equity portfolios with graduated degrees of portfolio characteristics perform under the FRTB framework. We then extend this stylized analysis by means of examining a range of industry portfolios, drawing from a suite of publicly investable Exchange Traded Funds (ETFs). This allows us to consider alternative real-world portfolio constructions that move beyond market capitalization–based weighting, and that employ more active management-type strategies. The PLA tests are designed to incentivize the alignment between front office trading data and middle office risk management modeling data (BCBS 2019). There were significant industry concerns regarding the introduction of PLA tests, particularly following the high failure rate of the original specification (normalized mean and variance ratios of the unexplained P&L left between the Risk Theoretical P&L [RTPL] and Hypothetical P&L [HPL]⁴) and its treatment of hedged portfolios (Mokhtari et al. 2018). The revised PLA tests are similar to those proposed by Spinaci et al. (2017): a combination of a correlation test and distribution test designed to assess the similarity between the RTPL and the HPL. While Mokhtari et al. (2018) argue that the revised test is more conceptually sound and addresses the intolerable high failure rate of the original specifications, Pogliani, Paganini, and Rata (2019) demonstrate the low probability of passing the revised tests. In this paper, we examine how portfolio characteristics influence the propensity to pass the tests and the resultant implications for portfolio design and management. The insights we provide around PLA test offer an important contribution to existing literature.

4. HPL is the front office derived P&L of the portfolio of assets, and RTPL is the middle office derived P&L of the risk factor(s).

We find that the PLA tests present a significant challenge to desk-level portfolios and will necessitate reforms in risk factor mapping and/or portfolio construction decisions. From our analysis of stylized equity portfolios, we find that market capitalization, value weighting, and the volume of stocks in the portfolio relative to the equity risk factor (i.e., the broad equity market index) are important factors leading to positive outcomes in the PLA tests. Furthermore, through the PLA tests, the FRTB framework appears to incentivize passive portfolio management, whereby the portfolio is designed to assimilate the performance of the market index. Any such incentivization of passive portfolio management may lead to increased systemic concentration in high market capitalization stocks and may create liquidity issues in low market capitalization stocks.

These insights from the stylized analysis motivate our extended analysis of industry portfolios. We present supporting evidence that broad-based passive portfolios perform well in respect of the PLA tests. In contrast, portfolios comprised of small-cap stocks in particular tend to enter the PLA test defined amber and red zones, generally reflecting that such portfolios are less representative of the equity index risk factor. When we move to portfolios that are constructed more in line with active management strategies, so-called “smart Beta” ETFs, we find that funds that include high capitalization stocks display similar patterns. Indeed, when we go further and examine funds that rely on selecting from reduced populations of stocks, such as sectoral focused and socially responsible investing focused ETFs, we again find that such funds tend to pass the PLA tests when there is sufficient representation of large capitalization stocks. Finally, as an important insight, we discriminate between two sample periods: a pre-pandemic sample period, which captures a time of more tranquil market activity; and a sample period covering the pandemic period, which is relatively more volatile given the economic uncertainty the pandemic creates. We show that there is generally a degradation in the PLA test performance of this suite of industry portfolios when we transition to the more volatile pandemic sample period. Such periods of market volatility may create a significant challenge for banks in passing the PLA tests.

As a complement to the above, and to address our second research question, we conduct alternative comparable tests to examine the relative strength of the proposed IMA criteria to incentivize improvements in risk management. We exploit the same suite of stylized equity portfolios to examine the performance of four different risk models of varying performance to determine if the introduction of the additional IMA criteria specifically incentivizes banks to deploy superior risk models and/or the supporting calculative framework, specifically the use of risk factor modeling.

There are a number of studies that examine the performance of different risk models in the determination of market risk capital (e.g., Rossignolo, Fethi, and Shaban 2012, Burchi 2013). Berkowitz and O’Brien (2002) observe the deployment of conservative models that lack the ability to respond to changing volatility. Minimizing capital requirements while adhering to the Basel II backtesting criteria was the primary rationale for the choice of risk modeling method according to Mehta et al. (2012). Interestingly, Lucas (2001) examines whether, under the Basel I framework, banks are appropriately incentivized to implement correct internal models. Finding

that this is not the case, they present evidence suggesting that banks are prone to underreporting their true market risk. Subsequently, Hermsen (2010) finds that Basel II does not incentivize the use of models with more reasonable assumptions as this leads to higher levels of capital. They find that banks benefit (through reduced capital requirements) from the use of inferior performing models, contrary to the aims of Basel II. O'Brien and Szerszeń (2017) find that Basel II has misplaced incentives with respect to the choice of internal model. Against the backdrop of this literature, our study contributes through providing new insights into whether the FRTB framework corrects these failings in respect of superior risk model selection on the part of banks.

In this respect, we find that the FRTB desk-level backtests are weak compared to other available backtests and do not incentivize improvements in risk management through the deployment of superior risk models. Our comparison of the backtest performance of four risk models finds that the FRTB desk-level backtests have a high tolerance for exceedances and comparably low power to reject underperforming models. The relative strength of the PLA tests compared to the desk-level backtests suggests a change of emphasis to influence portfolio design (*ex ante*). This may prove to be a more significant challenge to IMA banks than previously assumed. Azoulay et al. (2018a) find that the initial perception of FRTB underestimates its implications for the trading-risk infrastructure.

The remainder of the study is structured as follows: Section 1 outlines the methodology and data, while giving details of how the empirical study design addresses the research question. The empirical results relating to the PLA tests are presented in Section 2, first for the stylized portfolios and subsequently for the industry portfolios. Section 3 outlines the result of the desk-level backtests. Section 4 concludes.

1. METHODOLOGY AND DATA

To evaluate the impact of FRTB's additional IMA criteria, we first clarify what these criteria measure, providing appropriate technical details as required. We begin with an explanation of the PLA tests and detail the benchmarking that we conduct to evaluate their impact. We then discuss the data we use for our stylized portfolio analysis and our industry portfolio analysis, providing details of the portfolio constructions. We then close with a presentation of the desk-level backtests and the alternative backtests to which we compare them. Full technical details of the PLA tests and backtests prescribed by FRTB, along with the benchmark tests, are presented in Online Appendix A.

1.1 PLA Tests

The PLA tests advocated under FRTB measure the similarity between the (realized) profit and loss (P&L) distribution of the portfolio as measured under front office pricing, labeled the HPL; and the (realized) P&L distribution as modeled under

risk management models, labeled the RTPL. PLA tests analyze the appropriateness of the risk mappings deployed prior to the application of the risk estimate. Mehta et al. (2012) find that most banks use a risk factor mapping approach rather than a full revaluation, which would require modeling each component in the portfolio. Risk factor mapping is a technique used in risk modeling to map a large complex portfolio to a manageable number of appropriate risk factors using sensitivities to these risk factors, and typically the covariance of the risk factors. For equity portfolios, the appropriate sensitivity measure is Beta, which captures the sensitivity of the stock to the chosen equity index (risk factor), which is taken to represent the market (Alexander 2009, p. 26).

The PLA tests comprise two measures: (i) the Spearman's rank (SR) correlation test and (ii) the Kolmogorov–Smirnov (KS) distributional test. Both are explicitly defined in FRTB documentation. The PLA tests focus on the adequacy of the data going into the risk models to reflect the portfolio held, specifically evaluating the relationship between the full information used for (front office) pricing and the information used in (middle office) risk modeling. The tests require that these data sets are sufficiently correlated through the SR test and have the same distributional form through the KS test.

To test the impact of the PLA tests, we design a range of portfolios to examine whether particular characteristics promote or impede the likelihood of passing the tests. Without loss of generality, and to avoid generating results influenced by risk factor dependency modeling (such as covariance matrices or copulas), we design portfolios that can be mapped to a single risk factor. For equity portfolios, equity indices are typically used as risk factors. We select the S&P 500 as our risk factor and systematically construct alternative portfolios from constituent stocks. In this way, our experiment design tests the relevance of different portfolio characteristics on the propensity for the *portfolio-to-risk-factor-mapping* to pass the PLA tests. We provide insights into whether there is a need to change risk modeling practice and/or portfolio management practice.

With this test design, we examine empirically the challenge of passing the PLA tests. The SR measures the correlation between the RTPL and HPL using daily data for a 12-month period. A 12-month test period is specified for the PLA tests in the FRTB documentation (BCBS 2019). The final published document on FRTB acknowledges the potential impact of a mechanism, whereby failure of the PLA test would result in an automatic switch to SA, therefore they have devised a graduated mechanism using *traffic light* tiering for the PLA tests. According to FRTB documentation, this SR correlation test statistic is categorized by the criteria set out in Table 1. The results of our SR testing exercise are reported against this system.

For the FRTB-prescribed KS distributional test, the empirical cumulative distribution of the HPL is derived by taking each HPL observation, determining how many HPL observations are less than or equal to it, and dividing by 250 (approximate number of trading days in 12 months). The empirical cumulative distribution of the RTPL is similarly determined. The KS test metric is the largest absolute difference between these two empirical distributions at any P&L level. Online Appendix A provides the technical details. The KS test metric is categorized under the criteria

TABLE 1
SPEARMAN'S RANK CORRELATION TEST CRITERIA

Green	Amber	Red
$S > 0.8$	$0.7 < S < 0.8$	$S < 0.7$

NOTE: Test statistics in the green zone pass the Spearman's rank correlation test, those in the red zone fail, and those in the amber zone will be subject to further monitoring and testing. This means that (in conjunction with the Kolmogorov Smirnov test—see Table 2) desk-level portfolios with results in the green zone will be authorized to use their internal model. Conversely, those with results in the red zone will not be permitted to use their internal model. Desk-level portfolios with results in the amber zone will be subject to further monitoring and higher capital requirements.

