Part 1: Strategic

Is Basel turning banks into public utilities?
Anthony Saunders, John M. Schiff Professor of Finance, NYU Stern School of Business

Excessive leverage and bankers’ incentives: refocusing the debate
Emilios Avgouleas, Professor (Chair) of International Banking Law and Finance, University of Edinburgh
Jay Cullen, Lecturer in Corporate Law, University of Sheffield

The deep roots of the government debt crisis
Horst Zimmermann, Professor Emeritus for Public Finance, University of Marburg

Financial innovation oversight: a policy framework
Carlo Gola, Regulation and Macroprudential Analysis Directorate, Bank of Italy
Antonio Ilari, Regulation and Macroprudential Analysis Directorate, Bank of Italy

The governance structure of shadow banking: rethinking assumptions about limited liability
Steven L. Schwarcz, Stanley A. Star Professor of Law & Business, Duke University School of Law, Founding Director, Duke Global Capital Markets Center, and Senior Fellow, The Centre for International Governance Innovation

Rethinking the treasury operating model
Roy Choudhury, Principal, EY U.S.
Contents

Part 2: Tactical

Do big banks have lower operating costs?
Anna Kovner, Research Officer, Federal Reserve Bank of New York
James Vickery, Research Officer, Federal Reserve Bank of New York
Lily Zhou, Graduate Student, Sciences Po

The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing
Arun Muralidhar, Adjunct Professor of Finance, George Washington University, Founder, Mcube Investment Technologies and AlphaEngine Global Investment Solutions
Kazuhiko Ohashi, Professor of Finance, Graduate School of International Corporate Strategy (ICS), Hitotsubashi University, Tokyo, Japan
Sungwhan Shin, Professor of Finance, Hongik University, Seoul, Korea

Capital budgeting: a “general equilibrium” analysis
Bradford Cornell, Visiting Professor of Financial Economics, California Institute of Technology

New private equity models: how should the interests of investors and managers be aligned?
Joseph A. McCahery, Professor of International Economic Law, Tilburg University
Erik P. M. Vermeulen, Professor of Business & Financial Law, Tilburg University

Auditing estimates: what will the future bring?
Paul de Beus, Partner, EY, The Netherlands
Maarten Koning, Partner, EY, The Netherlands
Is Basel turning banks into public utilities?

Anthony Saunders
John M. Schiff Professor of Finance, NYU Stern School of Business

Abstract
This paper looks at the effects of Basel III on the long-run profitability prospects of banks. It is argued that the new restriction on maturity mismatching and leverage will mitigate the “special” role banks play in the economy. Indeed, it will turn them into over-regulated public utilities with limited profitability and ability to innovate. The major thesis is that regulators appear to have forgotten that banks are “special” in terms of the economic benefits they provide to an economy, and that such benefits enhance social welfare.
Introduction
As every commercial banker, savings banker and credit union officer knows, Basel III will be phased in over the 2015-19 period. These regulations will not only affect BIS member countries, but it is also likely that countries worldwide will adopt similar regulations – in large part because of IMF and World Bank “conditionality” for loans.

In this paper, I briefly discuss the major features of the new Basel III accord. I will also discuss the two key factors that will have severe effects on future bank profitability, in particular, as I will argue in more detail later, because they can be expected to radically reduce the ability of banks to leverage their equity and to engage in traditional forms of “financial intermediation,” i.e., borrowing short term and lending long term so as to earn a borrowing-lending spread. I will argue that in the next few years bank RoEs (return on equity) can be expected to be reduced to the 8%-10% range, a range common among “safe” public utilities, such as gas and electric companies. Indeed, this raises the question about how banks are expected to generate profit in the Basel III era.

In Section 1 of the paper, I will give an overview of the key features of Basel III, which will reduce bank leverage and ability to exploit, in full, the benefits of financial intermediation in order to generate interest spreads between their borrowing and lending. In Section 2, I will look at the economic implications of these regulations for bank profitability and their potential demise into “public utility” type firms with low RoEs and low risks.
1. Overview of Basel III reforms

1.1 Leverage

As most bankers are aware, under Basel I and II the minimum capital requirements for credit risk were set at 8% of risk weighted assets – where the so called tier-1 capital had to comprise a minimum of 4% of that total. Tier-1 comprised of retained earnings, common stock and a limited amount of perpetual preferred stock.

Under Basel III, the 8% ratio stays unchanged, but the tier-1 component is raised to 6%, and a new “common equity” tier-1 ratio of just common stock and retained earnings is imposed at 4.5%. In essence, regulators will focus considerable attention on the core or so-called tangible equity of the bank.

But in addition to the 8% ratio, four new additional “add-ons” to that ratio will be introduced.

The first is a so-called conservation buffer of 2.5% to both total risk-based capital and the common equity tier-1 ratio, increasing them to 10.5% and 7%, respectively. The idea behind this buffer is to create some extra capital reserves for truly unexpected financial events.

The second is a “countercyclical” buffer that may be introduced by individual central banks ranging from 0% to 2.5%. The basic idea of this add-on is to offset the so-called “pro-cyclicality” effect that became a feature of Basel II. Essentially, what was discovered during the 2007-09 crisis is that as countries entered recessions and firms were down rated by rating agencies, banks’ risk-weighted assets rose and pressure was put on their risk-based capital ratio. This problem was exacerbated by the fact that most banks could not offset this effect by generating increased earnings and/or issuing new common stock, i.e., by increasing capital. To offset this pressure, many banks cut their high-risk-weighted loans, putting increased downward pressure on the real economy and making the recession worse. To offset this type of built-in procyclical effect, the countercyclical buffer will be used as countries grow fast toward the top of a business cycle – say China or India growing at the 6%-8% per annum range. With an overheating of the economy, it becomes an increasing probability that a downward correction or recession might follow. This is especially true if the “overheating” is in the property sector.
It is in periods of high growth that banks are generally more profitable and can raise new equity capital more easily, so it is at this time, before a recession starts, that a central bank may impose the additional counter-cyclical buffer ratio of between 0% and 2.5%. The basic economics of this is to create a so-called “rainy day” fund to be used in bad economic times. While this is highly unlikely to be imposed in the current Eurozone economies, the capital ratios of banks in countries like China may be raised by an additional 2.5%, resulting in a total risk based ratio of 13% and a common equity tier-1 ratio of 9.5%.

The third additional capital ratio is relevant to the 29 or so largest, most systematically risky banks in the world that are central to the global stability of the international banking system and are viewed as essentially being “too big to fail.” On the current list are the largest European, American and Japanese banks. Interestingly, despite having seven of the largest banks in the world only one Chinese bank – the Bank of China – made the initial list.¹

To determine the level of additional capital required for the globally systematically risky banks (GSB), the Basel Committee has developed a systematic risk index that is based on five factors of equal weight (i.e., 20% each). These factors are based on a bank’s cross-jurisdictional activity, its size, its degree of interconnectedness, the substitutability of its products and complexity. Some of these 20% weights are subdivided. Table 1 provides further indication of these weights.

¹ Other Chinese banks have since been added to the list. Industrial and Commercial Bank of China (ICBC) was added to the list in 2013, and Agricultural Bank of China (ABC) was added in 2014.
Based on the size of a bank’s total systemic risk score it can be placed in one of the five buckets of additional capital ratio requirements, ranging from a low of 1% (bucket 1) to a high of 3.5% (bucket 5). For example, based on their systemic risk scores 4 of the 29 banks have been placed in bucket 4 with an additional 2.5% capital requirement: Citigroup, Deutsche Bank, HSBC and JP Morgan Chase. Adding this up, the maximum risk-based capital ratio for a bank would be: 8% + 2.5% + 2.5% + 3.5% = 16.5%, i.e., the minimum risk-based capital ratio plus a conservation buffer plus the maximum countercyclical buffer plus the maximum systemic risk buffer.

Is Basel turning banks into public utilities?

Table 1: Indicator weights – indicator-based measurement approach

<table>
<thead>
<tr>
<th>Category (and weighting)</th>
<th>Individual indicator</th>
<th>Indicator weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Cross-jurisdictional activity (20%)</td>
<td>Cross-jurisdictional claims</td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td>Cross-jurisdictional liabilities</td>
<td>10%</td>
</tr>
<tr>
<td>2. Size (20%)</td>
<td>Total exposures as defined to use in the Basel III leverage ratio</td>
<td>20%</td>
</tr>
<tr>
<td>3. Interconnectedness (20%)</td>
<td>Intra-financial system assets</td>
<td>6.67%</td>
</tr>
<tr>
<td></td>
<td>Intra-financial system liabilities</td>
<td>6.67%</td>
</tr>
<tr>
<td></td>
<td>Wholesale funding ratio</td>
<td>6.67%</td>
</tr>
<tr>
<td>4. Substitutability (20%)</td>
<td>Assets under custody</td>
<td>6.67%</td>
</tr>
<tr>
<td></td>
<td>Payments cleared and settled through payments systems</td>
<td>6.67%</td>
</tr>
<tr>
<td></td>
<td>Values of underwritten transactions in debt and equity markets</td>
<td>6.67%</td>
</tr>
<tr>
<td>5. Complexity (20%)</td>
<td>OTC derivatives notional value</td>
<td>6.67%</td>
</tr>
<tr>
<td></td>
<td>Level 3 assets</td>
<td>6.67%</td>
</tr>
<tr>
<td></td>
<td>Trading book value and available-for-sale value</td>
<td>6.67%</td>
</tr>
</tbody>
</table>

However, beginning in 1996, Basel introduced an “add-on” to the 8% ratio for market risk. This was based on one of the two models: a standardized model and an internal model of a bank adopted with the oversight of bank regulators. This capital requirement is only really materially important for the largest banks, which have perhaps 10%-15% of their assets held in their trading books available for sale. The key charge here is to move from the 99% value-at-risk (VaR) concept for the required market risk capital for the trading book to the concept of expected shortfall based on extreme value theory.
Intuitively, expected shortfall extends beyond the 99% VaR to measure the average of the worst losses on the worst days in each 100-day period. Depending on the loss distribution of a trading asset, this may more than double the required market risk-based capital, and simulations have shown that it leads to an additional $\frac{1}{2}$% of required risk-based capital, resulting in a maximum of 17% plus any additional add-on for operational risk. The operational risk add-on introduced under Basel II in 2006 currently remains unchanged under Basel III. However, it is currently under review for revision in 2015. If we add 2% additional required operational risk capital to the credit risk and market risk capital requirements, we are put in the 19% range.

To some extent, this calculation rationalizes the approach of the Swiss regulators who have imposed a standardized risk-based capital ratio of 19% on banks such as UBS and Credit Suisse, known as “the Swiss finish.”

Given the above, it can easily be seen that under Basel III some banks’ leverage ratios could be reduced to only 5 (i.e., $1/20\%$) from 12½ (i.e., $1/8\%$).

### 1.2 Liquidity ratios

Under Basel I and II, there was no attempt to introduce any formulistic liquidity ratios for banks. Essentially, it was left to individual regulators under Pillar 2 of Basel to oversee individual bank liquidity positions. The financial crisis and failure of banks, such as Northern Rock in the U.K., a U.S.$6$ billion bank that relied for 75% of its liabilities on short-term wholesale funding, brought to the forefront the links between bank liquidity and solvency.

To ameliorate these types of insolvencies, Basel III has introduced two liquidity ratios: a liquidity coverage ratio and a net stable funding ratio. The liquidity coverage ratio requires the stock of high quality liquid assets to match or exceed the total net cash outflows from a bank projected to occur over the next 30 days, while the net stable funds ratio requires the available amount of stable funding to match or exceed the required amount of stable funding.

An extreme interpretation of these two ratios is that short-term assets should be sufficient to cover short-term liabilities and long-term liabilities should cover long-term assets. In aggregate, banks will be forced to more closely match the maturities (durations) of their assets and liabilities.
The economic implications for bankers are clear: they have traditionally made spread income from mismatching maturities, i.e., borrowing short-term and lending long-term as the yield curve normally slopes upward. If banks are forced by the two new liquidity requirements to more closely match the maturities of their assets and liabilities, then how will banks make spread income from “normal” financial intermediation?

2. Are banks becoming like public utilities?
What are the potential implications for banks in a world in which their leverage ratios have shrunk and their ability to engage in traditional financial intermediation is constrained? Analysts have made their views very clear by lowering projected RoEs for banks from the traditional 15% range to the 8% to 10% range.

Thus, we will have safer but lower economic returns to bank shareholders. In such a scenario, why would investors buy common stock in a bank? Indeed, the issue of where or how banks are going to raise or earn the new required levels of capital have hardly been discussed. In my view, the net effect of these changes in Basel III is that banks will become something similar to electric companies and other public utilities, earning a safe but low return on capital for investors. In addition, which investors will be willing to set up new banks? Perhaps it should be no surprise that in the U.S., a country of 7,000 banks, no new bank charters have been granted in the last three years. This absence of capital flowing into the banking system can only have negative implications for the supply and growth of bank credit. Moreover, how long will the banking system as we know it survive?

One possible profit opportunity for banks to generate profits remains, namely bank fees. In my opinion, it is not surprising that in response to the above Basel III implications banks have increasingly turned to fee generation as a source of profit. Apart from credit card fees, late fees etc., some banks are actually charging a fee for “storing” deposits, i.e., an implicit negative interest rate. The growth of banks’ reliance on fees can be seen from their most recent income statements. For the five largest banks in the U.S., 40% or more of their revenue now comes from non-interest income, which is largely fees.
Is Basel turning banks into public utilities?

However, even this avenue may soon be restricted. In the U.S., the Dodd-Frank Act of 2010 set up a Consumer Protection Agency for banking that has controlling bank fees high on its agenda. Likewise, the European Commission is starting to take a serious look at bank fees. Any such restrictions can only be expected to push banks’ projected RoEs even lower.

Finally, three questions should be asked:
1. Has Basel gone too far?
2. Has it outlived its relevance?
3. Has the whole capital calculation for banks become too complex and cumbersome?

With respect to the first point, my view is yes. Banks are “special” because they engage in financial intermediation that has large potential benefits to the real economy via credit supply and liquidity/money supply provision. Any regulation that seriously inhibits this process cannot be beneficial to the health of the overall economy.

With respect to the second point, the answer is a partial yes. Historically, Basel I was introduced in 1988 in response to the rapid expansion in international bank lending by the Japanese banks by utilizing their then lower capital requirements. A primary objective of Basel was to create a single capital requirement across all banks globally and thus induce a more level playing field.

However, over time, this objective has been increasingly eroded by some countries raising their minimum capital ratios, e.g., 12% in Brazil, introducing and allowing new forms of capital as part of tier-2 capital, and in the case of Basel III giving individual country bank regulators power to introduce country-specific countercyclical capital buffers.
Finally, with respect to point (3), on whether Basel capital requirements have become too complex, my answer would also be yes. Andrew Haldane of the Bank of England and the U.S. FDIC have recently made the simple point regarding complexity and the informational and computational inputs necessary. For example, banks using the internal ratings-based approach to calculate the 8% credit risk requirement and internal model-based approach to calculate the market risk capital requirement may need to calculate 5,000 or more parameters to estimate Basel capital ratio compliance. Perhaps, it is no surprise that at the last moment Basel III included a simple minimum 3% leverage ratio for credit risk to run side by side with the new risk-based capital requirements. Indeed, such a leverage ratio, with the addition of risk-weighted off-balance sheet items to total assets, does not look very different from what existed in most countries pre-Basel 1988!

3. Conclusion
This paper has argued that Basel III could have serious negative effects on the banking system and indeed the real economy. By leveling new restrictions on liquidity and leverage, the RoEs of major banks are expected to fall below 10%. This, of course, raises the question of how banks can survive as profitable non-public utility type firms. One current trend is obvious – fee generation. But in the long run, competitors will enter to limit fee growth and/or regulators will impose constraints on banks-fees. Thus, from a long-run perspective of “social welfare” and recognizing that banks are indeed “special” in terms of the services they provide to the economy, regulators need to seriously reconsider Basel III implementation.
Excessive leverage and bankers’ incentives: refocusing the debate

Emilios Avgouleas
Professor (Chair) of International Banking Law and Finance, University of Edinburgh

Jay Cullen
Lecturer in Corporate Law, University of Sheffield

Abstract
High leverage levels can lead to virtually limitless expansion of bank asset size, which maximizes, in the short to medium term, banks’ return on equity. In the absence of regulatory controls on leverage, all it takes to assume excessive risks, even for benign bankers, is to imitate competitor business strategies and herd. This form of herding is not solely motivated by compensation considerations, but also by career (job retention/promotion) concerns. Namely, while bankers’ compensation has been a major factor behind bank short-termism and excessive risk-taking, the availability of high leverage entails serious agency costs even in the absence of compensation incentives. As a result, regulatory reforms that focus on regulation of private compensation contracts ought to be supplemented by well-calibrated leverage ratios. Otherwise, they are bound to produce, in the long-term, suboptimal results, notwithstanding the conspicuous political gains of such a strategy.
1. Introduction
Leverage is normally understood as employment of borrowed funds in a way that allows a financial institution to increase potential gains or losses on a position or investment beyond what would be possible through a direct investment of its own funds. Leverage is an inevitable feature of banks’ role in providing credit intermediation and maturity transformation. With the advent of securitization and the explosion of shadow banking as a means of short-term financing, bank leverage in the mid-2000s reached levels previously unseen.

Under modern finance theory, a debt-financed corporation is as robust as an equity-financed one [Modigliani and Miller (1958)]. In addition, debt has traditionally been viewed as an effective corporate governance tool [Jensen and Meckling (1976)]. However, as we shall discuss in detail, these traditional maxims rarely — if ever — apply to banks, at least in the absence of strong regulatory constraints. High leverage levels can lead to rapid expansion of the size of bank assets, maximizing, in the short to medium term, banks’ return on equity (RoE). At the same time, (short-term) debt-fuelled bank capital structures increase bankruptcy risk, since they are an important cause of bank failures.

A great deal of scholarly and regulatory work since the global financial crisis (GFC) has focused on bankers’ pay and the perverse incentives embedded in it. These works have lent a sound empirical and theoretical grounding to public “gut feeling” and outrage with bankers’ irresponsible behavior. As a result, the tenor of contemporary post-GFC regulatory reform reflects the view that bank corporate governance deficits and perverse incentives embedded in compensation packages were significant causes of bank failures, since market discipline clearly failed in the pre-GFC era.

However, the role of leverage and its influence over bank governance and bankers’ behavior has remained largely unexplored. Yet unchecked availability of leverage creates perverse incentives in connection with bankers’ rent-seeking which, in turn, creates strong agency costs. On this basis, an in-depth examination of the counter-factual in relation to the GFC, which is largely missing from the “bankers’ greed” narrative, reveals a more nuanced picture. Bankers pursued risky strategies because they could, while the motive could be either job retention/promotion, or enhanced compensation, or both.
Excessive leverage and bankers' incentives: refocusing the debate

But would bankers have been so free to adopt and pursue risky business strategies if their ability to excessively leverage bank balance sheets was restricted by regulation? Clearly, there is a symbiotic relationship between leverage and short-term performance based compensation packages. Yet leverage building is not only motivated by compensation incentives, but also by job retention and promotion concerns, which in the financial industry are highly prevalent [Lakonishok et al. (1994); Chevalier and Ellison (1999); Hong and Kubik (2003); Brown et al. (2013)].

There is strong evidence for example, that the highly leveraged buyout of ABN AMRO by RBS in 2007, which reduced RBS's reported tier-1 capital ratio to 4%, was mostly driven by the career concerns of its most senior executive rather than pay targets [FSA (2011)]. The career concerns rationale is the only possible way to reconcile the narrative of “greed” [Bebchuk et al. (2010)] which is, arguably, based on sound empirical research, with the equally methodologically flawless line of research that shows CEOs in most of the banks that failed in 2008 were major shareholders and their compensation incentives were perfectly aligned with the banks’ long-term objectives [Fahlenbrach and Stulz (2011)]. Namely, career concerns and other behavioral explanations of bankers' behavior complement the greed narrative. Thus, it provides a more complete picture of what motivated bankers' conduct, given shareholder short-termism and industry peer pressure, which pushed bankers at all levels – but especially senior management – to adopt herding strategies.

It is therefore debatable whether the regulation of bankers' pay can eliminate in itself risks to bank and financial system stability emanating from bank employee behavior; it might not even substantially curb such behavior. But in its normative implications, such public intervention with private contracting is far from being all-encompassing, since the availability of high leverage impacts bankers' investment behavior and the size and complexity of the institution’s balance sheet regardless of the structure of bankers' compensation.

In the absence of unrestricted leverage, building bank size is a time-consuming and costly exercise that requires strong focus on relationships with existing and prospective clients. On the other hand, employment of leverage as the principal ingredient of bank capital structure, rather than share capital increases, is the cheapest and fastest way to build size – both in a relationship banking but, even more so, in a transactional banking environment.
Increased bank size and complexity present an ideal environment for hiding excessive risks and posting “inflated” short-term profits, which also increase the size of performance-based executive compensation packages. Given that excessive leverage can lead to institutional failure, financial instability and allocative inefficiencies, it is clear that leverage-induced executive decisions have governance implications and entail high agency costs, as management adopts business strategies that are harmful to the long-term financial sustainability and profitability of the bank.

This situation is in turn prejudicial to the interests of bank creditors and shareholders with a long-term view and undermines social welfare. In this paper, we define bank management’s capital structure-based decisions to choose suboptimal business strategies as “leverage agency costs.” Axiomatically, since bankers’ motives to build leverage are much wider than mere greed, controls on bankers’ compensation are probably insufficient to contain bankers’ risk-seeking. Arguably, leverage controls are an effective remedy whether in addition to or even in lieu of strict regulation of private contracts.

Our analysis is based on existing empirical research and does not present new empirical findings. We are, however, of the opinion that by highlighting the implications of the relationship between bank leverage, on the one hand, and bankers’ rent- and risk-seeking behavior, on the other, we contribute to a more nuanced understanding of how excessive leverage impacts on bank governance, including bankers’ risk-seeking, and of the conditions that make it possible. We postulate that, as leverage leads to virtually limitless expansion of bank asset size, which, of course, maximized — in the short to medium term — shareholder returns, even benign bankers have no other option but to imitate competitor business strategies and herd in order to retain their jobs or further their career prospects. Such behavioral motives entail serious agency costs, since in their striking selfishness they induce risk-taking, which is detrimental to the long-term stability of the institution and of the financial system. At the same time, they are also much less reprehensible from a moral point of view than if these strategies were motivated by mere greed.
In Section 2, we discuss the leverage cycle and its relationship with bank strategy. In Section 3, we summarize existing research on the interplay between bankers' incentives and bank leverage. In Section 4, we analyze briefly the trend of increasing leverage at large global banks prior to the GFC. In Section 5, we survey reforms (both enacted and proposed) to bank leverage ratios which, we argue, will restrain the capacity of bankers to increase bank debt, both due to career concerns and the quest for increases in their compensation. Section 6 concludes.