TABLE 2
KOLMOGOROV-SMIRNOV DISTRIBUTIONAL TEST CRITERIA

Green	Amber	Red
$KS < 0.09$	$0.09 < KS < 0.12$	$KS > 0.12$

NOTE: Test statistics in the green zone pass the Kolmogorov–Smirnov distributional test, those in the red zone fail, and those in the amber zone will be subject to further monitoring and testing. This means that (in conjunction with the Spearman's rank correlation test—see Table 1) desk-level portfolios with results in the green zone will be authorized to use their internal model. Conversely, those with results in the red zone will not be permitted to use their internal model. Desk-level portfolios with results in the amber zone will be subject to further monitoring and higher capital requirements.

set out in Table 2. The results of our KS testing exercise are reported against this system. For benchmark purposes, we compare our KS test results against the Anderson–Darling distributional test.

1.2 Portfolio Constructions

For the PLA test dimension of our quantitative impact analysis, we first consider a range of stylized stock portfolios. The stock portfolios are selected to exhibit varying degrees of two characteristics: market capitalization and portfolio Beta. We have chosen to focus our analysis on stock portfolios related to the S&P 500 index. We begin with the construction of decile portfolios drawn from S&P 500 stocks, where the portfolios have the following characteristics: (A) stocks ranked by market capitalization and weighted in proportion to their index weighting, (B) stocks ranked by market capitalization and equally weighted, (C) stocks ranked by Beta and weighted in proportion to their index weighting, and (D) stocks ranked by Beta and equally weighted.⁵ We then systematically construct a set of portfolios comprised of alternative accumulations of the decile portfolios, which we collectively term *cumulated decile portfolios*. Table 3 summarizes these stylized portfolios, where there are 65 alternative portfolios (i.e., 10 decile portfolios plus 55 cumulated decile portfolios) for each ranking–weighting combination. We draw index prices, constituent mem-

5. As per convention, the 1st decile portfolio comprises the lowest ranked stocks based on either market capitalization or Beta value (stocks 451–500 from the index), while the 10th decile portfolio comprises the highest ranked stocks based on either market capitalization or Beta value (stocks 1–50 from the index).

TABLE 3
STYLIZED EQUITY PORTFOLIOS FROM S&P 500 STOCKS

Decile portfolios [ranking by (i) market capitalization or (ii) Beta]

10th (highest rank)	1–50
9th	51–100
8th	101–150
7th	151–200
6th	201–250
5th	251–300
4th	301–350
3rd	351–400
2nd	401–450
1st (lowest rank)	451–505

Cumulative decile portfolios [ranking by (i) market capitalization or (ii) Beta]

1–50	51–100	101–150	151–200	201–250	251–300	301–350	351–400	401–450	451–505
1–100	51–150	101–200	151–250	201–300	251–350	301–400	351–450	401–505	
1–150	51–200	101–250	151–300	201–350	251–400	301–450	351–505		
1–200	51–250	101–300	151–350	201–400	251–450	301–505			
1–250	51–300	101–350	151–400	201–450	251–505				
1–300	51–350	101–400	151–450	201–505					
1–350	51–400	101–450	151–505						
1–400	51–450	101–505							
1–450	51–505								
1–505									

NOTE: The table shows the systemic approach to the portfolio construction process, with stocks drawn from the S&P index constituents. Decile and cumulative decile portfolios are constructed as labeled. Constituent stocks are ranked either by market capitalization or by Beta. See Section 1.3 for further details.

bers, free-float shares, and share prices for a 2-year⁶ period from Bloomberg. Different holiday conventions for stocks and indices were taken into account. The effect of index constituent changes was minimal over the test period. Adjustments made to index levels due to corporate actions (e.g., share adjustments, initial public offerings, mergers, and acquisitions) complicate the mapping from constituent prices and weightings to the index level. However, we argue that this enigma occurs in practice, so sanitizing this issue would be unrepresentative.

In the case of the PLA tests, the rationale for the above design is to investigate in a controlled environment whether particular characteristics (market capitalization, stock Beta) and approaches to weighting (value or equal) have an impact on the propensity for the selected stock portfolios to pass the PLA tests. This allows us to comment on the implications of FRTB from a portfolio management perspective.

To further explore the practical implications of the introduction of PLA tests, we wish to consider portfolios that better reflect the reality of investment practice. Indeed, we would like to make some statements in respect of the performance of passively and actively managed portfolios, while considering alternative investment strategies. To this end, we draw on the universe of ETFs in order to capture portfolio constructions

6. The 2-year period consists of a 12-month calibration period (07/06/17 to 24/05/18) and a 12-month testing period (25/05/18 to 08/05/19).

TABLE 4
SELECTED EXCHANGE TRADED FUNDS (ETFs) FROM THE U.S. MARKET

Ticker	ETF name
Passive and size based	
IVV	iShares Core S&P 500 ETF
IJH	iShares Core S&P Mid-Cap ETF
IJR	iShares Core S&P Small-Cap ETF
Value	
IVE	iShares S&P 500 Value ETF
IJJ	iShares S&P Mid-Cap 400 Value ETF
IJS	iShares S&P Small-Cap 600 Value ETF
Growth	
IVW	iShares S&P 500 Growth ETF
IJK	iShares S&P Mid-Cap 400 Growth ETF
IJT	iShares S&P Small-Cap 600 Growth ETF
Momentum	
MTUM	iShares MSCI USA Momentum Factor ETF
XMMO	Invesco S&P MidCap Momentum ETF
XSMO	Invesco S&P SmallCap Momentum ETF
Minimum volatility	
USMV	iShares MSCI USA Min Vol Factor ETF
XMLV	Invesco S&P MidCap Low Volatility ETF
SMMV	iShares MSCI USA Small Cap Min Vol Factor ETF
Sectoral	
XLE	Energy Select Sector SPDR Fund
XLI	Industrial Select Sector SPDR Fund
XLK	Technology Select Sector SPDR Fund
Socially responsible investing	
ESGU	iShares ESG Aware MSCI USA ETF
SUSA	iShares MSCI USA ESG Select ETF
PBW	Invesco WilderHill Clean Energy ETF

NOTE: The table shows the Bloomberg tickers for the selected ETFs and their market names.

that differ notably from the stylized portfolios considered thus far. Extending the market capitalization focus, we wish to explore whether there is some differential effect between portfolios comprised of alternative sized companies. To this end, we consider the broad-based iShares Core S&P 500 ETF, and its two counterpart ETFs, the iShares Core S&P 500 Mid-Cap ETF and iShares Core S&P 500 Small-Cap ETF, that concentrate investment in mid-cap stocks and small-cap stocks, respectively. These ETFs are summarized in Table 4.

While the above ETFs are passive in nature, we wish to consider alternative investment strategies that at least augment the passive strategy with some form of stock selection-based active management. We focus in particular on factor-based investment strategies. We select a number of ETFs, commonly referred to as “smart Beta” funds, that follow (i) value, (ii) growth, (iii) momentum, and (iv) minimum volatility investment strategies. For comparison with the size-based passive ETFs, we consider a range of counterparts for each factor-based approach that concentrates on mid-cap and small-cap stocks. Table 4 summarizes the resulting 12 ETFs considered.

We develop our analysis further by means of considering a number of sectoral-level ETFs to provide insights into the performance of the PLA tests for portfolios constructed from a reduced, concentrated population of stocks. We chose a selection of sectors that are of particular importance to the U.S. economy. Specifically, we consider the energy, industrial, and technology sectors. The specific ETFs are the Energy Select Sector SPDR Fund, Industrial Select Sector SPDR Fund, and Technology Select Sector SPDR Fund. Table 4 again summarizes these.

Finally, to close our analysis, we pick up on the recent trend of socially responsible investing, and in particular the use of Environmental, Social, and Governance (ESG) metrics within portfolio management practice. ESG screening leads to a reduction in the population of stocks used for portfolio construction. This final analysis provides insights into whether ESG orientated stock portfolios perform differently relative to broader portfolios. We consider two prominent ESG-based ETFs. These are the iShares ESG Aware MSCI USA ETF and iShares MSCI USA ESG Select ETF. As a third, tangential ETF in the socially responsible investing domain, we also consider a clean technology-based fund, which allows for some overlap in scope with the sectoral analysis. We consider the well-established Invesco WilderHill Clean Energy ETF for this purpose.

1.3 Desk-Level Backtests

In the case of the desk-level backtests, we concentrate on the stylized portfolio constructions, allowing us to examine how FRTB backtests compare to other available backtests using the same data restrictions (FRTB specifies a 250-day rolling calibration period and 250-day backtesting period), and if the introduction of these desk-level backtests improve incentives for banks to deploy superior risk models. Varying the portfolio rankings and weightings as proposed in Section 1.1 allows us to consider portfolios that have varying degrees of market risk exposure. This aspect of the study allows us to comment on the implications of FRTB from a risk management perspective.

We wish to examine the ability of FRTB desk-level backtests to reject poorly performing risk models and thereby incentivize the use of superior forecasting risk models. We design our quantitative impact analysis to include four popular risk models: (i) Normal Linear (NL) VaR, (ii) Historical Simulation (HS), (iii) Exponentially Weighted Moving Average (EWMA), and (iv) GARCH(1,1). These risk models are chosen because they (or close variations) are the most popular models deployed by banks (Mehta et al. 2012), while offering sufficiently different performance levels for our analysis. Indeed, Hermesen (2010) notes that the NL VaR and EWMA models are premised on particularly unreasonable assumptions. We backtest these risk models under FRTB desk-level specifications at the two confidence levels 97.5% and 99%, along with eight other benchmark backtests (Table 5).

A concern regarding the specifications of FRTB desk-level backtesting framework is the threshold values for VaR at both the 99% and 97.5% confidence levels. The specified values of 12 and 30 exceedances in a 250-day test period are in excess of

TABLE 5
LIST OF BACKTESTS PERFORMED

Test	Label
FRTB Desk-Level Test (99% and 97.5%)	FRTB_99/FRTB_97.5
Binomial Test	BIN
Basel II Traffic Light Test	TL
Proportion of Failures Test (Kupiec 1995)	POF
Time Until First Failure Test (Kupiec 1995)	TUFF
Independence Test (Christoffersen 1998)	CC
Time Between Failures Test (Haas 2001)	TBFI
Combined POF and Independence Test	CCI
Combined Coverage Time Between Failures Test	TBF

NOTE: By the nature of their objectives, the backtests are primarily exceedance driven. This can incentivize the implementation of conservative risk models that will overestimate the risk forecast. However, these conservative models are often not reactive to emerging events, which can then lead to a clustering of exceedances such as were observed during the 2007–09 financial crisis. The Time Between Failures test is designed to capture the potential of a model to give rise to clustered exceedances but this would only be exhibited in a volatile test period. When the backtests are performed on a benign test period, model weaknesses are less likely to be revealed. The FRTB backtesting framework specifies using a test period of the most recent 12 months with calibration on the previous 12 months. However, in contrast, the calculative framework for determining market risk regulatory capital requires ES calibrated on a 12-month period within the timeframe of the 2007–09 financial crisis. Backtesting using a stressed period would be a more rigorous foundation for evaluating model performance.

the expected value of the binomial distribution of exceedances. At the corresponding significant levels of 1% and 2.5%, the expected number of exceedances is 2.5 and 6.25, respectively, with standard deviations of 1.57 and 2.47. A VaR exceedance of 12 at the 99% confidence level is equivalent to the expected value of 2.5 plus *six* standard deviations. A VaR exceedance of 30 at the 97.5% confidence level is equivalent to the expected value of 2.5 plus *nine* standard deviations.

The entity-level backtest deploys a traffic light system similar to the Basel II framework but using a different weighting system to that deployed in Basel II. FRTB documentation discusses the issue of Type I and Type II errors arising in backtesting. The additional leniency in the prescribed thresholds appears to be a means of reducing the likelihood of incurring Type II errors (the erroneous rejection of a sound risk model) but significantly increases the likelihood of a Type I error (erroneously validating an inadequate risk model). We suspect that the ability to reject an internal model at desk level on the basis of FRTB backtest specification is very weak. We will test this empirically by comparing the performance of FRTB backtest specification on the benchmark models.

2. EMPIRICAL RESULTS: PLA TESTING

2.1 Stylized Portfolios

As described in Section 1.1, the PLA tests comprise the SR correlation test and the KS distributional test. We begin with the former. Figure 1 presents a visualization of the results of the SR correlation tests for the individual decile portfolios, while Figure 2 similarly presents a visualization of the results for the cumulative decile portfolios. The associated tabulated results are presented in Online Appendix B. We have used a grayscale colour code in the plots to align with the defined red (dark

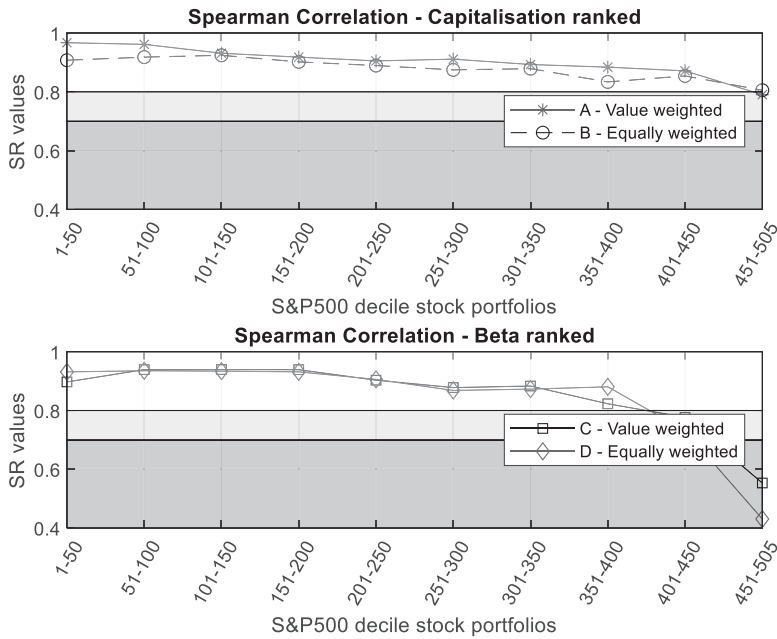


Fig 1. FRTB Spearman’s Rank Correlation Test for S&P500 Decile Portfolios.

NOTE: The figure shows the results of the Spearman Rank (SR) correlation test. The test examines the correlation between the risk factor and each of the constructed decile portfolios ranked by capitalization or Beta (see Section 3) over the test period. FRTB determines that the test is passed if the SR value is >0.8 (green zone), it fails if $SR < 0.7$ (red zone), and is in the amber zone for SR values between 0.7 and 0.8. Portfolio type A involves ranking by market capitalization and value weighting, portfolio type B involves ranking by market capitalization and equal weighting, portfolio type C involves ranking by Beta and value weighting, and portfolio type D involves ranking by Beta and equal weighting.

gray), amber (light gray), and green (white) testing zones as per the FRTB prescription (Tables 1 and 2).⁷

Each of the decile portfolios with a capitalization ranking and equal weighting show SR correlations test results within the green zone. This is true of the value-weighted portfolios too, except for deciles 451–500, which is marginally inside the amber zone with a result of 0.7903. However, all the other value-weighted deciles have a higher SR result than their equivalent equal-weighted deciles. Eight out of ten deciles (both value and equal weighting) that are ranked by Beta have SR test results in the green zone. Deciles (401–450) and (451–500) are amber and red, respectively. There is no significant dominance between value-weighted and equal-weighted portfolios SR test results for Beta-ranked portfolios. In both the capitalization and Beta-ranked cases, the SR test statistics decline as one progresses from the highest decile

7. Benchmarking against the Anderson-Darling (AD) distributional test, we find that the KS distributional test is somewhat more difficult to pass for the stylized portfolios. These AD test results are available upon request.

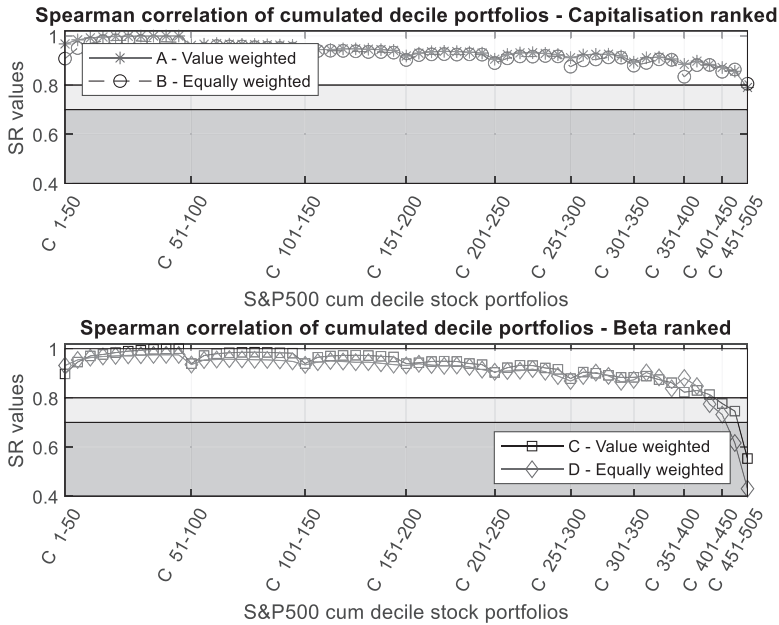


Fig 2. FRTB Spearman's Rank Correlation Test for S&P500 Cumulative Decile Portfolios.

NOTE: The figure shows the results of the Spearman Rank (SR) correlation test. The test examines the correlation between the risk factor and each of the constructed cumulative decile portfolios ranked by capitalization or Beta (see Section 2) over the test period. FRTB determines that the test is passed if the SR value is >0.8 (green zone), it fails if $SR < 0.7$ (red zone), and is in the amber zone for SR values between 0.7 and 0.8. Portfolio type A involves ranking by market capitalization and value weighting, portfolio type B involves ranking by market capitalization and equal weighting, portfolio type C involves ranking by Beta and value weighting, and portfolio type D involves ranking by Beta and equal weighting.

portfolio to the lowest decile portfolio,⁸ reflecting ever lower representativeness of the S&P 500 index.

When we examine the cumulative decile portfolios, for both the capitalization and Beta rankings, we see a very similar pattern. We note that as the decile portfolios cumulate, the SR correlation measure increases, reflecting that the portfolios become increasingly representative of the index. Moreover, excluding the lower decile portfolios recursively, the SR measure decreases, with the largest effect occurring when the two lowest decile portfolios (1–50 and 51–100) are removed. Although this effect is less pronounced as we progress to exclude higher deciles, this leads to amber and red zone results at the lower end cumulated portfolios. Beta-ranked portfolios show significantly poorer performance in respect of the SR test at the upper end of the cumulated deciles. The overall picture suggests that the SR tests are relatively easy

8. We adopt the convention of describing decile 450–500 as the 1st decile, 400–450 as the 2nd, etc.

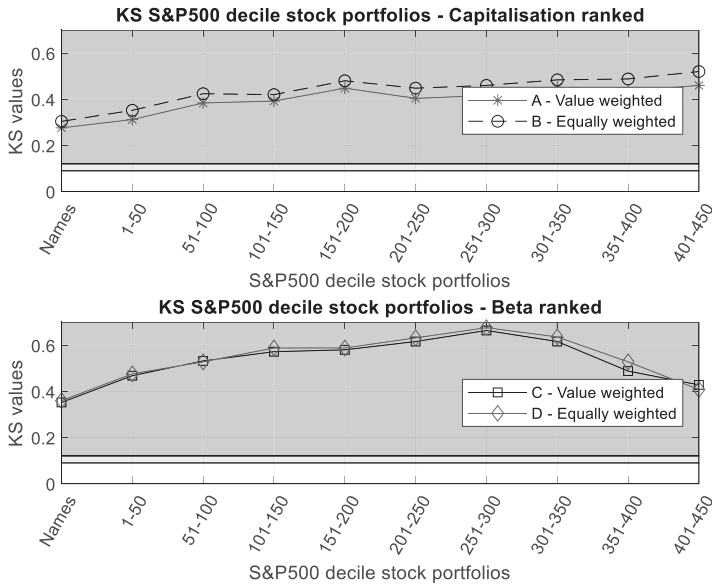


Fig 3. Kolmogorov–Smirnov Test for S&P500 Decile Portfolios.