2. The leverage cycle and bank strategy

In many ways, the cyclicality of debt and impact of leverage on the economy as a whole and on the behavior of individual economic actors as well as on financial stability was best described by the late Hyman Minsky [Minsky (1992)]. His work is further developed today by a number of respected scholars exploring the financial stability ramifications of the leverage cycle [Bhattacharya et al. (2011)]. One of the most worrying characteristics of the leverage cycle is that while the risks it creates are often Gaussian it can also give rise to “fat tails.” Moreover, as leverage is cyclical, so is the rise and fall of asset prices. While the credit cycle should not be confused with the so-called “leverage cycle,” the main feature of the leverage cycle is also rising asset prices in tandem with rising leverage, followed by falling asset prices and deleveraging [Fostel and Geanakoplos (2013)].

The most important risk associated with leverage is the speed of deleveraging in a downturn as it may often prove difficult to prevent a “leverage cycle crash,” which critically will lead to increased margin calls (a so-called ‘margin calls spiral’) and probably to an evaporation of liquidity and a credit crunch [Brunnermeier and Pedersen (2009)]. Therefore, one of the most harmful effects of excessive leverage and sustained rapid credit growth – for both individual financial institutions and the financial system as a whole – is that it induces financial instability.

Intuitively, one would expect that in a fair-value environment a rise in asset prices would boost bank equity or net worth as a percentage of total assets. Stronger balance sheets would result in a lower leverage multiple. Conversely, in a downturn, asset prices and the net worth of the institution would fall and the leverage multiple would be likely to increase.
Contrary to intuition, however, empirical evidence has shown that bank leverage rises during boom times and falls during downturns. Leverage tends to be procyclical because the expansion and contraction of bank balance sheets amplify rather than contain the credit cycle. Fostel and Geanakoplos (2013) explain that the reason for this phenomenon is that banks actively manage their leverage during the cycle using collateralized borrowing and lending. When monetary policy is “loose” relative to macroeconomic fundamentals, banks expand their balance sheets and, as a consequence, the supply of liquidity increases. In contrast, when monetary policy is “tight,” banks contract their balance sheets, reducing the overall supply of liquidity [Adrian and Shin (2010)].

In other words, during periods of economic prosperity and low loan defaults, banks' capacity to create credit, coupled with inflated collateral values, increases banks' profits, which if retained as capital removes constraints on further credit growth [Turner (2010)]. In addition, some financial assets become very popular among a certain class of buyers in relation to the rest of the public. Buyers in this case are willing to pay higher prices, or tolerate increased risk. This is often due to optimistic expectations concerning the future price trajectory of a given set of assets. Unsurprisingly, such procyclicality contributes to higher risk-taking by banks. Moreover, investors – including banks – will borrow more to fund asset purchases, thus driving those prices up. The capacity to leverage balance sheets acts in tandem with this optimism to facilitate greater speculation on asset prices than unleveraged investors [Minsky (1992)].

On this basis, there are four plausible drivers of the adoption of excessive leverage. Firstly, managers will engage in rent-seeking [Shleifer and Vishny (1992)]. In the absence of bail-in safeguards, the strong possibility of a bailout means that monitoring by debt-holders is weakened and increased bank leverage does not generate a commensurate increase in the cost of debt financing. The possibility of a creditor bailout creates a situation in which the adjusted costs of monitoring by debt-holders may exceed its benefits and market discipline breaks down, leading to increased use of uninsured debt to fund bank assets [Diamond and Rajan (2001)]. The perverse result of this is that banks’ ability to increase leverage is enhanced when they take excessive risk [Acharya et al. (2010)].
Excessive leverage and bankers’ incentives: refocusing the debate

Secondly, bank management may use asset substitution to shift risks, a process made much easier through leverage. The possibility of regulatory arbitrage may mean that banks select asset portfolios with higher risk in order to maximize return on capital [Koehn and Santomero (1980)]. Risk-weighting of assets in bank portfolios ought to mitigate these effects somewhat (assuming those risk weights are accurate), but it is clear that banks push many risks off-balance sheet and actively manage risk-weights [Goodhart (2011)].

Furthermore, because leveraged institutions are also likely to be more opaque, evaluating the riskiness of their operations may be difficult. Banks’ ability to borrow heavily to alter financial risks permits them to engage in asset substitution more readily than nonfinancial firms and hide problems in their asset books [Merton (1977)]. In the presence of significant managerial incentives to adopt a leveraged capital structure, the ease with which risks are shifted increases incentives for banks to grow their balance sheets [Morgan (2002)].

Thirdly, the relative riskiness of bank asset portfolios is driven largely by the preferences of the category of each agent class (deposit guarantor, shareholder, or manager) that dominates a banks’ decision-making [Jeitschko and Jeung (2005)]. Based on this classification, banks that are dominated by shareholder decision-making favor the highest level of risk, followed by manager-dominated banks, and lastly, by banks dominated by deposit guarantors. To the extent, therefore, that capital requirements may influence bank behavior, governance arrangements in those institutions will inevitably be affected. It follows that shareholder-dominated banks will, in the absence of leverage constraints, pile up leverage, since shareholders face a clear conflict of interests when choosing a bank’s capital structure and they normally prefer to finance balance sheet expansion through debt.

Admati et al. (2013) have vividly explained the reasons for shareholders’ persistent preference for high leverage. While premature debt redemption in good times favors creditors as the bank will become safer and less likely to fail in bad times, thereby lowering creditor risks, bondholders will inevitably ask to hand in their bonds to be redeemed for a price higher than the prevailing market price. Thus, early redemption will eat into bank profits, leaving shareholders worse off without any clear compensating benefit to the share price, which might even decline as lower leverage clearly points to lower levels of future profitability.
This makes the up-front cost of early tax redemption undesirable to shareholders. Shareholder preference for high leverage thereby leads to the so-called “leverage ratchet” effect in which worthy projects are ignored by an over-leveraged bank in favor of lower quality loans.

Finally, financial innovation and the ability to “optimize” capital structure through leverage have been the key factors behind the growth of bank business over the last 20 years. In the case of transactional banking, leverage can be the catalyst for rapid building of bank asset size [Blundell-Wignall and Atkinson (2012)]. But, as leverage reduces the impact of trading positions on cash flows, it enables banks to reduce the amount paid at the outset of a financial contract, particularly when derivatives are used. Financial innovation has enabled financial institutions to trade in notional amounts, and to leverage their trading positions, on the basis of thin margins or borrowed collateral. This fuelled the expansion of trading book and (to a lesser extent) lending book assets, on the basis of a thin capital base.

These developments increased risk-taking not only because there was no obvious limit to the kind and size of financial bets a bank could take, but also because trading positions are easy to bury in a big bank’s balance sheet; a possible explanation for financial industry’s lobbying toward conglomeration in the 1990s [Avgouleas (2012, 2015)]. It follows that in the absence of controls on leverage, all it required to expand bank asset size, even for benign senior managers, was to imitate competitor business strategies and herd.

3. Bank leverage and bankers’ incentives

Remuneration packages are viewed as risk-reducing solutions to corporate agency problems: by tying the rewards of executives to shareholder interests, compensation plans purport to reduce risk to firm equity. In principle, the less stock managers own in a firm, the greater their incentive to appropriate private benefits, since any profits made by the company are distributed to shareholders. Thus, stock-based compensation is held as a solution to such agency problems, as it may be designed in an appropriate fashion to align the interests of equity owners in the firm with those of company management [Morck et al. (1998)]. In the past two decades, the award of stock options was the most popular method to align shareholder and managerial interests, and to act as a counterweight to their innate risk-aversion [Jensen and Murphy (1990)].
On the other hand, stock option awards may exacerbate agency costs, given management’s tendency to use mechanisms available to them to capture short-term gains at the expense of long-term performance [Cullen (2014)]. This risk is ever stronger in corporations where shareholders are inadequate monitors and boards are weak and unwilling to rein in executives wishing to influence the structure of their compensation packages and, arguably, in which managers enjoy greater capacity to use leverage to magnify returns. As demonstrated in Figure 1, equity-based compensation and stock-based bonuses were used heavily at financial institutions prior to the GFC.

**Figure 1: Senior (named) executive compensation proportions at U.S. financial firms**

Source: Balachandran et al. (2010)
Despite the theoretical logic underpinning the use of stock-based compensation, its popularity at banks in the pre-GFC period was grounded in a major fallacy; when it comes to banks, agency theory, which suggests that firm ownership structure determines shareholders' capacity to influence risk-taking, does not hold much traction. First, agency theory is limited when it comes to explaining the interaction between parties with financial interests in banks, because it assumes that firms operate in competitive and frictionless markets, where all actors are strong-form rational maximizers [Ciancanelli and Reyes Gonzalez (2000)]. Thus, it is assumed that they are immune to departures from optimization induced by cognitive biases, bounded rationality, and other sociopsychological pressures leading to herding [Avgouleas (2009)].

Secondly, bank managers are constrained in their actions because strong external forces, independent of the market – the regulator, the deposit guarantee scheme and others – exert control over financing decisions, and thereby bank governance. Finally, in the absence of a bail-in regime that shifts the risk of bank failure to creditors, these important corporate monitors feel shielded from business risk by the implicit guarantee of rescue in the event of insolvency, and thus remain unconcerned about management’s risky behavior, if they were, in any case, in a position to understand the complex risks banks may load up on and off their balance sheets [Adams and Mehran (2003)].

In the aftermath of the GFC, it became obvious that banking sector compensation systems did not align managerial incentives with firm sustainability and long-term profitability. It is arguable that banks’ capacity to assume very high levels of borrowing to fund assets led to excessive short-termism. Bankers have strong incentives to increase leverage in the absence of strong sanctions for failure: if an executive cashes in their shares prior to a drop in value, for them there is no downside risk attached to excessive leverage. Managers rewarded in stock options, in particular, have incentives to expand the bank’s balance sheet through increased leverage, and benefit generally from changes in the volatility of their options [Berger et al. (1997)]. On this basis, bank size and compensation levels can become positively related. By assuming greater levels of debt relative to equity and by the use of stock options in remuneration systems, top executives may increase their compensation levels; as Bebchuk and Spamann (2010) have put it: “[e]quity-based awards, coupled with the capital structure of banks, tie executives' compensation to a highly levered bet on the value of banks' assets.”
Another set of empirical studies have demonstrated that financial institutions with high “residual compensation” — that is, relative compensation levels after adjustment for firm size — were most likely to fail [Cheng et al. (2010)]. Perhaps more significantly, the higher the stock-option wealth within financial firms, the higher the bankruptcy risk of that firm [Armstrong and Vashishtha (2012)]. Where CEOs of banks receive a greater proportion of their remuneration in salary and bonuses rather than stock options, they are less likely to take high risks [Palia and Porter (2004)]. Because baseline assets are a significant driver of firms’ share price and executives often control significant stock options, any increase in baseline asset prices benefits an option-holder, even where these price movements are short term. Therefore, any increase in bank’s leverage that feeds into a bubble that inflates asset prices also feeds into the size of stock-based compensation for bank executives. Namely, leverage is a great accelerator of the value of bankers’ compensation packages based on options, and such compensation packages give bankers a very strong incentive to increase leverage.

Persistent focus in the banking sector on RoE, rather than on future risks, supplies strong evidence for this trend [Haldane (2011)]. It is clear that many firm executives acknowledge that there is a trade-off between long-term shareholder value and the pressure to realize short-term performance targets, even where meeting these benchmarks results in excessive risk [Graham et al. (2005)]. On the other hand, RoE is a poor indicator of overall long-term performance. Strong focus on RoE implicitly encourages a reduction in the amount of equity employed, which incentivizes banks to adopt leverage, increasing bankruptcy risk. Of course, as has been well documented, the rewards on offer for top executives at large banks were massive [Bebchuk et al (2010)].¹ Leverage, therefore, has a certain allure for executives as expansion of the asset base indirectly increases the value of their option compensation, and any bonuses linked to share price performance, since the more highly levered the institutions, the greater the speed of asset expansion.

Assuming that growth is constant and increased leveraging becomes a general trend, per-share earnings in the banking sector will rise, and be reflected in bank value.

---

¹ For example, the top executives (defined as the top two to five executives) at Bear Stearns cashed out U.S.$1.1 billion in stock between 2000 and 2007, and those at Lehman cashed out U.S.$850 million during the same period.
Since the rate of profitability accrues mainly through increasing net earnings per unit of assets, assumption of leverage [Minsky (1986)] is the best way for banks to increase profitability rates. Thus, unrestricted leverage has tremendous governance implications for banks as it embeds short-termism into bank business models.

As compensation systems themselves are procyclical, particularly during asset price booms, bank executives are incentivized to link compensation to asset price levels. For example, in the period leading to the GFC, financial institutions operated within an environment where risk incentives significantly induced managers to implement a more aggressive financial policy, manifested by progressively high levels of leverage and leading to higher downside risk [Van Bekkum (2013)].

However, risk-seeking behavior on the part of senior executives with major shareholding is paradoxical. Debt-fuelled balance sheets do not just boost earnings, they also amplify equity risk and contribute to the build-up of asset risk within investment portfolios, which increase both bankruptcy risk and systemic threats [Armstrong and Vashishtha (2013)]. Asset write-downs during the GFC were more strongly related to asset volatility in highly leveraged financial institutions: there was “a positive and significant interaction between [firm risk] and leverage” [Chesney et al. (2011)]. Accordingly, excessive leverage increases the prospect of bankruptcy that will wipe out shareholders (invariably including top executives), and destroy senior management’s reputation and future employment prospects. One explanation in relation to the GFC is that while markets were stable and returns from these products were healthy, asset portfolio riskiness was of little concern to top executives [DeYoung et al. (2013)].

But such short-termism is at odds with another line of empirical research, most notably that compiled by Fahlenbrach and Stulz (2011), which has shown that the incentive arrangements at large financial institutions were not responsible for bank failures or the creation of excess risk within the financial system. Senior management at financial institutions held significant equity positions and suffered substantial paper losses once stock prices began to fall sharply. Indeed, banks whose CEO interests were most aligned with the interests of shareholders performed worst. Large shareholdings ought to have incentivized senior bankers to act in the long-term interests of their firms and prevent the building of a high-risk asset base.
Of course, these assets became toxic once markets became totally illiquid and market prices collapsed, but it is highly improbable that CEOs and other senior executives with significant equity positions would knowingly take risks that could destroy firm equity, as this would not only result in obliteration of much of their wealth but also cause severe reputational damage. Yet CEOs of firms with relatively high equity stakes in their firms assumed the same level of risk as CEOs of firms with commensurately lower equity stakes. There were also no significant reductions in equity positions among bank CEOs post-2006, which meant that they bore heavy losses in the market crash of 2008; in fact, net CEO shareholdings increased. This would appear to suggest that even as the risk profile of bank investments appeared to change for the worse, bank CEOs did little to hedge their exposure to reduce any potential wealth losses. Such behavior is clearly inexplicable and certainly inconsistent with compensation-induced short-termism.

It is thus likely that greed and perverse compensation incentives are only half the story. Two other less sinister (behavioral) factors could have been at play with different degrees of influence: job retention/promotion motivations influenced by shareholder short-termism, on the one hand, and bounded rationality on the other. Bankers at institutions that failed did not think that they were assuming massive risks [Acharya and Richardson (2009)]. It is, therefore, at least arguable that senior executives “managed their banks in a manner they authentically believed would benefit their shareholders” [Grundfest (2009)], succumbing, at the same time, to peer pressure and their cognitive limitations.

They neither understood the risks that complex securities posed to their firms, nor the extent to which correlations in certain asset markets had been established across banking institutions. As noted by the FSA (2009): “[I]n benign markets with no recent history of negative events, managers will potentially take high risks, as they essentially underestimate low-probability, high-risk events ...” Moreover, given bank management’s ability to optimize capital structure, in many cases lack of controls on leverage was the decisive (rational yet perverse) incentive, rather than naked greed. This view is, in fact, less contrarian than it sounds, especially in light of the limited investment horizons of many large institutional shareholders, who hold shares on average for little more than a few months [Haldane (2010)].
It is clear from the discussion above that CEOs who risked everything were not just motivated by greed. They were also boundedly rational having genuine ignorance of some of the risks [Avgouleas and Cullen (2014)]. Moreover, they clearly succumbed to shareholder pressure to take risks across the board, irrespective of their individual equity wealth, especially as pay packages that implicitly encourage the use of leverage often have efficiency-decreasing effects: increased risk and leverage reduce the pay-performance sensitivity within financial institutions [John et al. (2010)]. The losses these executives suffered imply that the excessive risks present in the system were as much part of the developing trend to increase leverage to imitate competitors' results and the product of errors of judgment, as of misaligned incentives [Mehran et al. (2011)]. Indicatively, the FSA (2009) notes: “Mimicking the actions of other players in the market can be a rational strategy in two ways. Employees can ‘hide in the herd’ in order to conceal their performance or they can ‘ride the herd’ to prove the quality of their decisions.” Namely, perverse compensation incentives were not the only explanation of managerial short-termism. For example, one of the key findings of the FSA in its investigation into the failure of RBS was based on a management philosophy, which, among other things, emphasized the importance of “an overt focus on capital ‘efficiency’, i.e., on high leverage” [FSA (2011)]. This, of course, in the short to medium term invariably leads to increased shareholder returns. Already very wealthy, bankers had to increase leverage since short-termist shareholders would surely reward those in the bank who increased their returns with job retention and/or promotion opportunities. In effect, they operated under the assumption that missing quarterly earnings targets would be a risk to continued employment [Mergenthaler et al. (2011)].

This explanation does not exonerate senior bankers from blame. It simply provides a more pluralistic explanation of bankers’ behavior. It is also true that the “too-big-to-fail” factor played a role in such senior executive behavior, since the bigger the bank size and interconnectedness, the more certain the possibility of some form of a public bailout. But it was not the decisive factor. Senior management could not have failed to understand that in the event of a spectacular failure, they would lose their jobs, status and reputation.
Too-big-to-fail properties would not shield bank management from the consequences of failure. What they did was to exploit a cheaper funding base for years as creditors charged bigger banks interest rates lower than their risky business would warrant. This subsidy enticed all banks of a certain size to enter the race for RoE [Jacewitz and Pogach (2013)].

As we noted earlier, peer pressure leading to herding is a long-standing problem in the financial sector and its role in creating the conditions that led to the GFC has arguably not been fully appreciated [Lo (2009)]. At the same time, absent undesirable bank re-nationalization or granting increased voting rights to long-term shareholders, bank shareholders will remain short-termist. As a result, leverage controls, apart from their macroprudential impact, are also an effective means of controlling banker and shareholder short-termism, thereby containing leverage agency costs.

4. Bank leverage in the pre-GFC era
   To adduce empirical credibility to the aforementioned discussion of leverage agency costs, it is vital to review leverage growth trends in the lead-up to the GFC. In this period, leverage levels increased markedly among large commercial banks and investment banks alike. The drive toward building ever-higher levels of leverage remained mostly unchecked under prevailing capital regulations. Basel II, especially, essentially allowed commercial banks to largely set their own leverage levels [Turner (2010)].

   Estimates show that the asset levels at the 10 largest publicly listed global banks doubled between 2002 and 2007. In contrast, risk-weighted asset levels grew much more moderately, resulting in much increased overall leverage in the banking sector [IMF (2008)]. Superficially, the data suggests that banks were investing in “safer” assets, and were, therefore, permitted to expand the asset base without commensurately large increases in capital. His trend shows clear signs of herding in banks’ approach to capital structure, as demonstrated in Figure 2.
While asset levels increased markedly in the years leading up to the GFC, reported leverage levels at large commercial banks were remarkably constant. This would normally suggest that while banks expanded asset levels aggressively they maintained capital levels and preserved stable leverage ratios [Kalemli-Ozcan et al. (2012)]. However, official data does not provide the full picture. Leverage increases were caused by poorly calibrated internal financial models [Simkovic (2009)], the poor performance of credit rating agencies [Johnston (2011)], and fraud [Valukas (2010)].

**Figure 2: Balance sheet profiles for the 10 largest publicly listed banks 2002–07**

![Graph showing growth in total assets and risk-weighted assets](image-url)

Source: IMF (2008)
Moreover, there is strong evidence that reported leverage levels at both commercial and investment banks were manipulated, or were inaccurate, due to exploitation of prevalent rules on bank capital by bank management. For example, banks switched away from loans into structured financial products, which benefitted from higher capital relief. The increased role of complex securitized credit and marketable securities provided additional avenues to augment bank capital structures [Stein (2010)]. Under the Basel Accords, the lower risk weights that securitized products attracted meant that banks did not have to hold the same levels of capital against those assets, as it would be the case if the underlying products were not securitized. Basel II, in particular, made few significant changes to regulatory capital requirements in relation to conduits, leading to a reduction in overall capital requirements.

Much risk-weighted optimization (RWO) was achieved through employment of securitization models, as it was assumed that by diversifying and spreading risk throughout the financial system through securitization, the financial system would be more stable and more resilient to shocks [Blair (2013)]. These conduits raised funds by selling short-term asset-backed commercial paper, with the assets concerned usually comprising mortgage pools and secured loans. Because these conduits funded themselves with short-term debt, any loss of confidence or liquidity pressures due to a reduction in buyers of commercial paper would quickly destroy their viability, indirectly exposing the sponsor bank to funding liquidity risk.

The dual advent of risk-weighted capital requirements and financial innovation for funding has hereto enabled banks to engage in RWO for the best part of the past two decades. Indeed, research confirms that RWO has not abated since the GFC [Blundell-Wignall and Atkinson (2012)]. This has resulted in several large European banks operating with relatively low levels of common equity, despite being “well-capitalized” in terms of tier-1 risk-based capital (Figure 3).
Excessive leverage and bankers’ incentives: refocusing the debate

The data presented in Figure 3 demonstrates the inadequacies of RWO capital adequacy requirements, which, prior to the GFC, were easily circumvented through accounting mechanisms. Proposals have, therefore, been made to augment financial reporting requirements to include a strict leverage ratio to be reported by global banks to reduce the heavy reliance placed on imperfect risk-weighted ratios in capital adequacy regulation. In the next section, we provide a brief analysis of the utilities and disutilities attached to introduction of leverage ratios, as well as a concise overview of global proposals to introduce leverage ratios as either a reporting tool or a backstop capital adequacy requirement.

Figure 3: Selected European bank leverage ratios (total assets/tier-1 capital)

Source: Blundell-Atkinson and Atkinson (2012)
5. Regulating bank leverage

As we noted in the introduction, banks’ ability to leverage is fundamental to financial system development and economic growth. Accordingly, the introduction of binding leverage ratios in the banking sector has to balance several conflicting concerns, especially in relation to the potential adverse impact that leverage restrictions might have on the macroeconomic outlook. Besides, risks inherent in modern banking could arguably be contained by other forms of capital requirements, liquidity standards and the lender of last resort facility.

A first important argument against strict leverage ratios is that they may encourage banks to increase the riskiness of their asset portfolio, not decrease it, a classic Goodhart’s Law outcome. In the view of many, this concern was precisely the rationale for seeking risk sensitivity in the Basel framework in the first place [Haldane and Madouros (2012)]. In many ways, reluctance to regulate leverage stems from a fear that it will affect economic growth without making banks safer, since banks will evade the “crippling” profit consequence of a strict leverage ratio by focusing on high-return projects, which are riskier in most cases. Namely, a strict leverage ratio is bound to create perverse incentives [World Bank Group (2009)]. In addition, the lack of risk weighting in the calculation of the leverage ratio would penalize prudent banks holding substantial portfolios of highly liquid, high-quality assets.

Another argument used against restrictions on bank leverage is that such restrictions are bound to have an adverse impact on credit flows, and thus economic growth. This argument is supported by some recent studies [IIF (2011); DeAngelo and Stulz (2013)].

In contrast, however, the majority of contemporary studies show that leverage controls are an overall welfare enhancing mechanism. Conceptual studies show that to be the case when lending markets are competitive [Christiano and Ikeda (2013)]. Moreover, other recent empirical studies show that even in less than perfectly competitive markets, a binding leverage ratio has a beneficial impact on institutional and systemic stability without affecting growth [Junge and Kugler (2012); Miles et al. (2013); Eliot (2009)].
At the same time, there are strong arguments in favor of the introduction of strict leverage ratios. Firstly, research on the relationship between bankers’ equity-based compensation and the state of the relevant economy at any particular point in time suggests that bank managers’ choice of asset risk and leverage ratio corresponds to prevailing economic conditions. In times of economic expansion, executives target the maximum possible level of asset risk to maximize RoE, regardless of optimality concerns. The opposite is the case in a downturn when bank executives target low risk and safer investment. This embeds procyclicality in the financial system [Raviv and Sisli-Ciamarra (2013)]. If bankers are incentivized by either compensation or career-based rewards to increase leverage, restricting leverage will assist in reducing short-termism in financial markets, particularly because, as we have noted, excessive leverage generates significant governance/agency costs, and due to the aforementioned “leverage ratchet” effect.

Secondly, an unweighted leverage ratio is simple to apply and monitor, and eliminates regulatory arbitrage; namely, banks’ ability to engage in RWO, thereby restoring confidence in bank capital data. The previously discussed ability of banks to manipulate risk-weights - and thus their capital requirements - in conjunction with credit rating agencies’ well-documented inability to accurately rate the riskiness of very complex structured financial products, will always give rise to uncertainty about bank’s capital levels. Research by the BCBS (2013) confirms considerable variation across banks in the reporting of risk-based measurement of assets. While some degree of variation may be due to differences in the composition of trading assets, there are also significant difficulties in comparing banks from separate regulatory jurisdictions. For instance, there is a considerable degree of variation in, inter alia, the market-risk measurement methodologies employed by global banks, banks’ modeling choices, and accounting requirements and practices [Ingves (2013)].

In fact, investors have reported losses of confidence in the risk-weighting system and in the capacity of banks to calculate their levels of RWAs, even among specific asset classes [Beltratti and Paladino (2013)]. Recently reported wide discrepancies in British banks’ capital ratios when non-weighted assets are measured against own funds provide enough evidence to make any confident investor or prudent bank regulator lose sleep [Masters (2013)]. This uncertainty severely undermines rather than reinforces market discipline.
Thirdly, there is a growing body of empirical evidence that suggests that RWAs are not a significant indicator of the possibility of bank default. As mentioned earlier, in the simplest form of rule-gaming, banks turned high-risk credits into highly rated structured securities, thereby mostly eliminating capital requirements, although they extended credit lines to requisite securitization vehicles, which attracted no capital charges. Yet, provision of liquidity facilities to these vehicles exposed them to appreciable risks. Moreover, they held structured credit instruments on their own balance sheet, exposing themselves to embedded leverage and increasing their asset-liability mismatch and their funding liquidity risk. Prior to the GFC, Basel tier-1 capital levels were not a statistically significant predictor of default risk. When analyzed on the basis of unweighted leverage ratios, however, a significant statistical link was present [Blundell-Wignall and Roulet (2013)]. Accordingly, a simple leverage ratio is a much better predictor of default risk than RWA measures, which are invariably subjected to RWO. In the same mode, Haldane and Madouros (2012) have compiled a persuasive set of data showing that leverage has been a better predictor of bank survival than capital.

Finally, as mentioned earlier, contrary to intuition, empirical evidence has shown that bank leverage rises during boom times and falls during downturns, producing bubbles and crashes. The leverage ratio is versatile enough to be used both as a macro- or micro-prudential policy tool and as a countercyclical instrument. More specifically, by targeting individual financial institutions, a leverage ratio as a front-stop can restrict leverage building up at the systemic level [Fostel and Geanakoplos (2013)]. Even if there are other credit intermediation channels in the economy, (for example, shadow banking channels), and financial institution leverage is not the only important factor in building up the credit cycle, the fact that individual institutions are not excessively leveraged would mean that the speed of deleveraging in the economy will become much slower in a downturn.
For the aforementioned reasons, there exists considerable support for the introduction of stricter leverage ratios to supplement existing RWA capital disclosure or as a front-stop measure. The Basel III Accord requires that banks have a minimum of 6% tier-1 capital (comprising 4.5% common equity and retained earnings). In an attempt to counter the effects of inadequate disclosure of banks’ capital positions, the BCBS also requires, as a backstop, that banks operate with a minimum leverage ratio of 3%. Any bank with an unweighted leverage ratio of less than 3% will be deemed to be undercapitalized. Banks have been given until 2018 to fully comply with the implementation of the leverage ratio [BCBS (2010)].

Partly due to the extended timescale for Basel III implementation and partly due to the fact that the Basel III leverage ratio of 3% of non-weighted assets is arguably too low, certain jurisdictions have drawn up their own plans for imposing a leverage ratio and, indeed, have opted to go further than the 3% ratio. The U.S., for example, despite requiring its banks to report on a simplified leverage ratio prior to the GFC, has opted to strengthen its required ratio, mainly given the failure of the previous U.S. leverage ratio to produce any warning signs. The leverage ratio itself for all banking organizations is set at 4%, although certain large banking organizations are subject to a lower leverage ratio if they utilize the advanced IRB approach to RWAs, requiring them to have a minimum total leverage exposure of 3%. Moreover, a further supplementary rule has been introduced to cover the largest and most interconnected bank holding companies (BHCs), categorized as those with more than U.S.$700 billion in consolidated total assets, or U.S.$10 trillion in assets under custody (covered BHCs).

These banks would be required to maintain a tier-1 capital leverage buffer of at least 2% above the minimum supplementary leverage ratio requirement of 3% (for banks that use the advanced IRB approach), for a total of 5%. Failure to maintain own funds in excess of 5% would subject covered BHCs to restrictions on discretionary bonus payments and capital distributions. This rule would currently apply to the eight largest and most interconnected U.S. banks, although it will not be implemented in its entirety until at least January 2018.
Among other jurisdictions with significantly large banking sectors, strong progress has also been made in the U.K. The Independent Commission on Banking (2011) recommended that U.K. banks should maintain capital of 10% of RWAs, and that global systemically important banks (G-SIBs) headquartered in the U.K. ought to maintain at least 17% RWAs. Moreover, it concluded that large ring-fenced banks ought to be subject to a supplementary leverage ratio of approximately 4%. In a similar vein, the U.K. Parliamentary Commission on Banking Standards (2013) recommended “the leverage ratio [in the UK] to be set substantially higher than the 3% minimum required under Basel III.” Accordingly, the U.K.’s Financial Policy Committee (of the Bank of England) has recently announced a binding minimum leverage ratio of 3%, a supplementary leverage ratio of 1.05% for systemically significant financial institutions, and a countercyclical leverage ratio buffer, activation of which may result in a maximum total leverage ratio of 4.95% for large banks [Bank of England (2014)].

In the E.U., the European Commission’s proposal for a Directive and Regulation (collectively known as CRD IV) was ratified by the E.U. Parliament in April 2013, and implementation across Member States began in early 2014. However, while CRD IV requires that institutions calculate and report on their leverage ratios, CRD IV does not propose a minimum leverage ratio. The European Commission has charged the European Banking Authority with monitoring the implementation of Basel III, and reporting on whether the 3% minimum leverage ratio is appropriate for European financial institutions. If agreement is reached on the precise definition and calculation of the leverage ratio, a standardized ratio shall be implemented across the E.U. by January 2018. Institutions must, however, report their simplified leverage ratios to regulators from 1 January 2015.

Notwithstanding the aforementioned reforms, we remain quite far from a globally agreed leverage ratio that is well calibrated, instead of Basel III’s 3%, which is widely seen as a very weak constraint on banks’ risk-seeking. Moreover, there is a strong industry and regulatory preference in favor of using leverage ratios as a backstop measure [European Banking Federation (2013)]. Such a ratio provides better results in terms of measuring bank riskiness [Wolf (2014)] and in terms of macroeconomic stabilization, institutional soundness and improving bank governance when it is used as a frontline measure.
6. Conclusion
In spite of the externalities caused by increased bank bankruptcy risks, including the possibility of a costly public bail out, and by the financial instability and economic recession risks associated with excessive leverage, bank executives cannot be incentivized to adopt on their own moderate levels of leveraging. Whenever leverage remains unrestrained, shareholder and peer pressure to maximize returns makes short-termist behavior the only way to save senior bankers’ careers, generating massive governance/agency costs, even if misaligned compensation incentives play little or no part. As Admati et al. (2013) have shown, bank shareholders exhibit constant preference for debt finance over equity and have no incentives to ask bank executives to reduce leverage.

On the other hand, even benign bankers are intent on furthering their careers or, at the very least, on keeping their jobs. This situation is exacerbated when competitors also pile up leverage to maximize RoE. In this scenario, even benign managers will follow the short-termist path and herd, notwithstanding an alignment of their compensation to the bank’s long-term performance. Namely, the natural consequence of excessive leverage is ever more risk-taking and rent-seeking. Moreover, excessive leverage, in combination with the general opacity of bank balance sheets and asset substitution, creates intense information asymmetries between bank management and their monitors, which merely adds to the ineffectiveness of market discipline in the banking sector [Avgouleas and Cullen (2014)].
Excessive leverage and bankers' incentives: refocusing the debate

References

Acharya, V. V., H. Mehran, and A. Thakor, 2010, “Caught between Scylla and Charybdis? Regulating bank leverage when there is rent seeking and risk shifting,” Federal Reserve Bank Of New York Staff Report no. 469


Admati, A. R., P. M. DeMarzo, M. F. Hellwig, and P. Pfleiderer, 2013, “The leverage ratchet effect,” Rock Center for Corporate Governance at Stanford University working paper no. 146


Avgouleas, E., 2015, “Large systemic banks and fractional reserve banking - intractable dilemmas in search of effective solutions,” in Avgouleas E. et al. (eds.), Reconceptualising financial regulation

Avgouleas, E., 2012, Governance of global financial markets: the law, the economics, the politics, Cambridge University Press


BCBS, 2013, “Regulatory Consistency Assessment Programme (RCAP) – analysis of risk-weighted assets for market risk,” Basel Committee on Banking Supervision


Brown, N. C., K. D. Wei, and R. Wermers, 2013, “Analyst recommendations, mutual fund herding, and overreaction in stock prices,” Management Science 60, 1-20
Excessive leverage and bankers' incentives: refocusing the debate


Cullen, J., 2014, Executive compensation in imperfect financial markets, Edward Elgar Publishing


EBK - Eidgenössisches Bankenkommission, 2008, “EBK und grossbanken einigen sich auf höhere eigenmittelziele und die einführung einer leverage ratio,” Medienmitteilung Der Eidgenössischen Bankenkommission

Elliot, D. J., 2009, “Quantifying the effects on lending of increased capital requirements,” Brookings Institution working paper

EBF, 2013, “EBF response to consultation on revised Basel III leverage ratio framework and disclosure requirements,” European Banking Federation

Ingves, S., 2013, “From ideas to implementation,” Speech at the 8th High Level Meeting organized by the Basel Committee on Banking Supervision and the Financial Stability Institute


FSA, 2009, “Reforming remuneration practices in financial services: feedback on CP09/10 and final rules PS09/15,” Financial Services Authority

Fostel, A., and J. Geanakoplos, 2013, “Reviewing the leverage cycle,” Cowles Foundation, Yale University, discussion paper no. 1918


IMF, 2008, “Containing systemic risks and restoring financial soundness,” International Monetary Fund
Jacewitz, S., and J. Pogach, 2013, “Deposit rate advantages at the largest banks,” working paper
Johnston, A., 2011, “Corporate governance is the problem not the solution: a critical appraisal of the European regulation on credit rating agencies,” Journal of Corporate Law Studies 11, 395-441
Junge, G., and P. Kugler, 2012, “Quantifying the impact of higher capital requirements on the Swiss economy,” Mimeo
Masters, B., 2013, “The leverage story banks want to hide,” Financial Times, April 8
Mehran, H., A. Morrison, and J. Shapiro, 2011, “Corporate governance and banks: what have we learned from the financial crisis?” Federal Reserve Bank of New York Staff Report no. 502
Mergenthaler, R., S. Rajpopal and S. Srinivasan, 2011, “CEO and CFO career penalties to missing quarterly analysts’ forecasts,” working paper
Minsky, H. P., 1986, Stabilizing an unstable economy, McGraw-Hill Professional

Excessive leverage and bankers' incentives: refocusing the debate
Excessive leverage and bankers’ incentives: refocusing the debate


Valukas, A. R., 2010, Final report, bankruptcy court examiner, in re: Lehman Brothers Holdings Inc. et. al., Chapter 11 Case No. 08-13555 (JMP) US Bankruptcy Court of the District of New York

Van Bekkum, S., 2013, “Inside debt and bank risk,” working paper


Wolf, M., 2014, “Europe's banks are too feeble to spur growth,” Financial Times, October 28

The deep roots of the government debt crisis

Horst Zimmermann
Professor Emeritus for Public Finance, University of Marburg

Abstract
The current differences between euro countries in terms of government debt, deficit and restructuring efforts have much deeper roots than many acknowledge. Behind them are apparently long-term differences in tax mentality and in attitudes toward one's government and its borrowing policies. A cautious effort in 1974 to build a “pyramid of tax mentalities” turns out to be a good predictor for the same differences between the north and south, which can be observed in an international study of tax mentalities in the 1960s, a 1651 description of European countries and in the forecast of the euro countries’ performance right up to 2017. The results do not imply that the Eurozone has to fail because of its fiscal problems. A clear message from these results is that the structural adjustments in the southern countries might take much longer than is acknowledged at present. Should at the end of the day the euro area be considered as a case for realignment, then the proponents of what may be called the northern euro or a “euro of the strong ones” could draw on these results.

I would like to thank Brigitte Preissl for helpful comments on the structure of the paper, Wolfram Richter for the information about the Habsburg experience, Bertram Schefold for pointing out the importance of Kaspar Klock for the subject of this paper, Linus Mattauch for the hint about the Mezzogiorno example, Johannes Becker for the information on the work of Herrmann et al., and Aloys Prinz for pointing out recent sources on the government debt crisis.
1. Background and hypothesis
1.1 Starting point and sets of observations
The crisis in the Eurozone is now mainly a government debt crisis. It is expected to last much longer than the recent financial crisis and the ensuing economic crisis, at least in Germany [Zimmermann (2012)]. At present, the amount of debt per capita varies very much between the euro member countries. But it is interesting that even before the financial crisis started, there was a marked difference between mainly northern and mainly southern member countries. This paper argues that the roots of the present government debt crisis are much deeper than many wish to acknowledge, and certainly date back long before 2007, which was the last year before the financial crisis of 2008 started.

The argument in this paper starts from an international comparison of tax mentality that Günter Schmölders and his Cologne group undertook in the 1960s. This is the main source for this subject until today, because a similar in-depth survey of this kind does not seem to have occurred since then. In 1974, a special study came out that aggregated the results of these Schmölders studies plus some other comparative efforts [Tretter (1974)]. Tretter was courageous enough to sum up his results in what he called a “Pyramid of tax mentality” (Table 1).

**Table 1: “Pyramid” of tax mentality**

<table>
<thead>
<tr>
<th>Switzerland</th>
<th>Scandinavian countries, England</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Luxemburg, Germany, Austria, Netherlands</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------</td>
</tr>
</tbody>
</table>
|                      | ---------------------------------
| Spain, Belgium       | France                          |
|                      | Italy                           |
|                      | (Greece)                        |

Note: In 1974 Germany meant West Germany. Greece and the North-South divide were added by the author H. Z. Source: Tretter (1974, p. 49)
Though visually it does not look like a pyramid, the term “pyramid” is used throughout this paper because of its origin. Greece was originally not part of the pyramid. But the information below should permit us to add Greece to this table as well, and at the lowest level. Interestingly, some evidence that fits the pattern of 1974 can be found in an 800-page Latin text of 1651, which looked at the government systems and fiscal performances of several European, as well as a few non-European countries [Klock (1651/2009)].

1.2 The hypothesis
The hypothesis for this paper is that tax mentality and the concomitant borrowing habits are long-lasting attitudes that only change over a very long time period. This paper should, therefore, be viewed as part of a recent strand of economic thinking, and especially fiscal matters, that looks at historical developments as a way of understanding present day phenomena. An example of this area of research is a recent study by Becker et al. (2011), which detected marked differences in the quality of local administration between the former Habsburg possessions in today’s Eastern European countries compared with nearby areas that were not part of the Habsburg area. Another example is the “Eight centuries of financial folly” subtitle of Reinhart and Rogoff’s book on financial crises [Reinhart and Rogoff (2009)].

To put the hypothesis in concrete terms, the “Pyramid” of 1974 is chosen because it transformed the debate from one based on heuristics to one based on an ordinal scale, from weak to strong tax mentalities. It turned out that with the exception of Belgium, the differences in tax mentalities could be broken down between northern and southern countries. Therefore, I added the north-south divide. This pyramid is used throughout the paper as the blueprint to discuss (1) the results of the studies in the 1960s, (2) the 1651 Latin text and (3) the recent performance of the euro countries and the forecast until 2017. It will be demonstrated that the positions of countries on the pyramid only change marginally and almost never between the northern and southern half of the table, thus highlighting the deep roots of the present day problems.
1.3 The relationship between tax mentality and public debt performance

Historically, it has been found that countries in which the relationship between the
government and its citizens is good are able to finance themselves during difficult times
comparatively well through taxes, whereas weaker or more authoritarian states have to
resort to debt financing. The French scientist Leroy-Beaulieu stated in 1906 that during the
Crimean War (1853-1856) Great Britain financed almost half of its additional costs by way of
taxes, whereas France had to rely on debt for more than 90% of its funding [Leroy-Beaulieu
(1906)]. The German public finance scientist Lotz generalized this notion in 1931 and stated
that: “Financial history is full of examples where only strong governments can impose a
burden from extraordinary need on the present, whereas weak governments in the fiscal
sense always have to rely on borrowing in such circumstances” [Lotz (1931, p. 862)].

It is thus the attitude of citizens toward their government that dictates tax mentality and
debt performance. The determinants of tax mentality and tax morale have been analyzed
thoroughly by, for example, Schmölders (1960/1970) and Torgler (2003/2007). It will also be
shown below that one indicator of tax mentality, and of the attitude toward the state, is the
way in which taxes are levied. This is something that can be traced far back and is valid today.

This paper will be structured according to the three points in time mentioned above. It starts
with a look at the 1960s, followed by an analysis of the 1651 Latin text, and then looks at the
current government debt crisis and the forecast until 2017.

2. Well-founded comparison for the 1960s

2.1 The underlying studies

The idea behind taking a long-term view of tax mentality, and the data that many refer to in
their studies of the subject, emanated from “the Cologne school of tax psychology,”
organized by Günter Schmölders and his “Office for Empirical Research on Social Economics”
(Forschungsstelle für empirische Sozialökonomik).²

² The importance of their work was recently acknowledged again, pointing out that these researchers were
well-versed in the sociological methods of the time in working with interviews from a representative sample.
See above all Torgler (2003), pages 5 and 172, but also Smekal and Theurl (1994) and Döring (2014).
Schmölders and his team undertook several interviews in different European countries, using roughly the same set of questions, to determine respondents’ attitudes toward their government, taxes and tax fraud and — more as a side — to fiscal soundness.

In parallel to the studies of tax mentality and tax morale, the team undertook studies into taxing technique and taxing practice. This made it possible to relate both sets of results to each other: a sophisticated tax system with a large share of personal and corporate income taxes is only manageable if tax mentality supports it. The written tax laws were often much more sophisticated than the way in which they were actually implemented, for instance in France.³

Of specific interest is Tretter’s 1974 report, which combined and summarized the results of the two volumes, as well as many other studies that had been undertaken on the subject by either the Cologne school or other institutions. His “pyramid of tax mentality” (Table 1), in specific, will be helpful for the brief interpretation of the Klock study of 1651, and provides a good foundation for interpreting the situation today. The first three lines are taken together here as northern countries and the last four lines as southern countries. Belgium belongs geographically to the first group, but the fact that more than half of the territory and 40% of the population is French-speaking, as well as some of the data provided below, justifies the decision to assign it to the southern countries. The following remarks for the circumstances of the various countries in the 1960s, mostly drawn from Tretter, are organized by the sequence in this pyramid.

2.2 The results for the northern countries

- Switzerland was put for good reason at the top of the pyramid. In all answers concerning tax mentality the country ranks first. This is based on a strong affinity toward the country’s government. This is helped by the country’s low tax burden, which is in general considered to be helpful for a positive tax mentality.

³ The empirical results of the two sets of studies were published in two volumes [Beichelt et al. (1969); Daviter et al. (1969)], along with some questionnaires and additional volumes of tables, which were deposited in the research center.
The Scandinavian countries are usually given the same rank as England (which in Tretter’s work apparently stands for Great Britain). Sweden demands an income tax return even for low income, thus using a taxing technique that can only be implemented with rather high tax mentality.

England was the subject of several studies at the time. Most of them assign a very positive tax mentality to the country, again based on strong affinity to government, comparable to Switzerland. But having the highest income tax rates of all the countries in the study means that England’s tax morale is not too great, which is why it is not ranked in the highest tier. Nevertheless, “every second citizen believes in the just distribution of the tax burden” [Beichelt et al. (1969, p. 184)], an unusually high proportion.

The three Benelux states Belgium, the Netherlands and Luxemburg show very different tax mentalities. The Dutch seem to have a positive tax mentality, only slightly less than Germany. There is no information on Luxemburg, and for Belgium see below.