NOTE: The figure shows the results of the Kolmogorov–Smirnov (KS) distributional test. The test examines the distributional similarity between the risk factor and each of the constructed decile portfolios ranked by capitalization or Beta (see Section 2) over the test period. FRTB determines that the test is passed if the KS value is <0.09 (green zone), it fails if the KS value >0.12 (red zone), and is in the amber zone for KS values between 0.09 and 0.12. Portfolio type A involves ranking by market capitalization and value weighting, portfolio type B involves ranking by market capitalization and equal weighting, portfolio type C involves ranking by Beta and value weighting, and portfolio type D involves ranking by Beta and equal weighting.

for stock portfolios to pass with capitalization ranked value weighted (A) portfolios performing better overall.

Turning our attention to the KS testing, we see a completely different picture emerge, with the KS test proving much more challenging to pass. Indeed, no decile portfolio is successful in passing the KS test. Figures 3 and 4 present the visualizations of the KS test results for decile and cumulative decile portfolios across the capitalization and Beta-ranked categories. Tabulated results are available in Online Appendix B. For the cumulative decile portfolios with capitalization ranking and value-weighting (Cumulated A portfolios), we see that the cumulation must reach portfolio (1–250) before the KS test is amber and (1–450) before results are in the green zone (two portfolios only). Amber KS test results are achieved for six other Cumulated A portfolios: (51–350, 51–400, 51–450, 51–500, 101–459, 101–500). None of the cumulated decile capitalization-ranked equally weighted portfolios (Cumulated B portfolios), nor any of the cumulated decile Beta-ranked portfolios (Cumulated C and D portfolios) pass the KS test. In general, we can see from the graphs in Figures 2 and 4 that capitalization ranked cumulated portfolios perform better in the KS test than Beta-ranked cumulated portfolios. Furthermore, weighting is not signif-

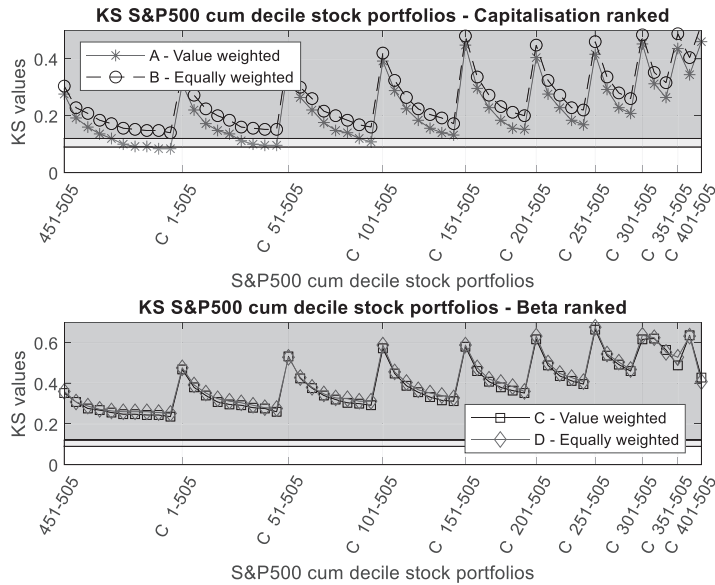


Fig 4. Kolmogorov–Smirnov Test for S&P500 Cumulative Decile Portfolios.

NOTE: The figure shows the results of the Kolmogorov–Smirnov (KS) distributional test. The test examines the distributional similarity between the risk factor and each of the constructed cumulative decile portfolios ranked by capitalization or Beta (see Section 2) over the test period. FRTB determines that the test is passed if the KS value is < 0.09 (green zone), it fails if the KS value > 0.12 (red zone), and is in the amber zone for KS values between 0.09 and 0.12. Portfolio type A involves ranking by market capitalization and value weighting, portfolio type B involves ranking by market capitalization and equal weighting, portfolio type C involves ranking by Beta and value weighting, and portfolio type D involves ranking by Beta and equal weighting.

icant in the performance of the Beta-ranked cumulated portfolios. However, value-weighted portfolios consistently perform better than those that are equally weighted for capitalization-ranked cumulated portfolios. This is interesting because it suggests that when the portfolios are constructed using value weighting rather than equal weighting (the former offering greater representativeness of the mapped index by construction) and capitalization-ranked, this leads to greater similarity between the empirical distributions of the portfolios and the risk-mapped index. Again, we have used a grayscale to code the regions: red zone (dark gray), amber zone (light gray), and green zone (white).

We have shown that none of the individual decile portfolios pass the KS test and only a small number of cumulated decile portfolios pass. The results indicate that value weighting is a significant characteristic for a stock portfolio's likelihood of passing the KS test. These findings imply that for equity portfolios to pass the PLA tests, they must be significantly aligned to the mapped risk factors. These results suggest that, for stock portfolios, index risk factor alignment should consider: (i) inclusion of high market capitalization stocks, (ii) adoption of a weighting mechanism similar to that of the index, and (iii) comparable diversification. This suggests that the PLA

tests (in particular the KS test) incentivize a passive form of portfolio management, strongly aligned to the performance of the market index.

These quantitative impact findings have important implications for the banking sector. The PLA tests may influence the constitution of portfolios through incentivizing passive portfolio management in preference to active portfolio management. Applied on a system-wide basis, the FRTB framework may encourage greater holdings of high capitalization stocks from an equity index, while instituting disincentives to hold lower capitalization stocks. This may impact the prices and liquidity in this latter market segment, potentially leading to the creation of systemic risk. Active portfolio management strategies may find it challenging to meet the requirements of the PLA tests if they deploy the standard risk factor mapping approach to risk modeling. Consideration of a full revaluation approach to risk modeling may be required. This would replace the practice of mapping the portfolio to relevant risk factors, modeling instead each individual asset in the portfolio and their respective correlations, a computationally intensive exercise. Bank portfolios have typically a large number of constituents so a full-revaluation approach presents additional costs through increased computational time and system requirements.

2.2 *Industry Portfolios*

The stylized analysis of the previous section raises some questions. As noted, the PLA tests may incentivize portfolio managers to prefer passive investment strategies over active, while promoting a bias for high capitalization stocks over lower capitalization stocks. Motivated by these findings, we move to examine the PLA tests in a more realistic setting, and so we consider the range of industry portfolios outlined in Section 1.2. We examine their performance under two time periods, one with low volatility (June 2017 to May 2019) and a second period with high volatility (December 2019 to December 2021). These ETFs span portfolio constructions based on particular investment features, whether company size (large-cap, mid-cap, and small-cap), investment factors (value, growth, momentum, and volatility), sectors (energy, industrial, and technology), or social responsibility (ESG, clean technology). This analysis allows us to extend beyond the market capitalization basis of the stylized portfolios. The results of their performance in the PLA tests are shown in Table 6, for the relatively tranquil pre-pandemic period, and in Table 7, for the more volatile pandemic period.^{9, 10}

For the pre-pandemic period (Table 6), the passive ETFs perform well under both components of the PLA tests, with a stellar performance evidenced by the iShares Core S&P 500 ETF for both the SR correlation and KS distribution tests. This ETF

9. S&P 500 10-day historical volatility is 22.375% in the period 06/12/2019 to 09/12/21 compared to 12.069% in 07/06/17 to 08/05/19. The VIX index has a 50-day moving average of 19.22% in the pandemic period compared to 14.04% in the earlier pre-pandemic period.

10. Benchmarking against the Anderson-Darling (AD) distributional test, we find very close consistency with the results of the KS distributional tests. These AD test results are available upon request.

TABLE 6
PLA RESULTS FOR SELECTED EXCHANGE TRADED FUNDS (PREPANDEMIC PERIOD)

Passive Ticker	Value			Growth			Momentum			Min vol.			
	SR	KS	Ticker	SR	KS	Ticker	SR	KS	Ticker	SR	KS	Ticker	
IVV	0.977	0.028	IVE	0.889	0.056	IVW	0.940	0.052	MTUM	0.878	0.068	USMV	0.835
IJH	0.850	0.076	IJJ	0.807	0.100	IJK	0.853	0.092	XMMO	0.793	0.152	XMLV	0.660
IJR	0.774	0.104	IJS	0.738	0.124	IJT	0.785	0.100	XSMO	0.744	0.152	SMMV	0.663
Sector													
Ticker	SR	KS											
XLI	0.814	0.076											
XLK	0.865	0.064											
XLE	0.646	0.144											
ESG													
ESGU	0.922	0.060											
SUSA	0.937	0.040											
PBW	0.658	0.164											

NOTE: The table shows the PLA test results for a range of ETFs in the U.S. market. Table 4 can be used to identify the ETFs from the Bloomberg tickers. The results relate to the prepandeemic period 07/06/17 to 08/05/19.

TABLE 7
PLA RESULTS FOR SELECTED EXCHANGE TRADED FUNDS (PANDEMIC PERIOD)

Passive	Value			Growth			Momentum			Min vol.		
	Ticker	SR	KS	Ticker	SR	KS	Ticker	SR	KS	Ticker	SR	KS
IVV	0.996	0.024	0.08	IVV	0.890	0.080	MTUM	0.794	0.104	USMV	0.830	0.064
IJH	0.762	0.092	0.124	IJK	0.795	0.100	XMMO	0.733	0.124	XMLV	0.700	0.068
IJR	0.637	0.144	0.172	IJT	0.696	0.128	XSMO	0.641	0.160	SMMV	0.755	0.116
Sector												
Ticker	SR	KS										
XLI	0.727	0.108										
XLK	0.858	0.104										
XLE	0.432	0.176										
ESG												
ESGU	0.992	0.032										
SUSA	0.979	0.040										
PBW	0.398	0.224										

NOTE: The table shows the PLA test results for a range of ETFs in the U.S. market. Table 4 can be used to identify the ETF from the Bloomberg ticker. This relates to the pandemic period 16/12/19 to 09/12/21.

has predominately high capitalization stocks and is value weighted, showing alignment to our conclusions from the stylized analysis. Of the passive ETFs chosen, only the iShares Core S&P Small-Cap ETF has results for both the SR and KS tests in the amber zone. This provides evidence of a divergence in performance when the portfolio comprises small-cap stocks that are less representative of the equity index risk factor. We see, however, for the more volatile pandemic period (Table 7), a significant deterioration in the performance of the mid-cap and small-cap based ETFs is observed, with the former falling into the amber zone and the latter falling into the red zone, failing both the SR and KS tests.