Germany has a good middle position for its tax mentality in international comparison. The interview results clearly suggest that the attitude toward government and the tax mentality are to be judged positively. This is the more astonishing since the German taxing technique was considered at the time to be characterized by sharp confrontation and a kind of perfectionism in taxing that would usually affect tax mentality negatively.

2.3 The southern countries

Among the Benelux countries with their different tax mentalities, Belgium is regarded as the worst and quite close to France in terms of results. This by the way is one of the reasons, apart from partly belonging to the Romance language group, why it was placed in the south, even though it is geographically in the north.

As for Spain, it might have been assumed at the time that because of its low socioeconomic development and its southern mentality it might have a negative tax attitude. But the studies demonstrated a relatively good tax mentality, and the government, as well as taxation, are mostly accepted. However, except for large businesses, no higher forms of taxation are applied, which would have meant a greater challenge for the loyalty of the taxpayers. Spain is, therefore, assigned the same level as Belgium in the pyramid. For the future of this euro country, it is important that “the tax mentality, as measured, should not stand in the way of success in taxation, which would be oriented toward ability-to-pay” [Beichelt et al. (1969, p. 184)].
France was at the time often mentioned as an example of bad tax mentality. Its history is rich in examples of fights between fisk and taxpayer. Tax code and comments on the income tax seem to indicate a modern income tax aimed at ability to pay. In reality, it is almost completely applied as a tax on estimated instead of actual receipts, i.e., the system is oriented toward indicators for cost and standard of living.

Italy is expected to be at the far negative end of the scale. Political culture, tax mentality and tax discipline are equally very bad. A large share of the direct taxes was not collected by the government tax authorities in the 1960s, but the right to collect them is given to third parties. This system is supposed to be abolished only very slowly, because it seems to be less costly than the administration of the government’s tax authorities.

Greece was not part of the international comparative studies. Schmölders mentions the country together with Spain, Italy and France as one, where sophisticated tax laws degenerate “to a mere farce” in daily tax administration [Schmölders (1970, p. 116)]. And the new Greek state in the early 19th century had started with high borrowing and early default [Schönhärl (2013)].

Taking the results of the observed countries together, they form the pyramid in Table 1. The researchers summarize their report by saying that: “The judgment on the fairness of the tax system in the three analyzed southern European countries [i.e., France, Spain and Italy] is devastating in international comparison” [Schmölders and Strümpel (1968, p. 149)]. In contrast, the judgment on the northern countries is generally positive.

3. A look further back
3.1 The origin in 1651
The voluminous work of Kaspar Klock in 1651 is well known and often quoted. However, it consists of about 800 pages of sophisticated Latin, which made it impossible for any scientist without a strong background in Latin to use, and sadly, according to Schefold’s private information, translations into German or other language does not exist. It was, therefore, very helpful that for the 2009 edition of the Latin text Bertram Schefold wrote a 100-page German introduction (plus many pages of notes). There he documented many of the detailed country-wise descriptions that Klock had given.
Klock concentrated on the Aerarium, which means the treasure of the ruler. At the time it mostly meant precious metals [Schefold, 2009, p. XLVIII], which fits the origin of “aerarium” from “aes” as metal. In the first volume, which is mainly referred to here, he deals with the various countries, “each with its particular institutions in the fiscal system, which are determined by topographical and cultural peculiarities and by the political setup, which in turn results from the specific historical development that Klock takes into account” [Schefold (2009, p. VIII)]. When Klock mirrors fiscal system and government system, he “adds the historical dimension, and if then one remembers that Klock tries to characterize national mentalities as well, one can speak of a comparison of economic styles” [Schefold (2009, pp. XI and LXVI)].

This seems to put Klock somehow in line with the analysis of tax mentality, as it was discussed in this paper. But whereas the studies in the 1960 analyzed the citizens’ tax mentality directly and then one had to link this to public borrowing habits, Klock’s context was different. He describes rulers in the first place and only occasionally adds a remark on the citizens. But here one can draw on the habits of rulers who may or may not pursue sound finances, and who ultimately are the ones to incur public debt.

The following remarks, which are exclusively based on Schefold (2009), are organized not by the sequence in Schefold’s introduction, but again by the arrangement in Table 1.

3.2 Klock’s remarks on the northern countries

- With regard to Switzerland, Klock notes that it does not seem to have any kind of fiscal system. However, “their big asset is parsimony.” They are poor, but “they thanked God for their freedom” [Schefold (2009, p. LXII)]. The economic base is diversified, and there are few beggars in the country. The city cantons, such as Basel, seem to run a reasonable system of excises. As for the soldiers, they are mainly mercenaries against pay and not hereditary subjects. Compared to the descriptions of the other countries, Switzerland certainly deserves a place in the upper part of the pyramid.

- The Scandinavian countries on average also receive positive marks. Sweden is dealt with most thoroughly. If this country has to carry the burden of wars, the money has to be approved within the corporate system. The state can rely on a rather diversified tax base, which is also extended to the nobility and to the church real estate. In general, its institutions are at the time more liberal than those of the German Reich.
In England, the king is in case of war supported by Parliament, which means that the ruler and the ruled are relying on each other. This had allowed the country to rely on taxes instead of borrowing in the Crimean War. On the revenue side, the country relies on a sophisticated system of excise duties. It seems obvious that England and Sweden (as the most important Scandinavian country) deserve a place in the upper tiers of the pyramid.

The Netherlands, which at the time included parts of modern Belgium, are dealt with extensively by Klock. In a summary statement, Schefold describes the country as the most interesting from an economic perspective and as the pioneering example of a republic with market economy structures in fields that one cannot find elsewhere. If taxes are sourced out to collecting agencies, then strict laws work against overstepping the competencies, which elsewhere seems to be a big problem with squeezing money from citizens and business.

Germany was not an easy country to judge at the time. It had barely survived the 30-year war (1618–1648), with large parts of the country devastated. The finances were split between the individual kingdoms, dukedoms, etc., and the emperor, with the emperor relying on little money in regular times and special allowances for special occasions. The “Reich,” as territorial unit under the emperor, did not possess enough tax base to keep it together and strong.

Looking only at Klock, one would probably put the Netherlands at the top of the pyramid. Taking the comments on the northern countries together, they would all be in the upper part of the pyramid, with Germany being for good reason in the third tier.

3.3 The southern countries

Spain was at the time the country with the widest colonial extension, and included Portugal. The Spanish king often went into debt hoping that additional revenues from the colonies would pay for it. Belgium is not mentioned as a separate country by Klock, because it partially still belonged to Spain.

France is considered to be of advanced statehood, and its riches are used parsimoniously.
Italy is not yet mentioned as a country, and Klock looks at various regions instead. Venice is discussed quite extensively. It borrows from rich citizens and seems to be well administered. With the exception of Venice, Italian public servants are considered to be particularly greedy and grasping, with Naples and Sicily being exceptionally bad.4 This north-south divide in the quality of the public sector might be taken as a precursor of present-day Italy, which seems to be divided between an economically strong north and a weaker south with its Mezzogiorno.5

Greece is at Klock’s time part of the Ottoman Empire and had been so for 200 years. The sultan relied on extraordinary revenue, and, though powerful, he is regarded as poor in comparison to Venice. Klock considers the Ottoman Empire as being typical of tyranny. In Schefold’s words “fear and subservience of the subjects, insecurity of the law, arbitrary distribution of sinecure, subservient presents instead of orderly taxation” [Schefold (2009 p. LXVII)].

Looking across the southern countries the picture is less clear. Venice could certainly be among the top countries, France receives good marks, Spain is prone to high public debt, and only Greece is difficult to judge because it is not yet a country of its own and will not be for another two centuries, so it remains to be seen whether the experience of more than 300 years of the Ottoman Empire could be overcome.

Of course, the results of Klock should be taken in context. More than 350 years have elapsed since then, so one cannot talk of continuity, given the changes in government systems and the enormous economic development in that time. Yet, the fact that they retain their positions on the pyramid throughout this period makes these remarks quite fascinating. They become even more interesting when one looks at the current situation – with the same geographic pattern as before.

4. The current government debt crisis

4.1 The situation before the financial crisis

After our look far back into European history, the discussion returns to the relationship between tax mentality and government debt performance, which in turn is at the heart of the government debt crisis.

---

4 Private information by Bertram Schefold.
5 This observation was referred to the author by Linus Mattauch.
As mentioned on the previous page, one might expect a rather strong correlation between a negative attitude toward government and taxation, and thus a bad tax mentality on the one hand and the readiness to accumulate large amounts of public debt on the other hand. Weck (1983) conducted a study of tax immorality for the period 1960 to 1978, with tax immorality standing for “the value of tax evasion” [Torgler (2003, p. 40)]. Table 2, which presents Weck’s findings, illustrates marked differences between north and south. The relationship was roughly 1:1.5 in 1960 and rose to more than 1.2 from 1970 onward, which means that the north-south divide has even widened somewhat during this period. The aforementioned Office for Empirical Research on Social Economics has continued to study tax morale in Germany and found that its already rather good tax morale even improved between 1997 and 2014 [Forschungsstelle (2014)].

**Table 2: Development of tax immorality, index figures**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Switzerland</td>
<td>0.6</td>
<td>1.0</td>
<td>1.2</td>
</tr>
<tr>
<td>Denmark</td>
<td>2.2</td>
<td>3.6</td>
<td>4.4</td>
</tr>
<tr>
<td>Finland</td>
<td>2.2</td>
<td>3.6</td>
<td>4.4</td>
</tr>
<tr>
<td>Norway</td>
<td>2.2</td>
<td>3.6</td>
<td>4.4</td>
</tr>
<tr>
<td>Sweden</td>
<td>2.2</td>
<td>3.6</td>
<td>4.4</td>
</tr>
<tr>
<td>Great Britain</td>
<td>2.2</td>
<td>3.6</td>
<td>4.4</td>
</tr>
<tr>
<td>Luxemburg</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td>5.5</td>
<td>9.0</td>
<td>10.9</td>
</tr>
<tr>
<td>Austria</td>
<td>5.5</td>
<td>9.0</td>
<td>10.9</td>
</tr>
<tr>
<td>Netherlands</td>
<td>5.5</td>
<td>9.0</td>
<td>10.9</td>
</tr>
<tr>
<td><strong>Average north</strong></td>
<td><strong>5.6</strong></td>
<td><strong>5.1</strong></td>
<td><strong>6.2</strong></td>
</tr>
<tr>
<td>Spain</td>
<td>7.1</td>
<td>11.6</td>
<td>14.0</td>
</tr>
<tr>
<td>Belgium</td>
<td>7.1</td>
<td>11.6</td>
<td>14.0</td>
</tr>
<tr>
<td>France</td>
<td>8.7</td>
<td>14.3</td>
<td>17.3</td>
</tr>
<tr>
<td>Italy</td>
<td>10.4</td>
<td>17.0</td>
<td>20.6</td>
</tr>
<tr>
<td>Greece</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Average south</strong></td>
<td><strong>8.3</strong></td>
<td><strong>13.6</strong></td>
<td><strong>16.5</strong></td>
</tr>
</tbody>
</table>

Note: Sequence of the countries as in Table 1. Averages unweighted because each government counts equally. Averages pertain to the countries listed with figures.

Given this background, the following arguments pertain directly to the government debt situation. In this last step of the paper, the findings of 1651 and of the 1960s are contrasted with the present situation. Table 3 presents the accumulated government debt from 1980 to 2007, the last year before the financial crisis started. By 1980, the welfare state, which in many cases began in the 1970s, had gained ground in European countries and was still progressing. The picture between the north and south is not very different, with the south being only four percentage points above the north in 1980. That changed in the 1980s. Until 1990, the figures for the north grew rather slowly whereas the south almost doubled its average accumulated debt as percentage of GDP. The negative outlier is Belgium, followed by Italy, but all others grew considerably as well.

Table 3: Government debt, % of GDP

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Switzerland</td>
<td>43</td>
<td>31</td>
<td>51</td>
<td>42</td>
</tr>
<tr>
<td>Denmark</td>
<td>39</td>
<td>62</td>
<td>52</td>
<td>27</td>
</tr>
<tr>
<td>Finland</td>
<td>11</td>
<td>14</td>
<td>44</td>
<td>35</td>
</tr>
<tr>
<td>Norway</td>
<td></td>
<td></td>
<td></td>
<td>52</td>
</tr>
<tr>
<td>Sweden</td>
<td>39</td>
<td>41</td>
<td>54</td>
<td>40</td>
</tr>
<tr>
<td>Great Britain</td>
<td>53</td>
<td>33</td>
<td>41</td>
<td>44</td>
</tr>
<tr>
<td>Luxemburg</td>
<td>10</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Germany</td>
<td>30</td>
<td>41</td>
<td>60</td>
<td>65</td>
</tr>
<tr>
<td>Austria</td>
<td>35</td>
<td>56</td>
<td>66</td>
<td>60</td>
</tr>
<tr>
<td>Netherlands</td>
<td>45</td>
<td>77</td>
<td>54</td>
<td>45</td>
</tr>
<tr>
<td><strong>Average north</strong></td>
<td><strong>34</strong></td>
<td><strong>40</strong></td>
<td><strong>48</strong></td>
<td><strong>42</strong></td>
</tr>
<tr>
<td>Spain</td>
<td>17</td>
<td>43</td>
<td>59</td>
<td>36</td>
</tr>
<tr>
<td>Belgium</td>
<td>74</td>
<td>126</td>
<td>108</td>
<td>84</td>
</tr>
<tr>
<td>France</td>
<td>21</td>
<td>35</td>
<td>57</td>
<td>64</td>
</tr>
<tr>
<td>Italy</td>
<td>57</td>
<td>94</td>
<td>109</td>
<td>103</td>
</tr>
<tr>
<td>Greece</td>
<td>23</td>
<td>72</td>
<td>104</td>
<td>107</td>
</tr>
<tr>
<td><strong>Average south</strong></td>
<td><strong>38</strong></td>
<td><strong>74</strong></td>
<td><strong>87</strong></td>
<td><strong>79</strong></td>
</tr>
</tbody>
</table>

Note: Sequence of the countries as in Table 1. Averages unweighted because each government counts equally. Averages pertain to the countries listed with figures.
The year 2007 has to be looked at more closely, because this was the situation in all of the countries just before the financial crisis started. Spain showed very sound finances, being below the average of the northern countries. At that point in time, there was no reason to think that Spain would not be able to face the crisis as well its northern counterparts did. It later turned out that it was not a basic structural problem that caused the crisis in Spain, but the large real estate bubble that had burst because of the effect of the crisis. France had increased its debt not beyond proportion, staying almost within the Maastricht limit of 60% of GDP.

Belgium had reduced its very high burden by about one-third between 1990 and 2007, and taking into account that this country had a sound economic basis, it could be expected that it would come out of the crisis in a rather good shape. It is the large increase in borrowing during the 80s and 90s that led the author to place this country among the southern countries. If during “normal” times a country can accumulate such high levels of debt, there must be something in the political and cultural system that makes it prone to high borrowing. An analysis for 1990 had found that “Flemish people … have a higher tax morale than the Francophone inhabitants” [Torgler (2003, p. 40)].

Things are different in Italy and Greece. These countries did not improve their debt ratios by any considerable degree until 2007 and entered the crisis with an amount of government debt that was above annual GDP. It was to be expected that these countries would run into difficulties when the financial – and subsequently economic – crisis would hit.

The general picture thus fits the pattern of the pyramid rather well. The most promising example is Spain, which in 1651 already had received rather good marks and did not do too bad in the surveys of the 1960s. Italy and Greece are the countries that should have caused most concern in that decisive year 2007.

4.2 Hope for the near future?
Today, it is of major interest whether the projections for the coming years show a different tendency in raising the south more than the rest of Europe and thus alleviating the government debt crisis in Europe.
Table 4 shows the development until 2017 for those countries in the pyramid in Table 1, which belong to the Eurozone. The picture is definitely not encouraging. To begin with government debt, the difference between north and south, which had been somewhat less than 1:2 (42:79) in 2007 as a percentage of GDP (Table 3), will not narrow, but in fact increase slightly (60:112).

Table 4: Budget deficit and unemployment rate, euro countries from Table 1

<table>
<thead>
<tr>
<th></th>
<th>Government debt % of GDP 2012</th>
<th>Budget deficit % of GDP 2012</th>
<th>Difference 2013-17</th>
<th>Unemployment rate % of labor force 2012</th>
<th>Average 2013-17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finland</td>
<td>53</td>
<td>-1.9</td>
<td>1.9</td>
<td>7.8</td>
<td>7.5</td>
</tr>
<tr>
<td>Luxemburg</td>
<td>21</td>
<td>-0.8</td>
<td>1.3</td>
<td>5.1</td>
<td>5.7</td>
</tr>
<tr>
<td>Germany</td>
<td>82</td>
<td>+0.2</td>
<td>0.0</td>
<td>5.5</td>
<td>5.2</td>
</tr>
<tr>
<td>Austria</td>
<td>73</td>
<td>-2.5</td>
<td>2.0</td>
<td>4.4</td>
<td>4.7</td>
</tr>
<tr>
<td>Netherlands</td>
<td>71</td>
<td>-4.1</td>
<td>0.7</td>
<td>5.3</td>
<td>7.1</td>
</tr>
<tr>
<td>Average north</td>
<td>60</td>
<td>-1.8</td>
<td>1.2</td>
<td>5.6</td>
<td>6.0</td>
</tr>
<tr>
<td>Spain</td>
<td>84</td>
<td>-10.6</td>
<td>4.1</td>
<td>25.1</td>
<td>27.0</td>
</tr>
<tr>
<td>Belgium</td>
<td>100</td>
<td>-3.9</td>
<td>1.6</td>
<td>7.6</td>
<td>8.2</td>
</tr>
<tr>
<td>France</td>
<td>90</td>
<td>-4.8</td>
<td>1.7</td>
<td>10.3</td>
<td>11.0</td>
</tr>
<tr>
<td>Italy</td>
<td>127</td>
<td>-3.0</td>
<td>1.3</td>
<td>10.7</td>
<td>12.2</td>
</tr>
<tr>
<td>Greece</td>
<td>157</td>
<td>-10.0</td>
<td>2.9</td>
<td>24.3</td>
<td>27.6</td>
</tr>
<tr>
<td>Average south</td>
<td>112</td>
<td>-6.5</td>
<td>2.3</td>
<td>15.6</td>
<td>17.2</td>
</tr>
<tr>
<td>Eurozone</td>
<td>93</td>
<td>-3.7</td>
<td>1.5</td>
<td>11.4</td>
<td>12.2</td>
</tr>
</tbody>
</table>

Note: Sequence of the countries as in Table 1. - Averages for north and south are unweighted because each government counts equally.
Data sources: EY (2013, p. 47); Government debt: German Ministry of Finance 2013, Table 16.

Of course, this figure does not change very fast. But it is driven by the budget deficits of the next few years. These deficits are more than three times higher in 2012 in the southern countries than in the north (Table 4). And the development of this indicator until 2017 sees the south still aggravating its problems.
Behind it are, of course, economic problems, for instance the high unemployment rate in the south, which is projected to be almost three times higher than in the north. And behind these economic problems are not short-term business cycle effects, as they could be seen in Germany for instance between 2008 and 2010. Instead, the problems in Italy and Greece stem from delayed structural reforms.

Structural problems might also be the reason behind Spain’s high unemployment, but as was mentioned before, the basic fiscal system of that country was sound before the crisis started.

It, therefore, looks as if the north-south divide, which was inherent in the pyramid of 1974, might persist far into this decade.

5. A comprehensive perspective
5.1 Analytical perspective
To begin with the facts: the initial hypothesis of the paper has been confirmed. Information from 1651, the 1960s and today demonstrate a similar international pattern in Europe. The northern countries have a more positive attitude toward government and taxation, and they lean more toward parsimony and fiscal austerity than southern countries. The method applied might be called the “recurrent pattern as research tool”: (1) observe a pattern, be it the “pyramid” here or be it an international pattern of frequent over-borrowing, then (2) search for a theoretical basis, for instance on the relationship between tax mentality and underlying attitude toward government, and finally (3) refine the hypothesis and then continue empirical analysis by observing the later patterns.

From an analytical perspective the question that has as yet not been addressed here is why the two types of countries are not spread around Europe, but instead are clustered closely in the north and, though less closely, in the south. A possible explanation, in addition to referring to a general notion like economic styles, could be provided by recent research on neighborhood effects [see Becker and Davies (2014), also for earlier literature]. Local proximity makes the exchange of information and experience easier. In the case of the north-south divide the similarity of language helps.
The south is, with the exception of Greece, characterized as Romance language, and the Germanic languages of the north play the same role. In addition, the greater proximity of the northern countries might also help explain why they are closer to each other in terms of tax mentality and fiscal attitude than the southern countries, which are geographically much wider apart. If one northern country has had a good experience with higher growth following fiscal austerity, the others – government and voters (or earlier rulers and the ruled) – could draw from that apparently successful experiment. “Countries react to each other” is an established result of recent research and could help to explain the north-south divide. Northern countries are also strong in terms of pro-social norms, which in turn supports a good tax morale, as opposed to – in this case – Greece and Turkey [Herrmann et al. (2008), table on p. 1363].

If the subject of this paper is regarded as interesting, then a research proposal might look promising. The results from the underlying studies in the 1960s could be taken as an opportunity for an interesting research today. At the time of the Cologne studies, Schmölders himself was involved in setting up the Central Archive for Empirical Social Research (Zentralarchiv für empirische Sozialforschung). It was meant to collect the methods and results of such empirical research in Germany. Some of the questionnaires for these studies on tax mentality in Great Britain, France, Italy and Spain have been preserved in this archive and in the Office for Empirical Research on Social Economics. In addition, separate volumes with basic univariate analysis (“Grundauszählung”) exist for some countries. This way, it should be possible to rerun these studies to see whether method and outcomes stand up to modern empirical research standards. In addition, and more interestingly, it would be possible to run a similar survey today. The results could be compared to those of the 1960s, so that the outcome might support the subject of this paper, namely that the roots of the present government debt crisis are deep.

5.2 Political perspective
Turning to the economic policy perspective, the results support the notion that the Eurozone is far from a Mundellian optimal currency area.

---

6 This source was referred to the author by Johannes Becker.
The paper shows that the economic differences between the member countries have much deeper roots than is usually acknowledged. Of course, not all of the countries included fit the pattern equally well. Spain, for example, was rather well-managed from the beginning; it entered the Eurozone with a very light debt burden, and its current crisis is more related to the property bubble bursting than to long-term fiscal burden. Germany, of course, had experienced two periods of very high inflation (1923 and 1948) following from an incredibly high debt burden, which at the time could only be relieved by currency reform.

Despite that, the fact remains that the northern countries remained in the upper part of the pyramid and the southern countries in the lower part, which makes the “Pyramid of 1974” a unique analytical tool.

Are there any conclusions from these findings that can be applied to the future crisis policy?

1. The results demonstrate that the Eurozone does not have to fail because of its fiscal problems. After all, the member countries are bound together by common values like the love for freedom and democracy, and the feeling that Europe belongs together, with Greece as the currently most problem-laden country being regarded as the cradle of European culture and civilization.

2. A clear message from these results says, however, that the structural adjustments in the southern countries might take much more time than is acknowledged at present. Attitudes toward the government and the concomitant tax mentality as well as the borrowing habits apparently have deep roots, and changes take longer than one thinks.

3. Should at the end of the day the euro area be considered as a case for realignment, then the proponents of what may be called the northern euro or a “euro of the strong ones” [Sinn (2013) and Sinn and Sell (2012)] could draw on these results. A recent perspective on this issue was provided by Allan Meltzer, who said that “the southern countries need their own euro.”

---

7 English version of the heading for his interview [Meltzer (2014)].
The deep roots of the government debt crisis

References

N. B. Quotations from German sources have been translated by the author.
Département fédérale des finances, Statistique financière, Bern
EY, 2013, “Eurozone forecast,” September
Lotz, W., 1931, “Finanzwissenschaft,” 2. ed., Tübingen
Reinhart, C. M., and K. S. Rogoff, 2009, “This time is different: eight centuries of financial folly,” Princeton
Schönhärl, K., 2013, “Geschichte eines Staatsbankrotts,” Frankfurter Allgemeine Zeitung, February 2, 21
Smekal, Ch., and E. Theurl, 1994, (eds.), “Stand und Entwicklung der Finanzpsychologie,” Baden-Baden
Financial innovation oversight: a policy framework

Carlo Gola
Regulation and Macroprudential Analysis Directorate, Bank of Italy

Antonio Ilari
Regulation and Macroprudential Analysis Directorate, Bank of Italy

Abstract
This paper proposes a policy framework for containing the unintended, and potentially harmful, effects of financial innovation. The total costs of conduct failures are often related to innovative products or services. The current approach, adopted by several authorities, makes extensive use of the tools of transparency and disclosure, mainly for consumer protection. There has been a growing recognition that this approach needs to be supplemented with more stringent organizational solutions, which mainly include corporate governance and risk management. In particular, we discuss the necessity to have in place a “new product committee” in order to promote a responsible risk culture and rigorous risk measurement, an appropriate remuneration policy, an explicit ethical standard, and, more generally, a set of internal controls related to both process and product innovation. In addition to retail financial products, the policy framework also deals with the harmful effects of complex and bespoke financial contracts for non-retail customers. This paper provides a number of practical examples of internal governance procedures, as well as discussing some more intrusive policy options, including guidelines, product pre-approval and restrictions. The suggested framework should be enforced through appropriate on-site visits.

We are very grateful to Domenico Albamonte, Riccardo Basso, Michele Leonardo Bianchi, Diana Capone, Agostino Chiabrera, John Kiff, Andrew Laidlaw, Marcello Minenna, Luca Zeloni and Luca Zucchelli for their very helpful comments and suggestions. This paper was prepared as a background document discussed at the EBA Task Force on Intervention Strategy – Standing Committee on Consumer Protection and Financial Innovation; it does not necessarily reflect the EBA’s policy stance on this topic. The opinions expressed are those of the authors alone and do not necessarily reflect those of the Bank of Italy. This paper is an abridged version of Gola and Ilari (2014).
1. Introduction
The recent financial crisis has highlighted the importance of financial consumer protection as a tool to promote trust and confidence in financial institutions, thus supporting financial stability and economic growth. At the same time, it has shown that the supervisory authorities need to consider the possibility of setting up a comprehensive system for capturing, monitoring and containing – possibly at an early stage – the unintended idiosyncratic and systemic negative effects of financial innovation on the financial sector and, more in general, on the economic system.

In the last decades, banking customers have gained broader access to increasingly complex financial products and the economic context has made consumers’ financial choices more difficult, calling for increasing attention from international organizations. Although there has been a great expansion in household assets and liabilities, often, individuals have not been supported by sufficient information about financial products and not adequately educated to make responsible decisions with full awareness of the associated risks. The phenomenon has often resulted in suboptimal choices and in the purchase of financial products not consistent with the individual’s needs. The situation has generated a number of problems, such as household over-indebtedness, excessive risk-taking, mistrust toward financial entities and in some cases litigation, with reputational and legal costs for banks, which include relevant compensations.

Behavioral economics can help to identify a range of cognitive biases that lead consumers to make choices that are not fully rational. Such biases affect financial choices especially in situations of marked uncertainty or complexity. Policymakers should help consumers to act in their own best interests, without becoming unnecessarily intrusive. Regulators are, therefore, moving in the direction of promoting more clear and effective information. Through specific public programs, financial literacy is enhancing consumer awareness and helping individuals make better financial decisions.

On the supply side, the creation and use of increasingly complex products can lead to excessive risk-taking and negative externalities. This behavior can be explained by various factors, such as wrong incentive structures that reward short-term profits or limited awareness of the risks associated with new, highly sophisticated financial products or contracts.
The complexity of financial products is often related to financial innovation, defined as the creation and development of new financial products or services for both consumers and producers. Financial innovation is driven by several factors, such as new forms of resource and risk allocation (risk sharing, portfolio diversification, hedging and risk management, etc.), as well as the search for temporary profits through price arbitrage, provision of liquid assets, reduction of transaction costs and so on. It can also be encouraged by regulatory arbitrage or strategies aimed more generally at circumventing existing legislation or regulations, including tax avoidance.

The net social welfare effect of financial innovation can, therefore, be either positive or negative depending on several factors. On the one hand, a more innovative financial system is superior in the sense that it helps to “complete the market,” to reduce frictions and transaction costs, and to increase liquidity. On the other hand, it can be a source of negative externalities (generating systemic risks such as excessive leverage or volatility), and suboptimal results for some classes of economic agents. Unethical behavior can further exacerbate such negative outcomes.

The suggested framework tends to shift attention from the product to the process. Consequently, in addition to retail financial products, the policy framework also deals with the harmful effects of complex and customized, bespoke non-retail financial contracts. This includes assessing the suitability of contracts that are likely to have a material impact on the counterparty, which could be either an intermediary or entities, such as nonfinancial corporations, municipalities, etc. Although simple and transparent financial products can also be a source of significant risks for the subscribers, our analysis mainly focuses on the risks created by excessive complexity and opacity. Conduct risk is, therefore, related to a much broader set of behaviors, affecting not only consumers but also producers and other financial entities.

In terms of policy options, in addition to the traditional transparency, disclosure, financial literacy policies and standards of conduct for financial providers (typical tools of financial market or financial conduct regulators), it is important to consider effective internal governance and organizational arrangements, including remuneration policies and, possibly,
the introduction of a “new product committee” (NPC). Where satisfactory results are not achieved by these means, more intrusive measures should, in principle, be considered. Specific legislation, warnings on certain products or services, and temporary prohibitions or restrictions should also be considered.

Before proceeding with the analysis, some methodological caveats are needed. Apart from a short passage on the necessity of adopting a paternalistic approach, and of introducing an ethics statement in the list of duties of the NPC, this paper does not pass any judgment on the social values of financial innovation outside the standard paradigm of utilitarianism.² Another important caveat concerns the point of view of the paper: while we underline the need to follow a consistent and coordinated approach with financial market regulators, our point of view remains that of the banking regulator, with the ultimate goal of preserving the soundness of the banking system.³

Unlike the field of consumer protection, where several policy papers are available, there are, to our knowledge, only a few comprehensive policy papers on financial innovation that discuss the nature of the phenomenon and why and how the regulator should intervene.⁴

² This analysis would require us to discuss in depth the relationship between duties and the social consequences of some decisions by also considering a non-utilitarian approach [see Sen (1991) and, for a discussion on non-utilitarianism, Sen and Williams (1982)]. A useful background for comparing and contrasting different ethical approaches can be found in Kutschera (1982). On ethics and finance see Boatright (2010). For a brief overview of the ethical aspects of financial innovation see Armstrong et al. (2011). For a specific analysis on ethics, banking and the recent crisis, see Koslowski (2011), part III.

³ For a recent discussion on complex financial products and transparency from the perspective of the financial market regulator, see Becker et al. (2012).

⁴ By contrast, the literature on the macroeconomic and social effects of financial innovation is extensive. For a recent contribution, see Arcand et al. (2012). Equally extensive is the literature on financial innovation and demand for money. For a recent paper on financial innovation and the role of information, see Piazza (2010). Among the few recent papers on the regulation of financial innovation, see Lumpkin (2009, 2010) and World Economic Forum (2012).
The paper is organized as follows. Section 2 provides a conceptual framework to describe and discuss the main aspects of financial innovation (definition, functions, evolution). Section 3 discusses how potential market failures and sources of risk can be identified, and Section 4 looks at their policy implications. Section 5 describes a set of general principles and the policy tools available (transparency and disclosure, financial education, internal governance, financial market technical standards, issue of guidelines or “warnings” in case of “serious threat,” restrictions or temporary prohibitions on financial products, product pre-approval) and Section 6 concludes the paper.

2. A conceptual framework for financial innovation

A broad definition of financial innovation: we define financial innovation as the creation and development of new financial instruments matching the needs of consumers, producers, financial intermediaries and governments. Innovation comprises both the “enhancement” of traditional financial instruments and the creation of “brand new” financial products or services. A financial instrument is a contractual right that has a monetary value, or represents a legally enforceable (binding) agreement between two or more parties; it can also be a physical product, specifically designed to support monetary or financial transactions.

Functions of financial innovation: the aim of financial innovation is to improve the utility of agents through various economic functions or objectives: 1) spatial and inter-temporal allocation of risks or financial resources (risk sharing, portfolio diversification, hedging and risk management, intermediation of resources between sectors); 2) extraction of information to support decision-making (pricing, rating); 3) search for temporary profits through price arbitrage both in the OTC and in organized markets; 4) provision of liquidity, safe assets and credit; 5) reduction of transaction costs (provision of medium of exchange, means of payment, services to support market trading and efficient price discovery); 6) reduction of agency costs and information asymmetries; 7) product differentiation or creation of new products, including in order to gain market power, and, 8) reaction to the normative environment (tax avoidance or evasion, accounting manipulation, regulatory arbitrage). A single innovation can simultaneously involve a bundle of such economic functions.

---

5 This classification is partially consistent with Tufano (2003), Lerner and Tufano (2011), Merton (1995) and Allen and Gale (1994).
Financial innovation oversight: a policy framework

**Policy implication 1:** Financial innovation involves a number of deeply intertwined economic functions, ranging from new forms of credit intermediation to various financial instruments or contracts. Such instruments can be traded both in organized or OTC markets and in the retail and wholesale sectors. Consequently, all traditional policy objectives (stability, efficiency, transparency, market integrity, orderly market functioning, fairness, depositors' and investors' protection), being closely interconnected, need an integrated and consistent policy framework and enforcement. For these reasons, in the area of financial innovation, strict cooperation between banking and financial market regulators is particularly important. This cooperation includes joint on-site visits aimed at verifying the enforcement of the policy framework proposed in this paper.

The scope of our analysis includes the above economic functions or objectives that are performed both by banks and by nonbank financial intermediaries generating bank-like risks (called the shadow banking system). The shadow banking system includes entities, such as broker-dealers, finance companies, asset management companies, investment funds (hedge funds, mutual funds, private equity funds, real-estate funds, money market funds, ETFs, etc.), securitization vehicles (SPVs, SIVs, etc.) and finance guarantors. It also includes activities that are a source of bank-like risks, such as nontraditional repo transactions or securities lending contracts. The suggested framework remains valid irrespective of possible structural transformations of the banking system (Volker rule, ring-fencing rule, etc.).

**Policy implication 2:** The scope of our policy action should include the financial innovation generated not only by banks but also by the shadow banking system. A pre-condition is a wide regulatory perimeter able to capture most of the shadow banking risks.

---

6 On the definition of shadow banking, see FSB (2011a). On shadow banking and the social value of financial innovation see Turner (2012).

7 This approach, recommended by the FSB to banking regulators, is also consistent with the following IOSCO principle: “The regulator should have or contribute to review the perimeter of regulation regularly” (Principle 7), IOSCO (2011).

8 On the regulatory perimeter, we believe the expansion of the area under regulation should be supported by the principle that entities undertaking similar risks should be subject to equivalent prudential standards, i.e., where the activity raises shadow banking risks, the same type of rules should apply consistently across financial sectors and jurisdictions, though tempered by the proportionality criterion.
Economic nature of financial innovation: innovation is the act of producing and developing new physical products, services or processes, among others, by means of organizational changes both within the firm and through its network. This includes the design, scaling-up and distribution of new financial products. In practice, this is a three-stage process: investigating potential demand, creating the product or service, and validating the company's internal procedures (to assess the product and minimize risks, including legal, operational and reputational risks). Financial innovation normally requires a combination of advanced quantitative finance, legal engineering and information technology.

While most financial innovation pertains to the evolution of existing financial instruments (e.g., increasingly complex forms of securitization), in some cases the innovative content is higher and more similar to a brand new product, for instance CDSs or ETFs. Consequently, as described by the standard theory of economic innovation, it is possible to distinguish between incremental and radical innovation. The former is built upon existing knowledge and simple product differentiation, while the latter refers to a major technological change or the supply of previously inexistent products or services.

The interaction between market structure and R&D intensity depends on several factors, such as large sunk costs, product differentiation, economies of scale, positive network externalities, etc. The possibility to patent financial innovation (for instance financial algorithms) is normally limited and, therefore, the dissipation of the appropriability advantages could be high. Supernormal profits for market leaders could leave room for transitory extra profit opportunities for followers, typically small dynamic firms. Financial innovation can involve both physical products (e.g., ATMs) and processes (electronic platforms for OTC markets, new financial algorithms for flash-trading, protocols for virtual currencies, etc.).

Some products, such as ATMs or smart cards, are truly innovative at their inception, but their characteristics and market diffusion soon stabilize, reaching the phase of maturity along the product life cycle S-shaped curve.

---

9 Among the few articles on the topic, see Lerner (2006) and Bhattacharyya and Nanda (2000).
10 On the diffusion process of innovation, see Hall (2003).
Other products are revitalized and mutate, creating a “jump” in the product’s diffusion in the market (Figure 1). Such mutations, which transform the characteristics of the original product, can be a source of new kinds of risk. This was the case for securitizations; initially, very simple and transparent financial instruments, they were subsequently transformed into more complex and opaque structures (e.g., the CDO square).

**Figure 1 - Product life cycle**
From a theoretical standpoint, recent literature underlines the importance of the evolution of a financial product’s life cycle. According to Gennaioli et al. (2012), several episodes of financial innovation share the following pattern: in some circumstances investors show a strong demand for a particular traditional financial asset; the excess of demand over supply stimulates financial intermediaries to financially engineer new securities with similar cash flows; as time passes, the new securities become increasingly complex and embed more risks. At some point, it could become apparent that the new securities are vulnerable to some unexpected risks, and in particular, are not good substitutes for the traditional securities. The result is that both investors and intermediaries are surprised by the news, and investors sell these false substitutes, moving back to the traditional securities with the cash flows they seek.

Policy implication 3: Monitoring financial innovation implies an effort to intercept new phenomena and their evolution at an early stage. To this end, the standard analysis based on public and supervisory information is necessary but not sufficient; it has to be complemented with targeted market intelligence and interviews with intermediaries, designed to give a clear understanding not only of the new products, but also of their qualitative evolution.

Financial innovation is often driven by normative and regulatory factors. For example, the development of constant-NAV money market funds (MMFs) in the U.S. in the early 1970s was a reaction to Regulation Q. Another example is the development of “special purposes vehicles” (SPVs) to reduce bank capital requirements in the 1980s (although securitizations have many other economic functions). What is interesting in MMFs and SPVs is their evolution: both were introduced decades ago, but only recently during the financial crisis did they become a source of major systemic risks.

11 From the early 1970s U.S. markets began to develop money market mutual funds with characteristics similar to remunerated sight deposits as a reaction to Regulation Q, which set a ceiling on interest rates that insured depository institutions were permitted to pay to depositors. Several European jurisdictions do not allow funds with the same characteristics (i.e., stable NAV based on amortization costs instead of mark-to-market accounting principles).
For example, constant-NAV money market funds are intrinsically unstable and prone to runs as they are forms of quasi-deposits without the fully fledged regulation of deposit takers (capital and liquidity buffers, deposit insurance, central banking facilities). However, for years they did not experience major systemic problems. It was only when they began to be part of a more complex intermediation chain with strong interconnections with the banking sector and the repo market (as well as with other shadow banking entities) that they emerged as a source of systemic risk. The lesson here is that the context matters and we should not limit our analysis to the financial instruments, but expand it to understanding the economic and legal environment as well. Moreover, the case of MMFs shows that abolishing Regulation Q did not reverse the market trend. Such inertia should stimulate the regulator to seek the deeper reason for this persistence, which could be rooted in some new form of regulatory arbitrage. Similarly, a specific macroeconomic condition (for instance prolonged periods of very low interest rates due to accommodative monetary policies) could force the market to develop new and more risky financial instruments.

Policy implication 4: Monitoring should not be limited to specific financial instruments or products, but should also be able to capture the interaction with the economic, legal and regulatory environment. Deregulation and prolonged periods of monetary expansion need strict monitoring and follow-ups.

Product and process innovation: in addition to the definition of traditional and innovative financial products (see Box 1), we believe that the distinction between product and process innovation is useful not only from the conceptual standpoint, but also for its policy implications. Indeed, since the crisis, regulators have increasingly supplemented the traditional approach of policy intervention (transparency, disclosure, point-of-sale monitoring, etc.) with an assessment of the design and the features of financial products, regardless of how they have ultimately performed for clients [FSA(2008, 2011)].

12 Recently, the Securities and Exchange Commission (SEC) introduced a mandatory transformation of constant-NAV into variable-NAV for some type of money market funds.
Box 1: Financial engineering and structured products

Traditional financial instruments are negotiable financial claims, such as standard debt securities, shares, and investment fund shares or units. Loans, deposits, trade credits and insurance technical provisions are examples of traditional nonnegotiable financial instruments. Typical innovative financial instruments are financial derivatives, which are not securities even if they are negotiable financial instruments. They are linked to specific financial or nonfinancial assets, or indices through which financial risks can be traded in their own right on financial markets. A combination of traditional financial instruments and derivatives is a source of incremental financial innovation. All combinations of financial instruments can be seen as bundles of cash flows. The aim of financial engineering is to understand how to price and hedge an instrument and assess the risks associated with it. In order to do so, it is necessary to consider the cash flows generated by an instrument during the lifetime of its contract. Then, using other (hopefully) simpler liquid instruments, these cash flows are replicated creating a new synthetic financial product or “contractual equation.” The evolution of structured products (defined as any security with a derivative component) is therefore part of financial innovation and should be monitored systematically. Complex structured products usually have the following features: (i) leverage; (ii) illiquidity; (iii) the potential for losses in excess of the initial investment; (iv) lack of price transparency; and (v) nonlinear payouts. On complex products, see ESMA (2014).

Product and process innovation are deeply intertwined. Despite the intrinsic difficulty of separating the two aspects, this distinction is very important from a prudential oversight perspective. As discussed in greater detail below, the creation of a new financial product can generate risks for the producers that are only indirectly (and not necessarily) transferred to the final user. Here again the example of the ETF fits very well: in principle, both physical and synthetic (swap-based) ETFs can generate similar or identical performances for investors (provided that, in replicating the same index, they have the same tracking error); however, they are based on very different processes with potentially different supervisory and prudential implications.

Policy implication 5: A single financial instrument and product can be produced through different processes, all of which the final user is unaware of. It is important, therefore, from the perspective of oversight of financial innovation to have a clear understanding of both product and process innovation. The latter can be a source of operational risk, as well as of other risks. On-site visits to regulated entities could be necessary to collect sufficient information on process innovation.

Theoretical developments are clearly a source of financial innovation. Markowitz’s (1954) portfolio theory, Arrow-Pratt’s (1963–64) measure of risk aversion, Sharpe’s (1964) Capital Asset Pricing Model (CAPM), Fama’s (1970) efficiency market hypothesis, Black-Scholes’s (1973) and Merton’s (1974) option pricing theory, and Ross’s (1976) arbitrage pricing theory are some examples.\(^\text{14}\)

All of these models are based on a very specific set of assumptions; for instance, they often rely on normal distributions. Until recently, such simplified models were widely used by finance professionals, despite clear empirical evidence that asset returns are often not normally distributed. Modern finance has developed a variety of models dealing with low-frequency high-impact events and with discontinuities; but these new theoretical frameworks (although more robust) are much more complex and computationally expensive [Rachev et al. (2011)]. Even the simplest model extensively used in finance, such as the possibility to capture investors’ attitudes towards risk with a single parameter (the Arrow-Pratt measure of risk aversion), has been put under scrutiny.

\(^{14}\) Pioneering contributions on several of these developments have been made by Paul Samuelson. See, in this regard, the paper by Merton (2006).
It has been argued that the risk is a much more complex concept that includes other characteristics, such as the degree of prudence and temperance.\textsuperscript{15}

**Market conditions and theoretical development:** theoretical developments are often driven by the emergence of new market conditions that make the adopted theory unsuitable [Brigo et al. (2013)]. For example, until recently, classical derivative pricing theories widely used the assumption that one can borrow and lend at a unique risk-free rate [Piterbarg (2010)]. In practice, this is no longer true. Since the crisis, different aspects are increasingly being taken into account to set up a proper pricing framework, such as funding, liquidity, credit and counterparty risk. Practitioners have begun to develop valuation formulas for derivative contracts that try to incorporate the modern realities of funding and counterparty risk, which deviate significantly from the risk-free textbook assumptions. Similarly, in response to the crisis, the classical pricing framework, based on a single yield curve used to calculate forward rates and discount factors, has been abandoned and a new modern pricing approach is prevailing among practitioners, one that takes into account the market segmentation as empirical evidence (e.g., Euribor-OIS spread, FRA rates-forward rates spread, and basis swap spread) and incorporates the new interest rate dynamics into a multiple curve framework (see Box 2) [Bianchetti and Carlicchi (2011)].

**Policy implication 6:** Regulators should have a deep understanding of the theoretical developments generated by financial innovation. In particular, while maintaining a frank dialogue with market players, they should develop an independent opinion of the model’s underlying assumptions and of their implications for intermediaries and the economy at large.