Turning attention to the four forms of actively managed ETFs considered, which again target value, growth, momentum, and volatility minimization strategies, we can see that each “smart Beta” ETF demonstrates similar patterns when the funds include high capitalization stocks. Such funds pass both the SR and KS tests for the pre-pandemic period. In contrast, for the same sample period, funds restricted to mid-cap or small-cap stocks fail one or both components of the PLA tests. Indeed, we observe that this is more pronounced for the momentum and volatility minimization-based investment strategies. This evidence corroborates further our observation from the stylized portfolios that the inclusion of high capitalization stocks has a strong influence over the likelihood of a stock portfolio passing the PLA tests, irrespective of what stock selection approach is taken.

Comparing the tranquil and more volatile pandemic period, we see a significant increase in the portfolios’ propensity to fail the PLA tests. Interestingly, this indicates that in periods of high volatility, there is a deterioration in the comparative attributional characteristics of correlation and distributional form between the portfolio and its risk factor mapped proxy. Higher likelihood of PLA test failure in periods of high volatility is a significant concern. Failure of the PLA tests means recourse to determining capital under the SA mechanism. Under the FRTB reforms, SA will also be a risk-sensitive measure, meaning that capital requirements will increase in periods of high volatility. Therefore, desks that have been ejected from the use of IMA because of failing the PLA tests will face a very significant increase in regulatory capital requirements. This could be characterized as a capital cliff effect (Lee 2013) and is reminiscent of the issue of procyclicality of VaR, which has been recognized for its destabilizing affects in requiring additional capital contemporaneously with stressed market conditions (Danielsson, Shin, and Zigrand 2012, Adrian and Shin 2013, Vasileiou and Samitas 2020).

We see also that the KS test is more challenging to pass than the SR test, as suggested by the analysis of the stylized portfolios. As an example of this, the mid-cap value ETF (iShares S&P Mid-Cap 400 Value ETF) and mid-cap growth ETF (iShares S&P Mid-Cap 400 Growth ETF) display amber values for the KS test while passing the SR test.

The results for the chosen sectoral ETFs provide further interesting insights. Both the industrial sector ETF (Industrial Select Sector SPDR Fund) and technology sector ETF (Technology Select Sector SPDR Fund) pass the PLA tests. The industry sector ETF comprises 9.46% of the top S&P 500 market capitalization decile, while

the technology ETF comprises 21.41% (source: Bloomberg). Notably, the industrial ETF focuses primarily on high-cap stocks. The XLE Energy Select Sector SPDR Fund fails both the SR and KS tests. The energy sector ETF, however, comprises of only 5% of the S&P index by market capitalization and less than 2% of the top market capitalization decile (source: Bloomberg). So again, while these ETFs are constructed from a reduced population of stocks, the level of large-cap stock representation appears to be an important factor in driving success with the PLA tests.

To conclude, we discuss the analysis of the socially responsible investing group of ETFs. The iShares ESG Aware MSCI USA ETF and the iShares MSCI USA ESG Select ETF (SUSA) both show an ability to pass the PLA test across the prepandemic and pandemic sample periods. This shows considerable alignment with the broad based S&P500 ETF (iShares Core S&P 500 ETF). These two ESG-based ETFs seek to track the MSCI USA Index with favorable ESG characteristics, as defined by the tracked index. Although this ETF excludes participation of stocks with unfavorable ESG characteristics, it has broad sectoral reach and focuses on large- and mid-cap stocks predominantly. This may explain to some extent the success of these ESG-based ETFs in respect of the PLA tests, but it is a notable observation that reducing the population of stocks on the basis of ESG screening does not lead to a statistically significant difference in the RTPL and the HPL. In contrast to the ESG-based ETFs, however, we see that the Invesco WilderHill Clean Energy ETF fails both components of the PLA tests across both sample periods. This likely reflects the narrow focus of this particular fund on companies that develop clean technology solutions, which leads to low representativeness of the equity index risk factor.

3. . EMPIRICAL RESULTS: DESK-LEVEL BACKTESTING

We now examine the results from the FRTB desk-level backtests for the same suite of stylized equity portfolios detailed in Section 1.2. Under these desk-level backtests, maximum exceedance levels of 12 and 30 for 99% and 97.5% confidence levels are tolerated. We are interested in establishing the power of the FRTB backtest to reject poorly performing risk models relative to the alternative benchmark backtests. As outlined in Section 1.3, we examine the performance of four different VaR models (NL, HS, EWMA, GARCH(1,1)), and determine whether the FRTB framework incentivizes banks to implement better performing models. Primarily we are interested in the impact and implications of the desk-level backtests on market risk management practice. We segregate the analysis into three key aims:

- (i) Assess the performance of FRTB desk-level backtests relative to alternative backtests.
- (ii) Assess the impact of deploying two confidence levels in an integrated approach.
- (iii) Assess the impact of portfolio characteristics on backtesting performance.

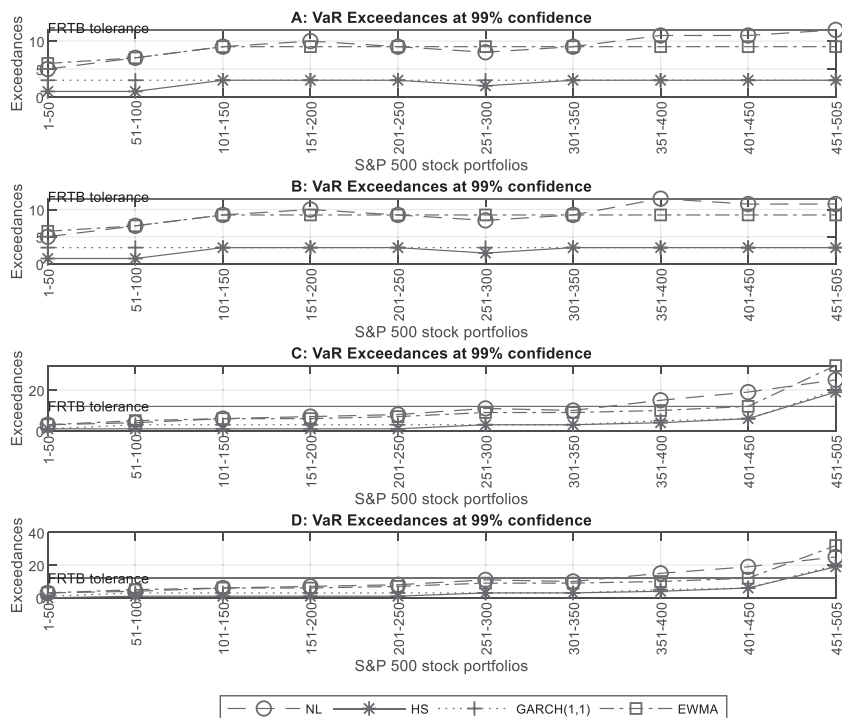


Fig 5. FRTB Backtesting Results for Decile Portfolios, 99% Confidence Level.

NOTE: The figure shows the exceedances for the four risk models considered (Normal Linear (NL), Historical Simulation (HS), GARCH(1,1), and EWMA) using a 250-day calibration period (09/06/17–06/06/18) and 250-day test period (07/06/18–06/06/19). FRTB exceedance tolerance of 12 for the 99% confidence level is also shown. Decile portfolios are considered (see Section 2). Portfolio type A involves ranking by market capitalization and value weighting, portfolio type B involves ranking by market capitalization and equal weighting, portfolio type C involves ranking by Beta and value weighting, and portfolio type D involves ranking by Beta and equal weighting.

Although our key focus here is the performance of FRTB desk-level backtests relative to alternative backtests, the backtests themselves assess the performance of the VaR models. We seek insights into the ability of FRTB desk-level backtests to identify and reject poorly performing risk models. A primary indicator of a poorly performing VaR model is a high number of exceedances. Exceedances occur when realized losses are higher than those forecast under the VaR model.

The backtest results are reported under FRTB desk-level backtests plus *eight* additional backtest specifications for each of the *four* risk models, while using the *two* confidence levels, 97.5% and 99%. Figures 5 and 6 show plots of the VaR exceedances recorded by the alternative risk models across the individual decile portfolios under the FRTB framework, while Figures 7–10 show similar results for the cumulative decile portfolios. Online Appendix B reports the results for the eight benchmark backtests.

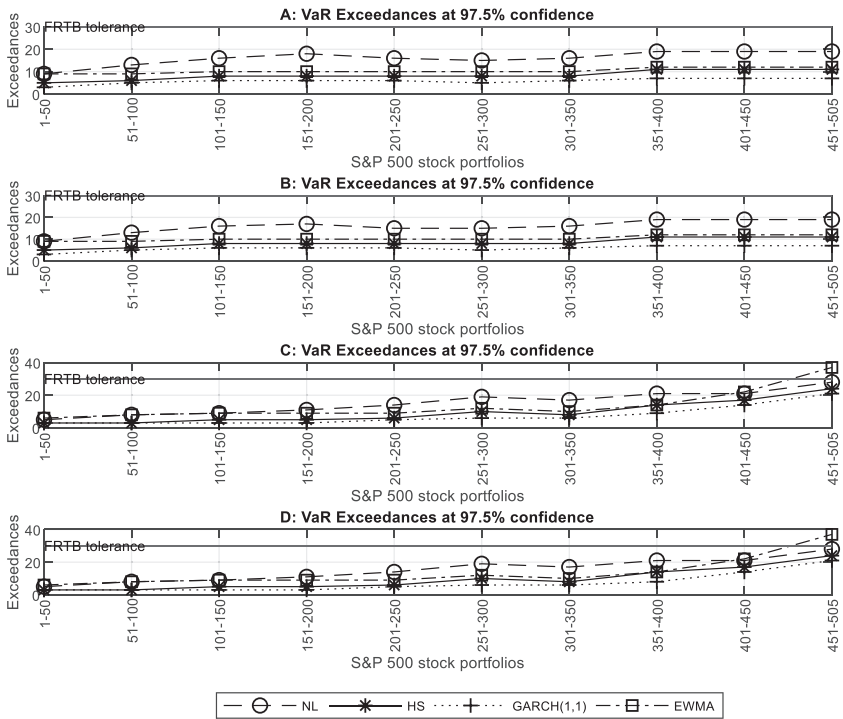


Fig 6. Backtesting Results for Decile Portfolios, 97.5% Confidence Level.

NOTE: The figure shows the exceedances for the four risk models considered (Normal Linear (NL), Historical Simulation (HS), GARCH(1,1), and EWMA) using a 250-day calibration period (09/06/17–06/06/18) and 250-day test period (07/06/18–06/06/19). FRTB exceedance tolerance of 30 for the 97.5% confidence level is also shown. Decile portfolios are considered (see Section 2). Portfolio type A involves ranking by market capitalization and value weighting, portfolio type B involves ranking by market capitalization and equal weighting, portfolio type C involves ranking by Beta and value weighting, and portfolio type D involves ranking by Beta and equal weighting.

In most cases, the individual decile portfolios pass the 99% and 97.5% backtests, with some exceptions evidenced for the lowest ranking decile portfolios in the case of ranking by Beta. We find that the number of exceedances under the NL VaR model in particular are consistently higher than the other risk models (Figures 5 and 6), indicating that it is a weaker resolution method. That is, its forecasts are more likely to be exceeded. Indeed, the NL VaR is recognized as being based upon unreasonable assumptions (Hermsen 2010), in particular the assumption that returns are normally distributed, a characterization that has been roundly criticized (Jansen and De Vries 1991, Danielsson et al. 2005, Ibragimov and Walden 2007). However, despite the apparent weak forecasting strength of NL VaR (particularly at the 99% confidence level), the FRTB backtests did not reject this model for any of (for example) the Type A decile portfolios. In contrast, when we look at the benchmark backtests (Online Appendix B), the NL VaR model for the 2nd decile portfolio (401–451) is in the

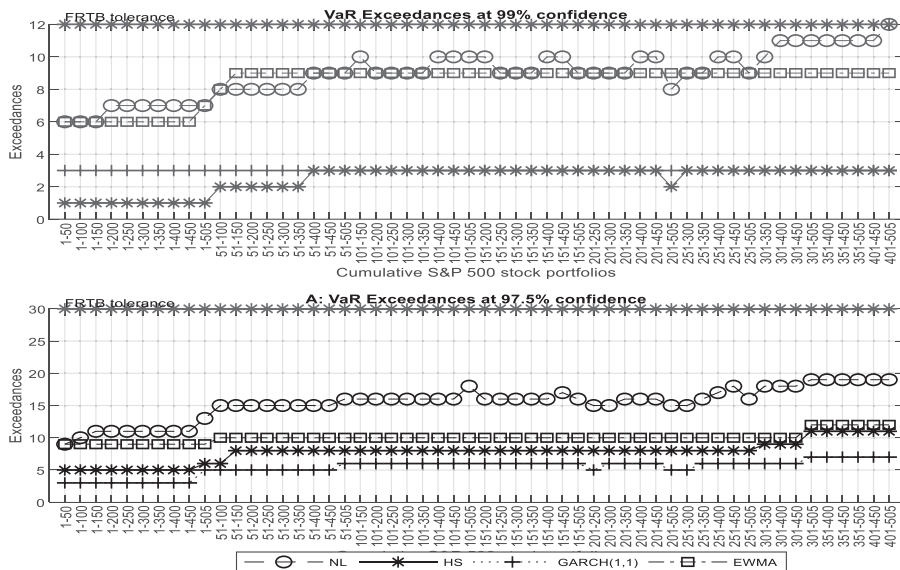


Fig 7. FRTB Backtesting Results for Cumulative Decile Portfolios (Portfolio Type A).
 NOTE: The figure shows the exceedances for the four risk models considered (Normal Linear (NL), Historical Simulation (HS), GARCH(1,1), and EWMA) using a 250-day calibration period (09/06/17–06/06/18) and 250-day test period (07/06/18–06/06/19). FRTB exceedance tolerances of 12 and 30 for the 99% and 97.5% confidence levels respectively are also shown. Cumulative decile portfolios are considered (see Section 2). Portfolio type A involves ranking by market capitalization and value weighting.

yellow zone¹¹ under the Basel II Traffic light backtest, while being in the red zone for the 1st decile portfolio (451-500). The same portfolios were rejected outright by the POF, BIN, CC, TBF, and TBFi tests. The results for EWMA VaR exhibit comparable exceedance levels to NL VaR and correspondingly, FRTB backtests are weak at rejecting this risk model as well. We provide clear evidence that the FRTB framework is inferior to many of the alternatives.

Furthermore, at the 97.5% confidence level, NL VaR, and EWMA VaR again exhibit the highest number of exceedances. However, at this confidence level, the deviation between their exceedances and those of HS and GARCH(1,1) are less pronounced. FRTB desk-level backtests rank in the lowest two backtests for exhibiting an ability to identify and reject poorly performing risk models at this confidence level. The introduction of a two-test combination for the desk-level backtests, with two different confidence levels was an opportunity to develop a combined test with stronger power to reject poorly performing VaR resolution models. Unfortunately, the high tolerance for exceedances (30) of the 97.5% confidence level and the unconditional (Gordy and McNeil 2018) pairing with the 99% confidence level backtest does

11. Indicates an increased multiplier for capital calculation

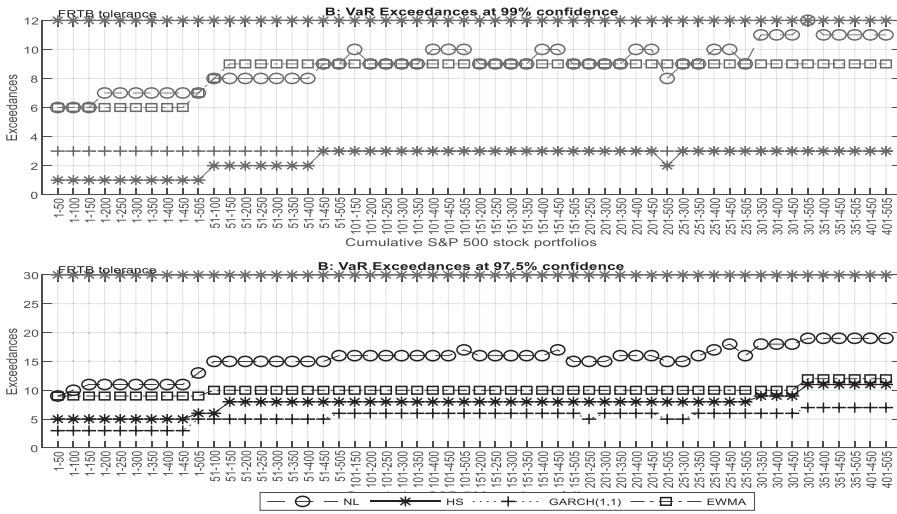


Fig 8. FRTB Backtesting Results for Cumulative Decile Portfolios (Portfolio Type B).

NOTE: The figure shows the exceedances for the four risk models considered (Normal Linear (NL), Historical Simulation (HS), GARCH(1,1), and EWMA) using a 250-day calibration period (09/06/17–06/06/18) and 250-day test period (07/06/18–06/06/19). FRTB exceedance tolerances of 12 and 30 for the 99% and 97.5% confidence levels, respectively, are also shown. Cumulative decile portfolios are considered (see Section 2). Portfolio type B involves ranking by market capitalization and equal weighting.

nothing to improve the power of the PLA tests. As described earlier, 30 exceedances corresponds to nine standard deviations from the expected value under the Bernoulli distribution. Therefore, we provide no evidence of a tangible incentive to deploy superior risk models under the FRTB methodology at two confidence levels at the desk level.

Examining the backtest results by portfolio characteristics, it is clear that portfolios that progressively drop higher capitalization-ranked stocks exhibit recursively declining results across all the backtests. This decline is common across each of the backtests and each of the risk models. However, the point at which the backtests identify and reject the portfolios and their risk models is at a lower decile (or cumulative decile) portfolio level for FRTB desk-level backtests than for the alternative backtests. This allows us to infer that FRTB desk-level backtests are inferior to several of the benchmark backtests (in particular the Basel II Traffic Light and Binomial tests) in their ability to identify and reject inferior performing VaR resolution models. This aligns with our findings thus far.