\textsuperscript{15} Analyses on the effects of risk attitudes on economic decisions have typically focused on the impact of risk aversion. Under standard expected utility theory, this amounts to assessing the impact of the second derivative of the utility function. However, many decisions crucially depend on higher order risk attitudes. For example, changes in precautionary saving due to changes in the distribution of a future income stream are determined by individuals’ prudence and temperance. Prudence is the sensitivity to risk of the optimal choice of a decision variable. The term “prudence” suggests the propensity to be prepared and forearmed in the face of risk, in contrast to “risk aversion,” which describes the extent of dislike of risk and the desire to avoid it where possible [Kimball (1990)]. “Temperance” reflects moderation in accepting risk. As observed by Kimball: we have a hedging position “when an unavoidable risk affects the freely chosen quantity of investment in another risk due to a correlation between the two risks. But it is reasonable to think that an unavoidable risk might lead an agent to reduce exposure to another risk even if the two are statistically independent. This tendency can be called temperance, in the sense of moderation in accepting risks” [Kimball (1992)]. See also Noussair et al. (2011).
Besides the specificities of various models, it is important that both intermediaries and regulators be aware of the limitations of the most widely used models and their underlying assumptions.

**Box 2: The current debate among “quants”**

Several financial engineers or quants, currently hold the view that we need a more holistic, comprehensive and realistic approach to finance.\(^\text{16}\) In addition to the difficulties of taking into account tail risks, skewed distributions and volatility clustering, the crisis has shown that it is critically important to properly model the correlations (or, more accurately, dependency structures) between risk factors.

It is now widely recognized that the copula function is not sufficient to model all types of interdependence, as it cannot explain the dynamics of dependence (i.e., how the dependency of all risk factors – market risk, liquidity risk, counterparty risk, etc. – changes over time). As underscored by Alex McNeil, “I still think some people really haven’t got the message that the correlation is just the first in an infinite sequence of numbers that drive interdependence.” “[What we need] is a proper understanding of the dynamic of dependence particularly between extreme events.” As noted also by Damiano Brigo, “the paradigm of western science – of analyzing pieces separately, then putting them together – does not work here. Everything interacts non-linearly, so we really need to understand the theory of multivariate processes.”\(^\text{17}\) Consequently, overconfidence in using new theoretical models should be tempered with an awareness that the models are necessarily based on simplified assumptions. This reflects the fact that the real world is complex and its properties cannot be fully explained in terms of individual components and their relationships. In addition, where probabilities are unknown and uncertainty dominates, a measurable risk metric is not available or is difficult to estimate.\(^\text{18}\) From the current debate, it is clear that there are many open questions, both theoretical and computational, and that a much more cautious approach to relying on current financial models is necessary.

\(^{16}\) Carver (2012) provides an interesting overview of the methodological problems that quantitative finance experts face in dealing with the complexities posed by the financial crisis.

\(^{17}\) For the quotations in this paragraph, see Carver (2012).

\(^{18}\) For an interesting discussion on the use (and limitations) of mathematics in finance, see Focardi and Fabozzi (2009).
3. Financial innovation and suboptimal outcomes: when to intervene?

Market and regulatory failures: like any regulatory framework, policy intervention should be underpinned by a comprehensive overview of various potential market failures and by an appropriate evaluation of possible regulatory failures, which refer to any post-implementation outcome that deviates negatively from what was expected.

Market failures arise when the free action of individuals in the market does not lead to an optimal allocation of resources (through the price mechanism) in the production process or in consumption; the most common forms of market failure are asymmetric information, externalities, imperfect competition generating excessive market power and supernormal profits, and market incompleteness for all contingent claims. In financial markets, such market failures are not only a cause of inefficiency for single entities, but also a source of instability (contagion) when they assume the form of systemic negative externality among intermediaries.

Market failures in trading innovative products: the retail sector: we know that if consumers are not fully informed or are not able to maximize their expected utility and inter-temporal plans, they can produce suboptimal outcomes. A notable case is when consumers lack access to information (incomplete or asymmetrical information), for example about the solvency of financial institutions. Incomplete information can lead consumers to buy unfair products, for instance, because they are mispriced or have an embedded implicit fee.

A very different situation arises when agents have bounded rationality. Even if fully informed, they are often not able to optimize their financial decisions, either because of computational difficulties (due to excessive complexity) or because of cognitive biases.\(^{19}\)

\(^{19}\) On consumer financial behaviour, see Guiso and Sodini (2012).
i) In standard economic theory, individuals make choices maximizing a utility function, using and processing appropriately all available information. Individuals’ preferences are assumed to be time consistent, affected only by their own payoffs, and are independent of the context (or framing) in which such decisions are taken. The literature on behavioral economics, however, suggests that individuals often deviate from the traditional models because they have nonstandard preferences, incorrect beliefs and systematic biases in their decision-making process.\(^{20}\) The context and the framing of the situation matter: two equivalent decision problems that are framed differently may elicit different responses.\(^{21}\) In many decision problems, “loss aversion” induces a bias that favors retaining the status quo over other options. Moreover, individuals often underestimate the probability of negative events; beliefs are affected by systematic overconfidence, overinference from past events and by overprojection of current tastes on future tastes. A better understanding of the above behavioral biases helps the financial conduct regulators detect and remedy problems arising from not fully rational economic agents. For instance, the regulator could require a more effective and fair disclosure, where different options are framed fairly and not so as to lead the consumer to make a wrong choice [FCA (2013)]. While this is an interesting path, we should be aware of the difficulties that the regulator could face in further expanding the scope of interventions. As observed by Tversky and Kahneman (1991): “We conclude that there is no general answer to the question about the normative status of loss aversion or other reference effects, but there is a principled way of examining the normative status of these effects in particular cases.”\(^{22}\)

\(^{20}\) For an overview, see S. DellaVigna (2009). This approach has recently been adopted by the Financial Conduct Authority (2013).

\(^{21}\) According to Tversky and D. Kahneman, the outcomes of risky prospects are evaluated by a value function that has three essential characteristics: “reference dependence” (the carriers of values are gains and losses defined relative to a reference point), “loss aversion” (the function is steeper for losses than for gains) and “diminishing sensitivity” (the marginal value of both gains and losses decreases with their size). An immediate consequence of loss aversion is that the loss of utility associated with giving up a valued good is greater than the utility gain associated with receiving it. [Tversky and Kahneman (1991)].

\(^{22}\) As observed by Shiller (1999), while it is difficult to find systematic patterns of behaviour that can be codified in a general theory, we cannot say that “anything can happen” in financial markets; moreover, “It is critically important for research to maintain an appropriate perspective about human behaviour and an awareness of its complexity” [Shiller (1999)].
ii) Nonrational behavior, such as over- and under-reactions, has been extensively considered by financial analysts as well and has implications for the efficiency of financial markets. We know that if the equity market is efficient, market prices should tend to be equal to the expectation of the present value of all future dividends, discounted at the appropriate risk-adjusted cost of capital and conditional on all available information. However, in the real world, a number of factors can explain short-term deviations from such highly stylized market conditions. In principle, even without relying on forms of fully optimizing behavior (but simply on trial and error) such deviations should disappear in the long run as the arbitrage mechanism should wipe out all profit opportunities. This mechanism, however, relies on the assumption that market forces are sufficiently powerful to overcome not only any type of behavioral bias (over- and under-reactions, herding behavior, sudden shifts of risk aversion, etc.), but also the uncertainties caused by an ever-evolving environment. In order to deal with some of these problems Andrew Lo and others, inspired by the work of Herbert Simon on bounded rationality and by the analysis of Edward Wilson on sociobiology, suggested using the “adaptive market hypothesis,” in which the cyclical changes in risk preferences (and therefore the risk premium) are shaped by the forces of natural selection (to explain or endogenize not fully rational behavior) [Lo (2005)].

iii) Another stream of behavioral models has tried to explain the wide and persistent deviations of asset prices from the discounted value of expected future cash flows (asset bubbles). For example, a new generation of behavioral models has tried to offer new insights into how bubbles start, under what conditions they might burst, and why arbitrage forces may fail to ensure that prices reflect fundamentals at all times [Scherbina (2013); Carvalho et al. (2012)].

---

23 For an overview of suboptimal behaviour and its implications for the efficient market hypothesis, see Shiller (1999). According to Fama (1998) most of the above anomalies (such as long-term under-reactions or over-reactions to information) can reasonably be attributed to chance. Moreover, according to Fama, these studies rarely provide a test of specific alternatives to the market efficiency hypothesis.

24 While these developments are very interesting, they do limit the analysis to finance theory. A policymaker has to overcome the serious limitation that, in the real world, wide and prolonged deviations from equilibrium in the financial markets (driven by behavioural biases) have serious effects on the real economy (through a number of channels, from welfare effects to public finance) which, in turn, can trigger second round effects on those markets.
The policymaker should help individuals by reducing at least some forms of behavioral bias. This can be done, for instance, by asking the financial providers to reorganize the context in which people make decisions (or choice architecture) so that they overcome their cognitive biases and are in a position to enhance their self-interest. According to this approach (called “soft libertarian paternalism”), the framework of the choice is such that cognitively constrained individuals are prompted to make better choices. This type of paternalism is called “soft” because individuals remain free to opt out of the choices deemed optimal by the benevolent designer of the policy framework [Thaler and Sunstein (2003)] (see Section 5, Box 5 for a discussion of a strong libertarian approach). The policy suggestion (called nudging) is to reorganize the context in which people make decisions so that their behavior is altered in a predictable way without ruling out any options. According to some authors, however, the limit of this approach is that nudging is not capable of sustaining long-term behavioral changes that are rooted in deep individual attitudes and values.25

Policy implication 7: Transparency, disclosure of information and consumer education are the typical policy responses to the above market failures on the consumption side. In order to minimize potential reputational risks and protect consumers, banks offering complex financial products need to have a good understanding of the main problems generated by insufficient awareness of the risks borne by retail investors; they should therefore perform a client suitability assessment. The regulator should encourage the bank to reduce unnecessary over-complexity and to frame different options.

Market failures in trading innovative products: the wholesale sector: in principle, financial innovation should not produce side effects for other banks, the risk of which should be mitigated by the fact that banks are (or should be) much better informed and better equipped to understand complex financial products. However, the crisis shows that even supposedly well-informed and sophisticated banks can buy large amounts of complex mispriced products (such as CDOs-squared), generating idiosyncratic and systemic risks. This can be explained, for example, by over-reliance on inadequate financial models and/or on ratings, caused in turn by moral hazard, wrong incentive schemes and unaccountability.

25 See (albeit in a different context) Salazar (2011).
Box 3: On mispricing and its effects
As previously mentioned, structured products are a combination of elementary instruments from the spot and futures markets (e.g., stocks, interest rate products, derivatives) promising tailor-made risk/return profiles for investors. They offer the feature of facilitating complex positions in options without the need to access options exchanges. When structured products are traded, transaction costs (e.g., bid-ask spreads) and commissions for the private investor are usually lower than those for the corresponding single trades. In practice, however, the price can deviate from the “fair value” along the life of the security. Some authors show that the mispricing is generally in favor of the issuing institution, and differs among the issuers as well as among the types of instrument. In general, more complex products incorporate higher implicit premia. In the secondary market, however, the level of overpricing tends to decrease as the products approach maturity [Stoimenov and Wilkens (2005)].

Structured securities or transactions are often bespoke products and their valuation depends on proprietary financial models and the inputs that drive them. Frequently, such inputs are not directly observable in the market. Even a valid model with accurate inputs will not always capture the immediate supply and demand profile of the market, meaning that the model price will not always determine the price at which a transaction occurs. In these circumstances, buyers and sellers may achieve price discovery only through actual transactions, but these, in stressful market conditions, may not occur because of the illiquidity of products. During the crisis, many high-risk complex financial instruments presented significant challenges for risk monitoring and management systems, which struggled to keep up with the complexities of product design and development and, in particular, encompass the risk that hedging strategies were ineffective, so generating additional, and sizeable, exposure in the form of basis risk. Once the market realizes that a class of asset is substantially mispriced, then the adjustment assumes the form of a sudden and wide correction. This can encourage behavior which, when taken across an industry as a whole, can prove highly pro-cyclical. This is particularly the case given industry participants’ tendency to mirror each other's trading strategies, and their requirement to unwind positions on a simultaneous basis during periods of market stress [CRMPG III (2008)].
Complex securitization products: before the crisis, several banks in some jurisdictions produced large volumes of complex securitization products through sponsored vehicles resulting in unexpected liquidity and credit risks for the sponsor. Large amounts of mispriced and misrated securities were sold to other banks or to consumers (through mutual funds). Structured products were re-used as collateral in the repo market to secure funding in order to set up highly leveraged positions. All these factors were a source of systemic instability among highly interconnected financial intermediaries both in the banking sector and in the shadow banking system.

The mispricing of structured products is not only detrimental to investors, it can also be a source of reputational and legal risks for banks. Mispricing is normally calculated as the deviation between the theoretical value of the security (sum of the value of the security plus the value of the derivative component) and its issuing value. This deviation can be the result of explicit fees (covering mainly production costs) plus a markup (implicit fees). Lack of competition or informational asymmetries between consumers and producers can be the main source of extra profits and implicit fees.

Policy implication 8: In principle, well-informed and sophisticated intermediaries should buy, and be able to manage, complex innovative financial products.

---

26 Rehypothecation (or re-use of a client's collateral) can in some circumstances be a source of risk. During the time the account provider exercises its rehypothecation right, the client's ownership right is replaced with a contractual right to the return of equivalent securities; this contractual right is not usually protected (MIFID, for instance, protects only the client's ownership rights). This works well until bankruptcy occurs. If the account provider defaults, clients with a mere contractual claim become unsecured creditors, meaning that their assets are, as a rule, tied in the insolvency estate and they are obliged to line up with all the other unsecured creditors to receive them back. This problem shows how important it is to have a comprehensive approach where both financial and legal aspects (and their evolution) are taken into consideration.

27 Rules on securitization, complemented by a wide regulatory perimeter, have allowed some jurisdictions to contain the harmful effects of financial innovation and limit the request for public support that has been particularly high in some countries (European Commission (2011)). For a recent short description of the crisis see Liikanen (2012).
Since, in reality, this was not the case – as emerged during the recent financial crisis – regulators should set up an adequate quantitative monitoring framework (eventually through centralized data repositories) to intercept anomalies or concentration of risks generated by complex innovative financial products [see Gola and Ilari (2013) on monitoring structured products], and use such information for developing early warning indicators.

**Table 1**

<table>
<thead>
<tr>
<th>Product innovation</th>
<th>Bounded rationality</th>
<th>Computational limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex financial products</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-retail sector</td>
<td>Asymmetric information (between producer and consumer)</td>
<td>Unfair practices (misselling)</td>
</tr>
<tr>
<td>Possible systemic effects (negative externalities) on other intermediaries or on the economy at large</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generation of idiosyncratic risks for a single intermediary: inadequate risk assessment or risk management (also through the acquisition of an innovative product offered by other intermediaries)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Process innovation: | | |
|--------------------|--------------|
| Type of problems | Poor risk analysis or mispricing of new products |
| Model selection | Mispricing, use of simplistic assumptions |
| Inadequate analysis of tail risks (counterparty risk, liquidity risk, etc. and their dependence structure and dynamics) | |
| Insufficient dataset, short-time series | |
| Model application | Miscalibration |
| Poor data quality (outliers) | |
| Algorithm/routines | Generation of errors |
| Legal design | Legal costs and reputational risks |
| i) Possible causes | |
| Misbehavior | Wrong incentives: poor internal governance |
| Overreliance on mathematical models (risk vs. uncertainty, complexity) | |
| Excessive risk taking/moral hazard/unaccountability | |
| Fraud | |

Market infrastructures (e.g., use of CCP to mitigate counterparty risk, reporting OTC derivative contracts to centralized trade repositories) close cooperation with financial market regulators.
Potential suboptimal behavior relating to process innovation: The financial crisis offers clear evidence that widely used market models have failed to properly consider correlations between extreme (rare) events. Market participants should be aware of the limits of models or product design, and not be overconfident in view of the following potential problems (Table 1):

i) Inaccurate selection of the financial model: with respect to the tasks or objectives, appropriateness of model selection is critical. If the model is not well-designed (because it is based on simplistic assumptions, is unable to deal with tail risks, volatility clustering, etc.), it can produce wrong valuations and mispricing. Complexity, excessive leverage, illiquidity and interconnectedness can exacerbate the negative effects of mispricing.

ii) Misuse of the selected model: even the most perfect financial model can produce undesirable results if not adequately used. Insufficient datasets or time series (small-sample problem\(^{28}\)), poor data quality, missing data or outliers, and miscalibration are some of the most frequent problems.

iii) Imperfect algorithms: a different set of problems relates to the generation of errors or malfunctions due to imperfect numerical routines (e.g., flash-trading programs).

iv) Imperfect legal design: this can produce significant legal or reputational risks and even systemic risks (for instance, ill-designed contractual triggers or covenants).

v) Non-robust market infrastructure: good market functioning is essential for preserving market confidence. More transparency and liquidity, which enhance market price discovery, improve the allocation of financial resources. A robust and well-regulated market infrastructure (supported by central counterparty clearing houses (CCPs)) helps to contain pro-cyclicality, cliff effects, and other forms of market misbehavior.

\(^{28}\) The use of optimal sample size in calibrating pricing or risk models is not trivial. Generally, large samples result in reduced sampling error; however, if market characteristics are changing, a small sample emphasizing recent data or (even better) the use of adaptive state-state models or GARCH models may be preferable. They are, however, computationally more demanding. See Sheedy (2009).
Overreliance on simple models: Complacency and overreliance on financial models have historically been related to the misbehavior of agents involved in sophisticated financial products, in some cases without top managers or boards being fully aware. This is of great importance if we consider that, according to some authors, the crystallization of tail risks may threaten the financial stability of even highly capitalized intermediaries [Ratnovski et al. (2011)].

Policy implication 9: It is reasonable to think that in a competitive and not excessively concentrated market, inefficient agents are progressively forced to reduce their role or even to leave the market. However, persistent use of wrong models and model misuse emerged as a standard at the outset of the financial crisis. Regulators should focus their attention on incentive misalignments, unaccountability, moral hazard and insufficient internal controls that may, through financial innovation, induce excessive risk taking. Sound internal governance, and in particular the “new product approval policy” (see Section 5), could reduce the appearance of new types of risk. Regulators should expect an intermediary to analyze all risk factors and consider all possible scenarios when designing a new product as well as to assess model assumptions. The assessment should be included in the information communicated formally to the board and senior management. Onsite supervisory visits should verify that the process is properly enforced.

4. Possible policy outcomes
From the overview of financial innovation provided above, it is possible to draw a set of general principles and policy recommendations.

A. General principles: A regulatory framework for financial innovation should respect the following four general principles:

i) Preserve the stimulus to innovate, maintain the positive aspects of financial innovation while minimizing unintended side-effects.

ii) Discourage unnecessary over-complexity of financial products in order to minimize market and behavioral failures. The regulation should be consistent with this principle.

iii) Promote regulatory consistency across institutions and instruments in order to minimize the risk of regulatory arbitrage.
iv) Maintain a critical, open-minded and holistic approach in order avoid being captured by conventional wisdom [FSA (2009)].

5. Policy tools
A strong and effective legal and supervisory system, designed to protect economic agents from unfair or illegal practices (fraud, abuses, etc.) and market misconducts is essential. This, in the long run, will pave the way for the sound and sustainable growth of banks. Competent authorities can use a broad range of regulatory and supervisory tools to deal with the above problems. The subject of this section, however, is not individual rights, but rather the definition of policy tools to deal with economy-wide market failures, generating inefficiencies and social welfare costs.

The following proposed taxonomy, consistent with the above overarching principles, is based on seven policy tools (PTs) able to capture the nine policy implications listed above. They are broadly consistent with the E.U. regulation establishing the EBA (Regulation (EC) No. 1093/2010, particularly Article 9), as well as the EBA Guidelines on Internal Governance and the EBA (2014) proposed Guidelines on Product Oversight and Governance. They are inspired by the G20/OECD high-level principles on financial consumer protection, by the FSB recommendations on good supervision and by the BCBS guidelines on corporate governance, adapted to take account of the specific characteristics of the financial innovation, which, as we have seen, also have several implications from a prudential point of view [OECD (2011); FSB (2012); FSB (2011b); BCBS (2010)]. They take into consideration recent ESMA (2012a, 2012b) guidelines and the ESMA (2014) “Opinion” on related issues as well. The taxonomy is ordered by the intensity of potential market failures, namely: asymmetric information, behavioral biases, principal-agent problems, non-systemic risks or limited negative externalities, systemic risks and serious market threats. Some of these policy responses are, to some extent, intrusive, possibly generating unintended side-effects, such as false alarms, altered preferences or risk-profiles and, more in general, inefficiencies. The policy intervention should, therefore, be calibrated on the basis of an adequate and sensible cost-benefit analysis. In the case of financial innovation, this is a particularly difficult exercise, as it is very hard to evaluate the long-term net benefits of policy intervention.
Transparency and disclosure (PT 1): this set of policies aims to reduce information asymmetries between consumers and banks. Greater disclosure and transparency is the precondition for any rational consumer to take responsible and well-informed decisions. Transparency and disclosure also help to enhance market discipline and reduce possible misalignments between fair and actual asset valuations (i.e., the level of mispricing), improving both asset allocation efficiency and consumer welfare. Standardized pre-contractual forms and comprehensive cost indicators should make information also accessible to less-educated consumers and allow a comparison between similar products. Specific disclosure mechanisms should be developed to provide information commensurate with the complexity and riskiness of the product or service. The provision of advice should be objective and based on the consumer’s profile, considering the complexity of the product, the associated risks and the customer’s financial objectives, knowledge, capabilities and experience. If the product is particularly complex and innovative, the financial agent or advisor should have an adequate and deep understanding of its features and potential risks.

Awareness and financial education (PT 2): the competent authorities should take a variety of measures to enhance the financial literacy of consumers. The objective is to help them make responsible choices in line with their attitude to risk, liquidity constraints, financial needs and inter-temporal plans. It could be useful to promote the understanding of basic economic concepts, as well as the diffusion of simple financial tests (such as the certainty equivalent test), to help individuals reveal their own risk preferences and level of risk tolerance. Consumers should be aware that financial markets can be highly volatile and characterized by the crystallization of rare, high-impact events. Appropriate policy responses should be developed to help consumers enhance their knowledge, skills and confidence to understand the risks and opportunities of a given product or service. Consumers should make informed choices and take effective action to improve their own financial well-being. Financial service providers should nevertheless assess the related financial capabilities, situation and needs of individual customers before agreeing to provide them with a product, advice or service. The policy intervention should also help consumers reduce at least some forms of behavioral bias and improve their financial responsibility. This, in turn, produces positive feedback on the intermediaries, which are the main objective of our analysis.
Financial innovation oversight: a policy framework

Box 4: The new product committee (NPC)
The NPC is an internal governance framework designed to foster responsible innovation and assess the appropriateness of new financial products and activities [Armstrong et al. (2011)].