This accumulation of evidence from our quantitative impact analysis raises major concerns over the inability of the FRTB backtest specification to reject poorly performing risk models. Hence, we provide an evidence-based argument that FRTB is unlikely to incentivize the use of optimally performing risk model within banks,

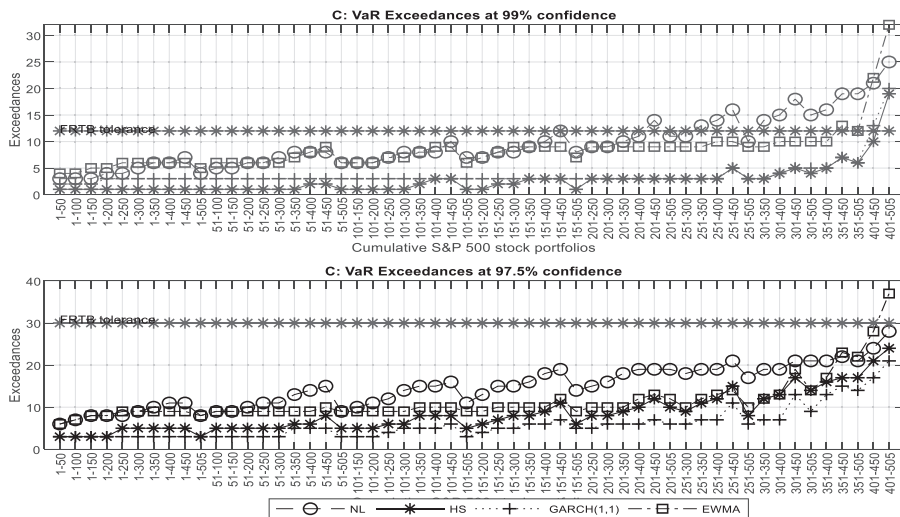


Fig 9. FRTB Backtesting Results for Cumulative Decile Portfolios (Portfolio Type C).

NOTE: The figure shows the exceedances for the four risk models considered (Normal Linear (NL), Historical Simulation (HS), GARCH(1,1), and EWMA) using a 250-day calibration period (09/06/17–06/06/18) and 250-day test period (07/06/18–06/06/19). FRTB exceedance tolerances of 12 and 30 for the 99% and 97.5% confidence levels respectively are also shown. Cumulative decile portfolios are considered (see Section 2). Portfolio type C involves ranking by Beta and value weighting.

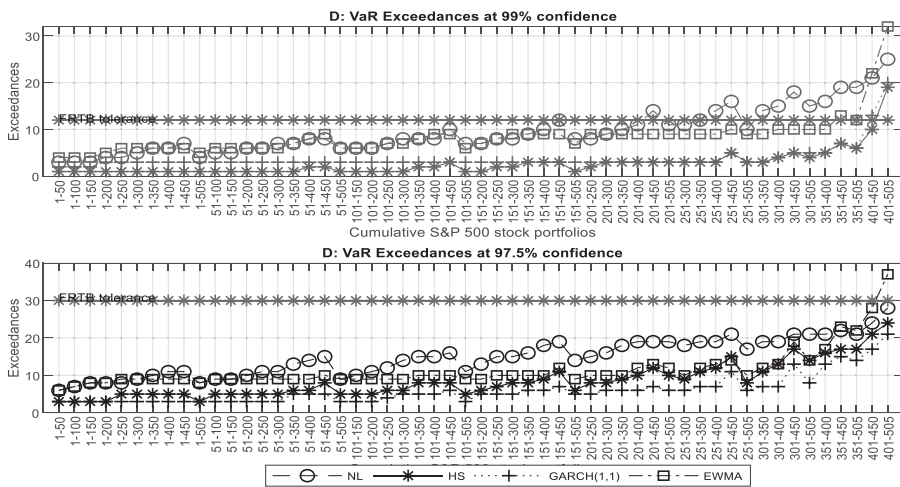


Fig 10. FRTB Backtesting Results for Cumulative Decile Portfolios (Portfolio Type D).

NOTE: The figure shows the exceedances for the four risk models considered (Normal Linear (NL), Historical Simulation (HS), GARCH(1,1), and EWMA) using a 250-day calibration period (09/06/17–06/06/18) and 250-day test period (07/06/18–06/06/19). FRTB exceedance tolerances of 12 and 30 for the 99% and 97.5% confidence levels, respectively, are also shown. Cumulative decile portfolios are considered (see Section 2). Portfolio type D involves ranking by Beta and equal weighting.

which has implications for capital adequacy requirements at the firm level and systemic risk at the sectoral level.

In summation, we find that FRTB desk-level backtests exhibit weak authority to identify and reject poorly performing models relative to many alternative backtests reviewed under the same data restrictions. The results from the backtests do not conflict with the results from the PLA tests. However, where the PLA tests show significant power to reject portfolios, the backtests do not exhibit similar power. The desk-level backtests are shown to be weaker than the entity-level backtests (similar to the Basel II Traffic Light tests). Portfolio diversification is naturally more challenging at desk-level than at entity level, which leads us to question why desk-level backtests are introduced under the additional IMA criteria of FRTB, particularly when, as we have demonstrated they are of weak power relative to viable alternative backtests. The answer may be that monitoring risk model performance at desk-level through these backtests may create the conditions for improvements in risk modeling through, what Borio and Zhu (2012, p. 238) describe as, the *framework effect*: “how banks actually perceive, manage and price risks.” Our study shows empirically, using stylized and industry portfolios, that the PLA tests present strong incentives for alignment of portfolios to the risk factors but that the desk-level backtests do not provide direct incentives for banks to deploy superior risk models. Furthermore, our examination of ETF portfolios reinforces the findings from the PLA tests on the stylized portfolios and highlights the significant impact of the PLA tests on portfolio design and/or the use of risk factor mapping in risk management.

4. CONCLUSION

The purpose of this study is to quantitatively investigate the impact of the introduction of two additional criteria for the approval of internal models in the determination of market risk capital, PLA tests and desk-level backtests, as prescribed under the FRTB framework. Specifically, we examine (i) their impact on portfolio management and (ii) if they incentivize the use of superior risk models.

The PLA tests are designed to assess the adequacy of the mapping of the portfolio to its risk factors. Risk factor mapping is a popular means of modeling large bank portfolios. We examine the performance of the PLA tests across a range of stylized and industry equity portfolios, the latter drawing on publically investable ETFs. The PLA test results confirm the importance of including high capitalization stocks in the portfolio for positive PLA test outcomes, ensuring adequate levels of representativeness of the equity index risk factor. Our study focuses on equity portfolios, but the approach we take can be extended readily to portfolios comprising other asset classes, and our study will prompt research in this direction.

There are several key findings from our study of the PLA tests. First, from our stylized analysis, we find that the FRTB-specified KS test is much more sensitive than the SR test to variations in market capitalization. Only portfolios that contain a

sufficient representation of stocks with the highest capitalization pass the KS test and only when they are value-weighted and reach a critical level of diversification (largest cumulated deciles). These results imply that passing the PLA tests will require the trading desk portfolios to be strongly aligned to the risk factors used for market risk analysis. Further, that they must hold the highest capitalization stocks of the index plus a critical mass of such stocks and that the weighting must be proportional to the stock's weighting within the index (value weighted). This means that passing the PLA tests will significantly affect construction of portfolios and encourage greater levels of passive management. Alternatively, the risk management practice of using risk factors to model the risk in the bank's portfolios may need to be replaced with full revaluation. The onerous nature of these options may cause banks to reconsider their use of internal models.

From our industry portfolios, we find corroborating evidence. We find that passive ETFs perform well under both components of the PLA tests, while active ETFs based on some form of stock selection generally pass the PLA tests when they include high capitalization stocks. Mid-cap and low-cap ETFs struggle to pass the PLA tests. This reinforces the findings from the stylized portfolios that the PLA tests will force a change in either portfolio construction through greater risk factor alignment or to risk management and the use of full revaluation. In either case, the introduction of PLA tests will have a significant impact.

We find furthermore that the performance of ETFs in the PLA tests deteriorates in the more volatile pandemic period chosen. This is most notable in the performance of the mid-cap and small-cap based ETFs, which fail one or both of the PLA test components. In a volatile period both the revised risk-sensitive SA and IMA implied capital will increase. If a period of high volatility also indicates a higher propensity for portfolios to fail the PLA tests, thus prompting a move to risk capital calculated on an SA basis, this presents a *double* capital hit for IMA desks. This is reminiscent of the issue of procyclicality of VaR, whereby increased market volatility led to increased capital requirements, which is particularly challenging if the volatility is indicative of an economic contraction (Danielsson, Shin, and Zigrand 2012, Adrian and Shin 2013, Vasileiou and Samitas 2020).

In complementary analysis, we also examine the performance of four simple risk models (NL VaR, HS, Equally Weighted Moving Average (EWMA), and GARCH(1,1)) under a spectrum of backtests. Hermsen (2010) characterizes the latter two models as having unreasonable assumptions. However, we find no incentive through FRTB desk-level backtests to discontinue the use of these flawed models. Further, we demonstrate that the FRTB desk-level backtests have low power to reject poorly performing models relative to alternative backtests under the same data restrictions. This indicates that the choice of risk model is not a priority concern in the FRTB framework. This is consistent with the findings of Burchi (2013), who argues that the increased complexity of the regulatory framework nullifies the significance of the choice of resolution model.

We conclude that the introduction of the PLA tests under the FRTB framework has the potential to significantly impact risk management and portfolio management

practice. The desk-level backtests do not exhibit evidence of their ability to incentivize the use of superior forecasting resolution models. This research provides interesting initial insights into the potential implications of the introduction of additional IMA criteria, in particular the introduction of desk-level PLA tests. This study lays the foundations for further research examining the complex risk factor mapping in the interest rate environment, further exploration of the universe of ETFs, and a systematic examination of the impact of volatility on portfolios' PLA test performance.