The NPC should be organized as follows: the NPC comprises permanent members (risk, compliance, legal, fiscal, IT, business conduct, remuneration and finance) and invited members (appropriate representatives appointed after discussion with the business sponsor). The chairman of the NPC, who should be fully independent, is responsible for the final decision (approval of the product or service). The business sponsor, in coordination with the risk manager, should prepare a document describing the features of the new product and identifying the resulting impact for the bank, including but not limited to the risk management, legal and reputational profiles. Based on this document, members of the NPC have to provide a written statement to the business sponsor justifying their position on the new activity or product under scrutiny. In order to get a new product approved, the sponsor first organizes informal meetings with the different functions in the bank to get their reactions. Initially, once the validation has been granted by the NPC, the approval is valid for a limited period (for instance, 12 months). A senior member of the “risk committee,” also a member of the NPC, is responsible for a preliminary assessment of the risks generated by the new product. If the new product has complex derivative components, the NPC should explicitly describe the robustness of the model and the underlying assumptions.

The NPC could be a useful tool to deal with several problems provided that 1) it is not considered a further add-on to the already overcharged internal governance structure of the banks; 2) its role is interpreted as an internal, critical, holistic assessment of potential undesirable risks – for the entity and for the contractual counterparts – of innovative products or activities; and 3) the assessment involves various areas of the bank. The final aim of the NPC framework is to enhance the staff’s and the board’s accountability and legal responsibilities.
Internal governance (PT 3): well-organized governance is essential for a safe and sound functioning of financial institutions and, if not implemented effectively, may adversely affect that institution's risk profile. Sound internal governance also helps to protect depositors, consumers and other clients; it can help to reduce moral hazard and principal-agent problems [BCBS (2010)]. The problems relating to mispricing, imprudent use of financial models, operational and legal flaws or fragilities, and inappropriate risk management (in particular, excessive leverage, liquidity and counterparty risks) associated with innovative financial products, require specific attention, above traditional standards. The two main internal governance functions dealing with financial innovation are “risk management” and “internal control.” The first requires an institution to have in place effective processes for identifying, measuring or assessing, monitoring, mitigating and reporting on risks. The second requires institutions to have an appropriate internal control framework to develop and maintain systems for ensuring effective and efficient operations, adequate control of risks, prudent conduct of business, reliability of financial and nonfinancial information reported or disclosed (internally and externally), as well as compliance with laws, regulations, supervisory requirements and the institution's internal policies and procedures [EBA (2011), Title III, Section 3 (33-34)].29

It is important for the internal governance and risk management functions to be performed at group or firm-wide level: a strong grip on foreign subsidiaries and branches, particularly when they are located in jurisdictions where the regulatory framework is less stringent, is critical. With specific attention to aspects related to financial innovation, we would like to underline that:

A. Senior management and directors should have a thorough understanding of innovative financial products and processes. The competent authorities should engage with board members, particularly independent non-executive directors and risk managers, to discuss and assess awareness of the characteristics of the most innovative or complex financial products. The board (especially the independent directors) should be selected on the basis of their understanding of financial innovation and related pricing models.

29 The approach suggested in this paper has been recommended by the European Supervisory Authorities [see Joint Committee of the European Supervisory Authorities (2013) and further elaborations in the recently published EBA's Guidelines under consultation (see EBA (2014))].
B. Distortion of managerial incentives or ill-designed remuneration schemes need appropriate policy responses. Compensation schemes should be consistent with prudent risk management and the company’s long-term results; the measurement of performance for bonus awards should include adjustments for all types of risk and the cost of capital and liquidity. We think that introducing “malus” or claw-back of past bonuses could be considered, for instance, when the characteristics, assumptions and implications (for the bank itself and for other contractual counterparties) of a new product or financial contract are not well-represented to the board and are a source of serious negative effects. The remuneration of staff operating through an external network (i.e., tied financial agents) shall not be paid on the basis of their performances. Well-designed incentives can help minimize the negative effects of excessive risk-taking and moral hazard, for instance, when traders generate risks through the production of innovative and complex financial instruments or contracts of which the board or the shareholders are not fully aware.

C. It is particularly important to review, at the earliest opportunity, “outsized profitability” and market share gains to ensure that this does not reflect a problem with the original pricing or risk assessment of the product. Similarly, it is advisable to introduce internal mechanisms to encourage early disclosure of problems related to innovative products.

D. Treating both retail and non-retail clients fairly and ensuring that they are not affected by firm’s conflict of interest should be an integral part of good governance.

E. It is important that the assumptions underlying financial models are clearly articulated and subject to frequent, comprehensive review. Alternative measures should be presented to the board to demonstrate the sensitivity of the calculated metrics to changes in underlying assumptions. Moreover, intermediaries should “think creatively” about how stress tests can be conducted, including the scenarios where a significant stress event is generated by contagion.

---

30 This aspect has also been widely considered in the context of the fund industry regulation (UCITS IV - European Commission Directive proposal (COM, 2012, 350 final). See also EU Recommendation (2009/384/EC).
31 See EBA (2011), Section 21 (5) p. 34.
32 The use of claw-backs has also been considered by the BCBS (2010), principle 11 (112).
F. The intermediary, after offering complex, innovative or highly risky products or services, should promote a post-transaction monitoring and follow-up, such as sending monthly evaluation reports or improving clients’ ongoing monitoring systems.

G. All particularly innovative financial institutions should have a “new product committee” (NPC) (see Figure 2 and Box 4). The NPC should regularly (at least once a year) produce a written statement containing, in addition to the above mentioned aspects of good corporate governance, the following information:

- A formalized description of the functions of the new class of products or type of transactions (as described in Section II). Unnecessary over-complexity of products or transactions should be avoided.
- A formalized quantitative representation of the risk profile of the new class of products, including stress testing.
- A synthetic formalized description of the rationale and assessment of the characteristics of all the classes of designated structured transactions. Designated structured transactions are transactions, series of transactions or products where (i) one of the client’s principal objectives appears to be to achieve a particular legal, regulatory, tax or accounting treatment, including transferring assets off balance sheet; (ii) the proposed legal, regulatory, tax or accounting treatment is materially uncertain; (iii) the product or transaction (or series of transactions) has substantially offsetting legs or lacks economic substance; or (iv) the product or transaction (or series of transactions) could be defined as financing, but is structured in another manner.
- A synthetic and formalized description of the characteristics of all classes of strategic transactions. These include all transactions that are sufficiently large and important to the client or sufficiently large in the context of the market to warrant heightened scrutiny. They often have several of the following features: (i) losses or gains from the transaction could reasonably be expected to impact materially on the client’s financial position or adversely on

---

33 This and the following paragraph are based on Goldman Sachs (2011).
34 “Offsetting legs” refers to cash flows under different parts of a transaction (or group of related transactions), which from an economic perspective cancel each other out. Transactions with offsetting legs may lack economic substance.
the firm's reputation; (ii) the transaction is likely to have a material impact on the market; (iii) the transaction requires the approval of the client's CFO, CEO or Board; (iv) the transaction hedges a material acquisition, disposition or other business combination transaction by the client, and the hedge is material; (v) the transaction requires separate disclosure in the client’s financial statements or will otherwise be disclosed through a public filing; or (vi) the transaction represents a large financing commitment by the client. Strategic transactions may not involve complexity or unique structural features, but they nevertheless merit heightened review because of the above factors.

A statement regarding the suitability assessment\textsuperscript{35}: the NPC should assess the suitability of the product or transaction for the following classes of clients: i) retail; ii) professional investors (e.g., banks, broker-dealers, investment advisers and hedge funds); iii) other institutional clients (e.g., municipalities, sovereigns, sub-sovereigns, pension funds, corporations, charities, foundations and endowments); and iv) high net worth accounts (e.g., natural persons, family businesses). The suitability assessment has recently been recommended by IOSCO (2013a), which advises that intermediaries be required to develop and maintain internal processes and procedures with a view to ensuring compliance with the suitability requirements,\textsuperscript{36} the management of conflicts of interests, the proper conduct of business and the fair treatment of all customers, including in the distribution of complex financial products.\textsuperscript{37} According to IOSCO, effective internal reporting and communication of information should be established at all relevant levels.

\textsuperscript{35} See European Commission, 17179/12, Proposal for a Directive on markets in financial instruments repealing Directive 2004/39/EC.

\textsuperscript{36} According to IOSCO's (2013a) suitability requirements, an intermediary should “assess whether the product being sold matches the customer's financial situation and needs.” [This] “may include an assessment of the customer's investment knowledge, experience, investment objectives, risk tolerance (including risk of loss of capital), time horizon and/or capacity to make regular contributions and meet extra collateral requirements, and understanding of the product in question where appropriate” [IOSCO (2013a)]. See also Article 19 of Directive 2004/39/EC (MiFID).

\textsuperscript{37} According to IOSCO (2013b), complex financial products are “financial products, whose terms, features and risks are not reasonably likely to be understood by a retail customer (as that term is defined in individual jurisdictions) because of their complex structure (as opposed to more traditional or plain vanilla investment instruments), and which are also difficult to value (i.e. their valuations require specific skills and/or systems, particularly when there is a very limited or no secondary market).” [see IOSCO (2013b)].
More specifically, “intermediaries should put in place and enforce written strategies, processes and controls in view to ensure that any financial products they intend to distribute, especially complex financial products, are suitable for the type of customers they intend to solicit. When developing or selecting complex financial products for distribution, intermediaries should establish appropriate ongoing internal processes for identifying, periodically reviewing and approving (or rejecting) the products in order to promote their compatibility with the characteristics and needs of the prospective customers they intend to target.”

Box 5: A strong paternalistic approach
Product restrictions or bans designed to protect consumers could be considered in cases where specific products can generate serious losses to a large number of consumers or investors. As discussed above, people do not always make rational choices; individuals can make poor decisions, especially if they have behavioral biases or limited capabilities with respect to the complexity of the problem. In such circumstances, a policy option could rely on the strong paternalism principle, according to which a policymaker intervenes in the choices of individuals (with no opt-out) for their own good, helping them make better decisions, free of behavioral bias. However, proper calibration is needed to avoid a situation in which the policymaker interferes in individuals’ decisions in a coercive or restrictive manner, reducing their fundamental choices and freedoms. Such intervention should be considered only in cases of clear behavioral bias or excessive risk-taking. While this policy option almost certainly creates a negative effect on suppliers of financial products (at least in the short run), it should produce a net social welfare gain. Such developments, however, would open the door to a number of very complex issues beyond standard utilitarianism. The following observation by Mirrlees (1982) is interesting in this regard “[i]t must be legitimate, in principle, to advance arguments in favor of modifying the utility function that exactly represents my existing tastes. It cannot be wrong in principle to try to get someone to do what would be better for him even though he does not recognize it, but there must be some basis for saying that, with full understanding, he would come to accept the rightness of the altered utility function, or, rather, the underlying preferences.”

Financial innovation oversight: a policy framework

**Figure 2: Financial innovation: internal governance**

The board of directors: establishes and oversees the new product policy embedded in the "risk appetite framework" (RAF) and related ethical standards, approves the NPC written statement, holds the NPC’s permanent members accountable for the integrity and enforcement of the WSNA’s, obtains an independent assessment by the internal audit on the process behind the NPC.

**New product committee (NPC)**
- Permanent senior members from the following committees: risk, legal, fiscal, compliance and business conduct, remuneration and finance.
- Invited members from the business units.
- Functions: the NPC is accountable for ensuring the implementation of the RAF, produces the WSNA and approves the suitability and appropriateness of the new products or contracts and the ethical standards.

**Risk appetite framework (RAF):** The RAF should establish a process for communicating the policy stance on risks within the firm and to external stakeholders. The RAF should be linked to the firm’s long-run strategies, capital financial plans and compensation schemes. The RAF should be adaptable to changing and potentially adverse market conditions [FSB (2013)].

**The written statement of new activities (WSNA),** in addition to the standard risk and conduct assessments, for each innovative class of products, activities, processes, series of transactions or contracts (including bespoke financial contracts) should include 1) a simple description of the financial product and process, including the economic rationale or functions (or bundle of functions); 2) a representation of the managerial incentives (bonus/malus); 3) a description of internal early warning indicators; 4) an assessment of a fair treatment of clients and absence of conflict of interests; 5) a description of the post-transaction monitoring process; 6) clear information on model assumptions; 7) probabilistic scenarios; 8) designated structured transactions; 9) strategic transactions; as well as 10) suitability; 11) the ethical statements; and, finally, in case of structured products 12) the price decomposition or unbundling [Gola and Ilari (2013)].
An ethical statement: it is widely recognized that reasonable ethical standards are part of good corporate governance. The board should base its ethical standards on principles shared with other stakeholders. A critical, open-minded attitude is important, as conventional wisdom is not necessarily correct and is potentially affected by ignorance, confusion and prejudice.\(^\text{39}\) Note that, while it is irrational or improper to ignore well-known facts, rational decisions can be taken in situations of uncertainty or partial ignorance (see below). Institutions can adopt different ethical rules, provided that they are simple, transparent and designed in a manner that is well-specified and verifiable ex-post. Ideally, in order to be universally accepted, the company should maximize the satisfaction of preferences of all fully informed and aware stakeholders, considered impartially (i.e., giving equal weight to all preferences). Some boards could decide to go further, where the long-run marginal consequences of the company’s actions for the society are taken into consideration as well. This suggested approach is very flexible: the board could decide to adopt a strict view (i.e., only the effects of the on specific financial markets are considered) or rather a broad view (where the perceivable implications for the real economy are included). While restricting the business possibilities for the company, this very strict policy stance could be beneficial in the long run, not least in terms of public reputation.

In practical terms, the NPC should translate the boards’ principles into specific ethical standards. It could decide, for instance, to forbid all speculative transactions unless for hedging a risky position of the firm. Similarly, the NPC could decide to discourage the production of an ETC that, for instance, increases the volatility of a primary good, causing very negative economic effects to some producers (for instance farmers in poor countries specialized in the production of that good). Note that this rule excludes relevant non-utilitarian information (such as motivations, rights or other non-utilitarian ethical values). This means, for instance, that it does not preclude the company selling a very risky product to a fully informed and aware retail customer.

In many circumstances the economic environment is characterized by relevant uncertainty and it is difficult for the NPC to define ex-ante the set of possible events and consequences

\(^{39}\) This section adapts some ideas developed by Hare (1989), particularly chapters I, VII and XIII.
Financial innovation oversight: a policy framework

and to apply respective probabilities. In these situations it would be problematic for the NPC to be accountable for the negative externalities generated by the supply of some financial products. The NPC could, however, reduce this problem by releasing a statement based on the following simple application of Harsanyi's criterion, which deals with ethical rules with uncertainty.

According to this criterion, if the NPC wants to decide between two alternatives, A and B, all the NPC's members have to do is to ask themselves the question: “Would I prefer to live in a society conforming to standard A or in a society conforming to standard B? - assuming I would not know in advance what my actual social position would be in either society but rather would have to assume to have an equal chance of ending up in any one of the possible positions” [Harsanyi (1977)]. This rule is fully rational, consistent with utilitarianism, and able to deal with an uncertain environment, where only an ex-ante intuition of possible consequences is available.

This framework roots its moral stance on the obligation to consider all available information and to critically assess the nature and implications of the financial innovation. This process needs to integrate two aspects; the first is based on a fully informed and rational analysis, while the second uses intuition in order to obtain a prudent evaluation of all risky factors. The intuition is particularly important in situations of uncertainty, tension or emergency [Hare (1989)], for instance due to the perception of potentially unstable market conditions and, more in general, when the NPC is dealing with complex situations.

The ultimate goal of the suggested policy framework is not to promote any specific ethical rule, but rather to force the firm and its bodies to disclose their own criterion and be socially accountable for it. However, should the market, on average, decide to take an ethical stance considered unsatisfactory by the regulators (which should be accountable and appointed through a transparent and democratic process), a more intrusive and paternalistic approach is needed (see Box 5 and policy tools 6 and 7).

Internal governance could be insufficient: even assuming that all the types of risks mentioned above are properly considered and managed, that all best internal governance procedures have been adopted and that the design of all incentive schemes perfectly aligns the interests of various stakeholders, there is still a possibility that the firm may be a source
of negative unintended systemic externalities for other entities or economic agents. Three circumstances are highly relevant: 1) the single intermediary is unaware of the effects of its behavior on the system; 2) the single intermediary is aware that it is generating a possible negative externality, but it considers that such action has negligible consequences; and finally 3) the intermediary behaves as other players in the market and is aware that the market is generating a negative externality even on its economic condition, but a coordination failure prevents self-correcting action. More policy interventions are therefore needed, and in particular:

**Warnings and guidelines (PT 4):** in some circumstances, specific banking and financial products can generate serious and undesirable risks for a significant portion of the household sector or other intermediaries, directly or indirectly affecting a certain part of the economic system. The competent authorities should consider a more incisive approach, releasing a general warning or enforceable guidelines regarding the diffusion of a specific financial product. The objective of such warnings or guidelines is to encourage financial prudence (on the part of both consumers and intermediaries). This policy option should be considered where a specific product is a source of serious threat\textsuperscript{40} for some market segments or the financial market as a whole. In the case of a false alarm or welfare loss for the producers, account should be taken of the potential reputational and legal risks to the regulator.

**Financial market standards (PT 5):** in some cases, the competent authorities should consider developing appropriate technical standards. The case of securitization offers a clear example: as we have already discussed, the market evolution of this instrument degenerated, in some jurisdictions, into the production of complex, opaque and mispriced products. Regulators and standard setters are now discussing several policy responses [IOSCO (2012a); BCBS-IOSCO (2014)]: i) better transparency by issuers about verification and risk assessment practices; ii) review of investor suitability requirements and development of tools to assist investors in understanding such instruments; and iii) better incentive schemes, based on the principle of more retention of risks by the issuers (called “skin-in-the-game”).

\textsuperscript{40} While preserving flexibility, regulators should provide a shared view of the meaning of “serious threat” (terminology used, for instance, by both EBA and ESMA).
Financial innovation oversight: a policy framework

This approach should align the compensation arrangements with long-run performance and asset quality. Where appropriate, the regulator should also develop standards to promote simple, comparable and uniform securitization products and, where appropriate, to support the market in developing a trading platform. This should enhance liquidity and price discovery.

**Restrictions or temporary prohibitions (PT 6):** In specific circumstances, the competent authority, after an in-depth empirical analysis, could consider whether to restrict or even (temporarily) prohibit a financial product. The main policy objective of the restriction or ban should be to prevent systemic risks and to protect the consumers or depositors. The theoretical justification for this form of “product intervention” is to preserve the system from economy-wide negative externalities. This rather coercive approach is particularly useful in circumstances where the potential serious threat consists in the sudden crystallization of risks, with a rapid and significant dislocation of assets, runs or wide price changes (cliff effects); the restriction on naked short selling is an example.

**Product pre-approval (PT 7) [FSA (2011)]:** A second type of product intervention is product pre-approval. In principle, the policymaker could contain the harmful effects of financial innovation by implementing a process of regulator pre-approval of all innovative financial products. On the basis of the discussions so far, we think that the regulator should be very cautious in deciding to be responsible for vetting products or assessing pricing models. There are various reasons for this: i) the evolutionary nature of the phenomenon implies that the regulator always lags behind market developments; ii) intervention could stifle market initiatives to promote innovative financial products; and iii) enforcement could be very difficult and costly. More importantly, even assuming that all these problems are overcome, two critical aspects remain unresolved: the policy option is not compatible with the incentives (generating moral hazard by both producers and consumers); the reputational and legal risks for the regulator could be very high. For these reasons, while not ruling out the possibility of introducing pre-approval in specific cases, we are more in favor of making extensive use of transparency, monitoring, on-site inspections and very stringent internal governance rules or other policy tools.
The introduction of a financial innovation authority: some authors have proposed the introduction of a specific authority in charge of approving each single financial product, building on the analogy of the American Food and Drug Administration (FDA) [Crotty and Epstein (2009)]. Under this proposal, the sponsoring financial institution would submit an application to market a new financial product: “It will be required to pay a significant fee that will fund the cost of testing. The sponsor will provide a safety and effectiveness statement, which will include comprehensive information concerning the nature of the product, the marketing plan (e.g., to whom it will be marketed, etc.), what the functions of the product are, and then evidence that the product will serve these functions. In addition, the sponsor will provide the results of safety tests based on its internal models, including the structure, inputs and assumption guiding these models. Unlike current practice, however, the models cannot be proprietary” [Crotty and Epstein (2009)]. Moreover, the sponsor would be required to provide the authority with full code information about the models, so that the latter could replicate the studies and understand their meaning. The sponsor would also be asked to provide information for a “financial stability impact statement.” Whereas, the “safety and effectiveness statement” would focus on the risk impact on the sponsoring institution and on the buyer of the financial product, the financial stability impact statement would focus on the impact of this product on the financial stability of the system as a whole. The proposal is based on the precautionary principle adopted by various international bodies (such as U.N. agencies), which states that “if an action or policy has a suspected risk of causing harm to the public or to the environment, in the absence of scientific consensus that the action or policy is harmful, the burden of proof that it is not harmful falls on those taking an act.”

6. Conclusion
The net social welfare effect of financial innovation can be either positive or negative depending on several factors. On one hand, a more innovative financial system is Pareto-superior in the sense that it helps to complete the market (for instance, by generating state-contingent claims for any possible future state of the world), reducing frictions and transaction costs and increasing liquidity.

41 See Crotty and Epstein (2009). Other authors have had the idea of setting up a similar authority which should, however, adopt the rule of approving only products that are socially beneficial, defined as financial instruments that help people insure against or hedge risks (in contrast with products used for gambling, considered socially detrimental). See Posner and Weyl (2012).
Financial innovation oversight: a policy framework

On the other hand, it can be a source of negative externalities, generating systemic risks, and suboptimal outcomes for some classes of economic agent (such as uninformed consumers).

Financial innovation involves consumer protection as well as systemic risk concerns. Regulation needs to be based on the identification of potential sources of threat for consumers and for the stability of markets; it is necessary to identify problems and implement solutions without curbing the stimulus to innovate. The regulator should have a strict and frank dialogue with market players, while maintaining a critical, independent and open-minded attitude.

If an innovative product has opaque or unclear returns and characteristics, transparency and disclosure could be a sufficient policy solution. If it is expected to generate not fully understood risks and consumers are affected by major behavioral biases, financial education policies should provide further support. More effective policy options (such as internal governance) can prevent the undesirable effects of certain types of financial and banking products, particularly if they are of a complex and risky nature, without being excessively intrusive. Warnings, restrictions and, in extreme cases, prohibitions are policy options that need to be carefully considered in terms of social costs and benefits, including the risk of generating regulatory arbitrages and reputational or legal costs for the regulator. The prohibition of financial innovation is a strong policy option and could create inefficiencies for the financial system.