LITERATURE CITED

- Adrian, Tobias, and Hyun Song Shin. (2013) "Procyclical Leverage and Value-at-Risk." *The Review of Financial Studies*, 27, 373–403.
- Alexander, Carol. (2009) *Market Risk Analysis, Value at Risk Models*. United States: John Wiley & Sons.
- Angelidis, Timotheos, and Stavros Degiannakis (2009) "Econometric Modeling of Value at Risk." In *New Econometric Modelling Research Nova* (pp. 9–60), W. Toggins (ed.), New York.
- Armstrong, John, and Damiano Brigo (2019) "Risk Managing Tail-Risk Seekers: VaR and Expected Shortfall vs S-Shaped Utility." *Journal of Banking & Finance*, 101, 122–35. <https://doi.org/10.1016/j.jbankfin.2019.01.010>.
- Azoulay, Mark, Daniel Härtl, Yuri Mushkin, and Anke Raufuss (2018a) "FRTB Reloaded: Overhauling the Trading-Risk Infrastructure." 5: McKinsey & Company, Available at: <https://www.mckinsey.com/business-functions/risk/our-insights/frtb-reloaded-the-need-for-a-fundamental-revamp-of-trading-risk-infrastructure> [accessed 31/01/19].
- Azoulay, Mark, Daniel Härtl, Yuri Mushkin, and Anke Raufuss (2018b) "FRTB Reloaded: Overhauling the Trading-Risk Infrastructure." *McKinsey & Company McKinsey Working Papers on Risk*, 5, 24.
- BCBS (2018) *Regulatory Consistency Assessment Programme (RCAP)—Handbook for Jurisdictional Assessments*. Bank for International Settlements. Available at: www.bis.org. Accessed 31/05/2019.
- BCBS (2019) *Minimum Capital Requirements for Market Risk*. Bank for International Settlements. Available at: <https://www.bis.org/bcbs/>, Switzerland. Accessed 31/05/19.
- Beder, Tanya Styblo. (1995) "VaR: Seductive but Dangerous." *Financial Analysts Journal*, 51, 12–24. <https://doi.org/10.2469/faj.v51.n5.1932>
- Berkowitz, Jeremy, and James O'Brien (2002) "How Accurate Are Value-at-Risk Models at Commercial Banks?" *The Journal of Finance*, 57, 1093–111. <https://doi.org/10.1111/1540-6261.00455>
- Borio, Claudio (2003) "Towards a Macroprudential Framework for Financial Supervision and Regulation?" *CESifo Economic Studies*, 49, 181–215. <http://doi.org/10.1093/cesifo/49.2.181>
- Borio, Claudio, and Haibin Zhu (2012) "Capital Regulation, Risk-Taking and Monetary Policy: A Missing Link in the Transmission Mechanism?" *Journal of Financial Stability*, 8, 236–51. <https://doi.org/10.1016/j.jfs.2011.12.003>

- Burchi, Alberto. (2013) "Capital Requirements for Market Risks: Value-at-Risk Models and Stressed-VaR after the Financial Crisis." *Journal of Financial Regulation and Compliance*, 21, 284–304. <https://doi.org/10.1108/JFRC-10-2012-0042>
- Christoffersen, P.F., (1998). "Evaluating interval forecasts." *International economic review*, 841–862.
- Danielsson, Jon, Bjørn N Jorgensen, Sarma Mandira, Gennady Samorodnitsky, and Casper G De Vries (2005) *Subadditivity Re-examined: The Case for Value-at-Risk*. Ithaca, NY: Cornell University Operations Research and Industrial Engineering.
- Danielsson, Jon, Hyun Song Shin, and Jean-Pierre Zigrand (2012) "Procyclical Leverage and Endogenous Risk." *Available at SSRN 1360866*.
- Danielsson, Jon, and Chen Zhou (2016) "Why Risk Is so Hard to Measure." *De Nederlandsche Bank Working Paper*, No. 494. Available at: <http://doi.org/10.2139/ssrn.2597563>
- Demirguc-Kunt, Asli, Enrica Detragiache, and Ouarda Merrouche (2013) "Bank Capital: Lessons from the Financial Crisis." *Journal of Money, Credit and Banking*, 45, 1147–64. <https://doi.org/10.1111/jmcb.12047>
- EBA. (2019) *Basel III Monitoring Exercise—Results Based on Data as of 30th June 2018*. European Banking Authority. Available at: <https://eba.europa.eu/documents/10180/2551996/Basel+III+Monitoring+Exercise+Report+-+data+as+of+30+June+2018.pdf>, Paris, France. [accessed 31/01/2019].
- Farag, Hany (2018) "A review of the fundamentals of the Fundamental Review of the Trading Book II: Asymmetries, Anomalies, and Simple Remedies." *Journal of Risk*, 20(6), 99–128.
- Farag, Hany M. (2017) "Bracing for the FRTB: Capital, Business and Operational Impact." *Journal of Securities Operations & Custody*, 9, 160–77.
- Gordy, Michael B, and Alexander McNeil (2018) "Spectral Backtests of Forecast Distributions with Application to Risk Management." *Board of Governors of the Federal Reserve System*, No. 2018–021, 1–40.
- Haas, M. (2001). New methods in backtesting. Working Paper, *Financial Engineering Research Center*, Bonn.
- Hermesen, Oliver. (2010) "The Impact of the Choice of VaR Models on the Level of Regulatory Capital According to Basel II." *Quantitative Finance*, 10, 1215–24.
- Ibragimov, Rustam, and Johan Walden (2007) "The Limits of Diversification When Losses May Be Large." *Journal of Banking & Finance*, 31, 2551–69. <https://doi.org/10.1016/j.jbankfin.2006.11.014>
- ICMA. (2020) Regulatory responses to the market impact of COVID-19. Charlotte Bellamy. Available at: www.icmagroup.org [accessed 31/08/20].
- Jansen, Dennis W., and Casper G. De Vries (1991) "On the Frequency of Large Stock Returns: Putting Booms and Busts into Perspective." *The Review of Economics and Statistics*, 73, 18–24.
- Kupiec, P.H., (1995). Techniques for verifying the accuracy of risk measurement models (Vol. 95, No. 24). Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Board.
- Lazar, Emese, and Ning Zhang (2020) Market risk measurement: preliminary lessons from the COVID-19 crisis. In: Billio, M. and Varotto, S.(eds.) *A New World Post COVID-19 Lessons for Business, the Finance Industry and Policy Makers*. Innovation in Business, Economics & Finance 1. Edizioni Ca'Foscari, pp. 97–107. ISBN 9788869694424.
- Lee, Emily. (2013) "Basel III: Post-Financial Crisis International Financial Regulatory Reform." *Journal of International Banking Law and Regulation*, 28, 433–47.

- Li, Luting, and Hao Xing (2018) "Capital Allocation under the Fundamental Review of Trading Book." *arXiv preprint arXiv:1801.07358*.
- Lucas, André. (2001) "Evaluating the Basle Guidelines for Backtesting Banks' Internal Risk Management Models." *Journal of Money, Credit and Banking*, 33, 826–46. <https://doi.org/10.2307/2673897>
- Mahfoudhi, Ridha. (2018) "A Statistical Study of the Revised FRTB's P&L Attribution Tests." SSRN, available: <http://doi.org/10.2139/ssrn.3160327>
- Mehta, Amit, Max Neukirchen, Sonja Pfetsch, and Thomas Poppensieker (2012) "Managing Market Risk: Today and Tomorrow." 32. *McKinsey & Company*. Available at: https://www.mckinsey.com/~media/mckinsey/dotcom/client_service/Risk/Working%20papers/Working_Papers_on_Risk_32.ashx [accessed 31/01/19].
- Mokhtari, Mohamed, Robert Smith, Ridha Mahfoudhi, and Laurent Duvivier (2018) "A Statistical Study of the Newly Proposed P&L Attribution Tests." FRTB White Paper. Available at: <https://assets.kpmg/content/dam/kpmg/xx/pdf/2018/10/frtb-white-paper-july-2018.pdf> [accessed 31/08/20].
- Nield, S (2017) "The P&L Attribution Test—Going, Going, Gone?" Available at: <https://ihsmarkit.com/research-analysis/22062017-In-My-Opinion-The-P-L-attribution-test-going-gone.html> [accessed 08/04/2019].
- Nocera, Joe. (2009) "*Risk Mismanagement*." The New York Times, 4th January. Available at: <http://www1.idc.ac.il/Faculty/Kobi/RiskMGT/riskmgmt%20nyt%20nocera.pdf> [accessed 31/01/2019].
- O'Brien, James, and Paweł J. Szerszeń (2017) "An Evaluation of Bank Measures for Market Risk before, during and after the Financial Crisis." *Journal of Banking & Finance*, 80, 215–34. <https://doi.org/10.1016/j.jbankfin.2017.03.002>
- Pederzoli, Chiara, and Costanza Torricelli (2019) "The Impact of the Fundamental Review of the Trading Book: A Preliminary Assessment on a Stylized Portfolio." *CEFIN Working Paper*. Available at: http://doi.org/10.25431/11380_1197773
- Pogliani, Alessandro, Federico Paganini, and Marilena Rata (2019) "The Implicit Constraints of Fundamental Review of the Trading Book Profit-and-Loss-Attribution Testing and a Possible Alternative Framework." *Journal of Risk*, 21, 1–16.
- Pritsker, Matthew (1997) "Evaluating Value at Risk Methodologies: Accuracy versus Computational Time." *Journal of Financial Services Research*, 12, 201–42.
- Rossignolo, Adrian F, Meryem Duygun Fethi, and Mohamed Shaban (2012) "Value-at-Risk Models and Basel Capital Charges: Evidence from Emerging and Frontier Stock Markets." *Journal of Financial Stability*, 8, 303–19. <https://doi.org/10.1016/j.jfs.2011.11.003>.
- Soobratty, Essan, Eugene Stern, and Vicky Cheng (2020) *Revised Deadline Poses Further Challenges for Asia-Pacific Banks*. risk.net: Bloomberg. Available: www.risk.net [accessed 31/08/20].
- Spinaci, M., M. Benigno, A Fraquelli, and A Montoro. (2017) "The FRTB's P&L Attribution-Based Eligibility Test: An Alternative Proposal." *Risk*, 142–147.
- Thompson, Peter, Hayden Luo, and Kevin Fergusson (2016) "The P&L Attribution Test." SSRN. Available: <http://doi.org/10.2139/ssrn.2897911>.
- Thompson, Peter, Hayden Luo, and Kevin Fergusson (2017) "The Profit-and-Loss Attribution Test." *Journal of Risk Model Validation*, 11, 37–55. <http://doi.org/10.21314/JRMV.2017.180>.

Vasileiou, Evangelos, and Aristeidis Samitas (2020) "Value at Risk, Legislative Framework, Crises, and Procyclicality: What Goes Wrong?" *Review of Economic Analysis*, 12(3), 345–369.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A

Appendix B