Financial engineering, and in particular financial modeling of derivative instruments, is moving toward a more comprehensive and holistic approach, where liquidity, credit and collateral risks, as well as funding costs, are considered. This is a formidable task, and several methodological problems are not fully solved yet. From a theoretical standpoint, further progress is needed not only to find proper statistical models able to capture discontinuities, dependences structures beyond the multivariate normal distribution, and volatility clustering, but also to have a better understanding of the underlying economic factors. In the meantime, industry and regulators should work to substantially reduce risks and complexities with more robust market infrastructures and better regulation.
Financial innovation oversight: a policy framework

References
BIS, 1986, “Recent innovation in international banking,” Bank for International Settlements, April, Basel
BCBS, 2010, Principles for enhancing corporate governance, Bank for International Settlements, October
BIS, ECB and IMF, 2009, Handbook on securities statistics, second draft, February
Boatright, J. R., 2010, Ethics in finance, New Jersey: John Wiley and Sons
Brigo, D., M. Morini and A. Pallavicini, 2013, Counterparty credit risk, collateral and funding: with pricing cases for all asset classes, Wiley & Sons
Carver, L., 2012, “No more heroes in quantitative finance?” Risk, August
EBA, 2011, EBA Guidelines on Internal Governance (GL 44), London, September
EBA, 2014, Draft Guidelines on product oversight and governance arrangements of banking products, Consultation Paper, November
ESMA, 2012a, Guidelines on certain aspects of MiFID compliance function requirements, final report, ESMA/2012/388, July
ESMA, 2012b, Guidelines on certain aspects of MiFID suitability requirements, final report, ESMA/2012/387, July
ESMA, 2014, MiFID practices for firms selling complex products, opinion, ESMA/2014/146, February
Financial innovation oversight: a policy framework

FCA, 2013, “Applying behavioural economics at the Financial Conduct Authority,” Occasional Paper, no. 1, April
FSA, 2008, “Consumer responsibility,” discussion paper no. 8, December
FSA, 2011, “Product intervention,” discussion paper no. 11, January
FSB, 2011a, Shadow banking: strengthening oversight and regulation, October
FSB, 2011b, Consumer finance protection with particular focus on credit, October
FSB, 2012, “Increasing the intensity and effectiveness of SIFI,” Supervision Progress Report to the G20 Ministers and Governors, November
FSB, 2013, “Principles for effective risk appetite frameworks,” July
Hare, R. M., 1989, Essays in ethical theory, Oxford University Press, Oxford
IOSCO, 2012b, “The credit default swap market,” FR05/12, June
IOSCO, 2013a, “Suitability requirements with respect to the distribution of complex financial products,” FR01/13, January
IOSCO, 2013b, “Regulation of retail structured products. Consultation report,” CR05/13, April
Joint Committee of the European Supervisory Authorities, 2013, “Joint position of the European Supervisory Authorities on manufacturers’ product oversight & governance process,” JC 2013/77, EBA, EASMA, EIOPA, November
Financial innovation oversight: a policy framework

Kimball, M. S., 1990, “Precautionary saving in the small and in the large,” Econometrica 58, 53-73
OECD – Bank of Italy, 2010, Symposium on financial literacy, Rome, 9 June
OECD, 2011, “G20 high-level principles on financial consumer protection,” October
Piazza, R., 2010, Financial innovation and risk, the role of information, No. 10-266, International Monetary Fund
Financial innovation oversight: a policy framework

Sen, A., 1991, Money and value: on the ethics and economics of finance, Banca d'Italia, Paolo Baffi lectures on money and finance, Edizioni dell'Elefante
Sheedy, E., 2009, “Can risk modeling work,” MAFC research papers 35, June, Macquarie University
Turner, A., 2012, “Securitization, shadow banking and the value of financial innovation,” The Rostov Lecture on International Affairs, SAIS, Johns Hopkins University, April
Walmsley, J., 1988, The new financial instruments: an investor’s guide, John Wiley and Sons
The governance structure of shadow banking: rethinking assumptions about limited liability*

Steven L. Schwarcz
Stanley A. Star Professor of Law & Business, Duke University School of Law, Founding Director, Duke Global Capital Markets Center, and Senior Fellow, The Centre for International Governance Innovation

Abstract
In an earlier article, I argued that shadow banking – the provision of financial services and products outside of the traditional banking system, and thus without the need for bank intermediation between capital markets and the users of funds – is so radically transforming finance that regulatory scholars need to rethink their basic assumptions. This article attempts to rethink the corporate governance assumption that owners of firms should always have their liability limited to the capital they have invested. In the relatively small and decentralized firms that dominate shadow banking, equity investors tend to be active managers. Limited liability gives these investor-managers strong incentives to take risks that could generate outsize personal profits, even if that greatly increases systemic risk. For shadow banking firms subject to this conflict, limited liability should be redesigned to better align investor and societal interests.

* The original version of this article was first published at 90 Notre Dame Law Review 1 (2014). For helpful comments, I thank Brad Bernthal, Bill Brown, Anupam Chander, Peter Conti-Brown, Andromachi Georgosouli, Erik Gerding and Robert Hillman; participants in faculty workshops at University of California Davis School of Law, University of California Hastings College of Law, and University of Colorado School of Law; and participants at a public lecture at Queen Mary, University of London. I also thank Seth Bloomfield and Andres Rodriguez for valuable research assistance.
In a prior article, I argued that shadow banking is so radically transforming finance that regulatory scholars need to rethink their basic assumptions.¹ In this article, I argue that the governance structure of shadow banking should be redesigned to make certain investors financially responsible, by reason of their ownership interests, for their firm’s liabilities beyond the capital they have invested.² This argument challenges the longstanding assumption of the optimality of limited liability.

Shadow banking is a loose term that refers to the decentralized provision of financing outside of traditional banking channels, and thus without the need for traditional modes of bank intermediation between capital markets and the users of funds.³ The shadow banking system is immense, recently estimated at U.S.$67 trillion worldwide.⁴ Numerous types of firms make up the shadow banking system. They include special purpose entities (SPEs) used in securitization and structured finance transactions to raise financing indirectly through the capital markets,⁵ as well as finance companies, hedge funds, money-market mutual funds, nonbank government-sponsored enterprises, securities lenders and investment banks.⁶

² Ibid. See, e.g., Black’s Law Dictionary 426 (9th ed. 2009) (defining limited liability as the “liability of a company’s owners for nothing more than the capital they have invested in the business”).
⁵ See Schwarcz, supra note 3, at 621. These SPEs include asset-backed commercial paper (ABCP), conduits, and structured investment vehicles (commonly known as “SIVs”). Ibid.
⁶ Banks themselves can be regarded as shadow banking firms to the extent they facilitate customer financing in an agency capacity, as opposed to making loans or otherwise extending credit to their customers. For example, banks often create SPEs, ABCP conduits and SIVs. Banks also are important players in repo markets. Ibid. at 621-22.
Limited liability, this article contends, is not always optimal for firms that make up the shadow banking system (hereinafter “shadow banking firms”). Limited liability can sometimes make the governance structure of these firms uniquely subject to a market failure that externalizes the systemic costs of taking a risky action. To repair this failure, managers of shadow banking firms should sometimes be required to put more “skin in the game,” in order to better align incentives\(^7\) between their firms and society.\(^8\)

This article proceeds as follows. In part I, the article sets forth the background for its analysis, including a short history of how corporate limited liability became the norm and an overview of the general academic debate on whether it should be the norm. In part II, the article analyzes whether limited liability should be the norm for the governance structure of shadow banking. The article starts that analysis by proposing a normative framework for rethinking the corporate governance assumptions about limited liability. It then applies that framework to shadow banking. Although limited liability has always been a potential source of externalities, the article finds that it is a uniquely fertile source of systemic externalities for shadow banking firms, and that current law does not — nor are adaptations to traditional legal remedies likely to — adequately internalize those externalities. To mitigate those externalities, limited liability should be redesigned for shadow banking firms that are governed by conflicted investor-managers. Part III of the article explores how limited liability could be redesigned, establishing goals for the redesign and testing redesign proposals against those goals. Annex 1 to the article provides practical guidance for assessing the costs and benefits of any particular redesign.

This article does not directly engage whether — or the extent to which — shadow banking should be subject to substantive capital and solvency regulation, or even prohibited. As to the first, it has not historically been customary, at least in the U.S., to engage in solvency

\(^7\) Cf. White, E. N., 2011, “Rethinking the regulation of banking: choices or incentives?” 2 (unpublished manuscript) (on file with author) (arguing that bank regulatory “reform should focus on changing the incentives that parties face to insure that they are correctly aligned to induce the development of less fragile institutions”).

\(^8\) See Schwarzc, supra note 3. This article does not argue that such managers necessarily have conflicts with their firms. I have separately argued, however, that certain managers, especially secondary or middle managers, may well have such conflicts. See Schwarzc, S. L., 2009, “Conflicts and financial collapse: the problem of secondary-management agency costs,” 26 Yale Journal on Regulation, 457, 458-59.
regulation of firms that are not traditional banks. The Dodd–Frank Act is beginning to change that, with its regulation of certain systemically important financial institutions. Nonetheless, even though the limited-liability rule of corporation law can cause externalities, the government does not generally “take a more active role in assuring the solvency of corporations.” In part, this is because government micromanagement of the private sector is not always efficient. Likewise, it might well be unwise to attempt to prohibit shadow banking. Even if that were feasible, shadow banking “has the potential to create both benefit and harm. Empirically, we do not yet know which effect is likely to dominate.” Therefore, “financial regulation of shadow banking should ... strive to examine ... how to mitigate the potential harm while preserving the potential benefit.” is this article’s goal.

I. Background
As background for this article’s analysis of limited liability in the context of shadow banking, the discussion begins by examining, from a historical standpoint, how corporate limited liability became the general norm and then, from a scholarly standpoint, whether it should be the general norm.

A. History of limited liability
Limited liability has been called “a distinguishing feature of corporate law – perhaps the distinguishing feature” of corporate law. Yet early corporations did not have limited liability. Because their histories are different, first consider the evolution of limited liability for shareholders of nonbank corporations, then consider that evolution for shareholders of banks. Finally, compare these with the evolution of limited liability outside the U.S.

10 See infra note 124 and accompanying text.
11 Posner, supra note 9, at 413.
12 Posner, supra note 9, at 413-14 (describing the “continuous regulatory scrutiny” of an agency as a “statist” solution with only “arguab[ly]” positive benefits).
14 Ibid.
1. Nonbank corporations

During the early 19th century, for example, unlimited shareholder liability was the norm.\textsuperscript{16} The rationale for such unlimited liability was that creditors assured of repayment from shareholders would lend the corporation additional capital.\textsuperscript{17} Indirect shareholder liability, which resulted from the corporate power to make assessments, was also a common feature of early 19th century corporations.\textsuperscript{18}

Legislators initially were willing to allow corporations performing public functions, such as operating turnpikes, toll bridges and canals, to organize under corporate charters with limited shareholder liability.\textsuperscript{19} But they were unwilling to permit limited liability for shareholders of manufacturing corporations.\textsuperscript{20} The movement toward general limited liability started in the courts, when judges had to determine “whether shareholders were directly liable for corporate debts if the [corporate] charter was silent on shareholder liability.”\textsuperscript{21}

\begin{footnotesize}

\textsuperscript{17} Mendelson, supra note 16, at 1210.


\textsuperscript{19} Not only were legislators initially reluctant to grant corporate charters with limited liability to manufacturing corporations, they were generally reluctant to grant any corporate charters to manufacturing corporations until 1809. Blumberg, supra note 16, at 590–91; Dodd, supra note 18, at 1352–55.

\textsuperscript{20} Blumberg, supra note 16, at 591.

\textsuperscript{21} This classification is partially consistent with Tufano (2003), Lerner and Tufano (2011), Merton (1995) and Allen and Gale (1994).
\end{footnotesize}
By the mid-19th century, most courts “presum[ed] limited shareholder liability in the absence of any legislative rule.”22 Different courts had different rationales. Some courts, for example, reasoned that because some corporate charters contained express statements imposing direct liability, the absence of those statements in a charter implied intent not to impose such liability.23 Other courts wanted to avoid an injustice to shareholders who were both innocent and ignorant of a corporation's mismanagement.24 Federal courts often relied on a trust-fund theory: shareholders are merely residuary owners, reimbursed only after the corporation pays its debts;25 thus a corporation is like a trust fund in which the capital stock is used for the payment of corporate debts while stockholders are liable only for the amount of capital stock they contribute.26

Around the same time, legislatures began allowing limited liability, even for shareholders of manufacturing corporations.27 This change appears to have been more driven by pragmatic factors.

---

22 Mendelson, supra note 16, at 1210 n.17 (quoting Morton J. Horowitz, M. J., 1992, The transformation of American law, 1870-1960 94. Massachusetts courts as early as 1808 and Pennsylvania courts by 1816, for example, held that shareholders, absent an express provision in the charter or in a statute, were not directly liable for corporate debts. Blumberg supra note 16, at 592. Federal courts, by 1824, similarly recognized limited liability for corporate shareholders as the default rule. Ibid.

23 Ibid.

24 Spear v. Grant, 16 Mass. (1 Tyng) 9, 14 (1819) (“[I]f [a stockholder] were equally liable to each holder of the notes (which he must be if liable at all; for if the facts agreed create a promise to one, they create a promise to all), then the most palpable injustice would take place. For a stockholder, wholly innocent and ignorant of the mismanagement, which has brought the bank into discredit, might be ruined by reason of owning a single share in the stock of the corporation.”).


26 Ibid. at 436–37.

It was in part responsive to the increasing political influence of industrialists, resulting from
the rapid growth of the manufacturing industry,\(^{28}\) and in part due to a “flight-of-capital
argument,” which suggests that states failing to legislate limited shareholder liability would
suffer a flight of corporate capital to other states that had limited liability.\(^{29}\)

Limited liability did not become universal in the U.S., however, until a century later. California,
for example, imposed pro rata shareholder liability for all corporate debts and obligations
until 1931.\(^{30}\) That pro rata liability did not appear to cause a flight of corporate capital to
other states or otherwise impede California’s economic growth.\(^{31}\)

2. Banks
Unlike nonbank corporations, in which shareholders initially were subject to unlimited liability,
bank shareholders initially were subject to (only) double liability – liability for corporate
obligations in an amount equal to the par value of their shares.\(^{32}\) States imposed double
liability either by express provisions in state bank charters or in their state constitutions.\(^{33}\)
Congress followed the example of the states and provided for double liability in the
National Banking Act of 1863.\(^{34}\)

\(^{28}\) Ibid. at 592-93.

\(^{29}\) Dodd, supra note 18, at 1367-68, 1369. There were other, lesser, factors. In Massachusetts, for example,
the governor feared that unlimited liability made shares of Massachusetts manufacturing corporations
nearly worthless, and therefore brought financial ruin to Massachusetts’s families. Ibid. at 1368. A
Massachusetts senator also argued that unlimited liability created a moral hazard problem by enabling
manufacturing corporations to obtain credit in an amount in excess of the value of their business assets.
Ibid. at 1370.

\(^{30}\) Blumberg, supra note 16, at 597.

\(^{31}\) Ibid. at 599; see also Hansmann, H., and R. Kraakman, 1991, “Toward unlimited shareholder liability for
corporate torts,” 100 Yale Law Journal 1879, 1924 (noting that California imposed unlimited pro rata
liability by statute on shareholders, “evidently without crippling industrial and commercial development”) (Hansmann & Kraakman).

Wake Forest Law Review, 31, 36 (Macey & Miller).

\(^{33}\) Ibid

\(^{34}\) Ibid. The National Banking Act of 1863 specifically provided that “each shareholder shall be liable to the
amount, at their par value, of the shares held by him in addition to the amount invested in such shares.”
National Bank Act of 1863, ch. 58, 12 Stat. 665, 668. The senator that proposed the provision explicitly
recognized that the provision “tracked the laws of ‘most of the States of the Union.’” Macey & Miller, supra
note 32, at 36 (citing Congressional Globe, 37th Congress, 3d Sess. 824 (1863)).
According to the senator that proposed the provision, its purpose was to provide additional protection to bank creditors and, in effect, to also prevent the bank from engaging in excessively risky operations.  

Double liability for bank shareholders quickly fell out of favor, however, following the Great Depression. Several factors contributed to its rapid fall from grace, including a perception of unfairness caused by liability of bank shareholders who did not contribute to management decisions and public questioning of its ability to reduce risk, given the widespread bank failures. In response, Congress ended double liability for shareholders of national banks and states ended double liability for shareholders of state-chartered banks. Today, “double liability for bank shareholders is a dead letter everywhere.” Limited liability is now the general default rule in the U.S. for shareholders of banks and nonbank corporations. As explained below, the U.S. is not alone in adopting limited liability.

---

35 Macey & Miller, supra note 32, at 36 (citing Congressional Globe, 38th Congress, 1st Session 1869 (1864)).
36 Ibid. at 37.
37 Ibid. One commentator remarked that double liability “disappeared in the wake of the Great Depression where shareholders, who were hit by the huge unanticipated losses, pleaded with federal and state legislators to move to single liability.” White, supra note 7, at 8. The public’s disgust with double liability during this time is exemplified by the popular press’s harsh criticism of the Supreme Court in 1944 for upholding as assessment of shareholders of a holding company for the liabilities of a failed subsidiary bank. Macey & Miller, supra note 32, at 37-38.
38 See Macey & Miller, supra note 32, at 38.
39 Ibid. at 38-39 (“In 1933, Congress repealed double liability for newly-issued national bank shares; and in 1935, it prospectively extinguished all double liability for national bank stock provided that a bank gave six months’ notice of termination.” (footnote omitted)); see Blumberg, supra note 16, at 601 (noting actions by Congress and state legislatures that ended double liability for shareholders of national and state banks).
40 Macey & Miller, supra note 32, at 39.
41 According to Section 6.22 of the Model Business Corporation Act, a shareholder is “not personally liable for the acts or debts of the corporation except that he may become personally liable by reason of his own acts or conduct.” Model Business Corporation Act § 6.22(b) (2002); see Mendelson, supra note 16, at 1211 n.25 (detailing the adoption of limited liability provisions in several states). Most states have adopted alternative corporate forms such as the Limited Liability Corporation (LLC) and the Limited Liability Partnership (LLP), which also provide equity owners with limited liability for corporate debts and obligations. See Mendelson, supra note 16, at 1209 n.14. For further discussion of the development of the LLC corporate form both in the United States and in foreign jurisdictions, see, e.g., Gevurtz, F. A., 2008, Business planning 64 (4th ed.); Gevurtz, F. A., 2011, “The Globalization of corporate law: the end of history or a never-ending story?” 86 Washington Law Review 475, 487; Lutter, M., 1998, “Limited liability companies and private companies,” in 13 International encyclopedia of comparative law, ch. 2, at 4-6 (Detlev Vagts ed.).
3. Limited liability’s development in foreign jurisdictions

Civil law countries in Europe adopted limited liability even earlier than the U.S.\(^{42}\). Canada\(^{43}\) and England\(^{44}\) adopted limited liability after the U.S.

The move toward limited liability in England is instructive. Joint stock associations, in which members were liable for association debts, were the dominant form of business organization during the early 19th century, because corporate charters were difficult and expensive to obtain.\(^{45}\) Members attempted in various ways to limit their liability, including by inserting the term “limited” in the association's name.\(^{46}\) At least partly in response, the Parliament passed a law facilitating general incorporation but imposing unlimited liability on shareholders, with rights of contribution from fellow shareholders.\(^{47}\) That law also prohibited the corporate charter from limiting liability.\(^{48}\)

Wealthy investors were outraged, being concerned “that creditors would first proceed against them in preference to less wealthy members.”\(^{49}\) They sparked a political debate, arguing that the law would inhibit investment.\(^{50}\) As a result, Parliament first enacted limited liability for investors in railways, then one of England’s most successful and growing industries.\(^{51}\) That soon led to a legislative broadening of limited liability, not only for shareholders of new corporations generally, but also for members of new and existing joint-stock associations.\(^{52}\) Shortly thereafter, limited liability was extended to shareholders of banks\(^{53}\) and insurance companies.\(^{54}\)

\(^{42}\) See Blumberg, supra note 16, at 596 (discussing France’s experience as early as 1807 with limited liability, and how limited liability then spread throughout Europe).

\(^{43}\) Ibid. at 594 (discussing Canada’s adoption of limited liability in 1850).

\(^{44}\) Ibid. at 584 (discussing England’s adoption of limited liability in 1855).

\(^{45}\) Ibid. at 581-82.

\(^{46}\) Ibid. at 582.

\(^{47}\) Ibid. at 583 (discussing the Joint Stock Companies Registration, Incorporation, and Regulation Act, 1844).

\(^{48}\) Ibid. (observing also that the law was silent as to whether contracts with third parties could limit shareholder liability).

\(^{49}\) Ibid.

\(^{50}\) Ibid.

\(^{51}\) Ibid. at 584.

\(^{52}\) Ibid. (discussing the Limited Liability Act, 1855 and the Joint Stock Companies Act, 1856).

\(^{53}\) Ibid. (discussing legislation in 1857-1858).

\(^{54}\) Ibid. (discussing legislation in 1862).
4. How history informs the limited liability debate
The historical trend toward limited liability can help to inform the debate. In the U.S. and England, the legislative trend was heavily influenced by lobbying. In the U.S., the legislative trend was also influenced (in a nonbank context) by fear of capital flight to other states.\(^{55}\)

The judicial progression toward limited liability has more of a “fairness” rationale: to protect innocent shareholders who are not in a “capacity to control” or influence management decisions.\(^{56}\) This rationale is also part of the justification for shifting from bank-shareholder double liability to limited liability: imposing double liability on bank shareholders who did not contribute to management decisions created a perception of unfairness.\(^{57}\)

The shift from bank-shareholder double liability to limited liability also reflected disappointment that double liability did not prevent the bank failures of the Great Depression. The magnitude of the Great Depression, however, does not make it an appropriate test; its impact overwhelmed even many prudent banks.\(^{58}\)

\(^{55}\) Subsequent experience suggests, however, that unlimited liability may not necessarily cause capital flight. See supra notes 30–31 and accompanying text (observing that California’s pro rata liability rule did not appear to cause a flight of corporate capital to other states); see also Blumberg, supra note 16, at 594 (“In fact, there is little sign that the different legal rules on shareholder liability adversely affected economic development in Massachusetts and Rhode Island during this period.”); Dodd, supra note 18, at 1378 (“There is some indication that the early adoption of the limited-liability principle by New Hampshire and Connecticut stimulated the development of the cotton-textile industry in those states. Yet both states continued to lag, as producers of cotton textiles, behind Massachusetts, which did not adopt the principle until 1830, and Rhode Island, which did not adopt it until seventeen years later.” (footnotes omitted)).

\(^{56}\) See supra note 24 and accompanying text.

\(^{57}\) See supra notes 37–38 and accompanying text. At least some bank shareholders were also bank insiders, however. Macey & Miller, supra note 32, at 56.

\(^{58}\) A positive result of the double liability regime was that it encouraged prudent bank practices including voluntary liquidation of failing banks. Macey & Miller, supra note 32, at 57. Notably, however, involuntary liquidations approached the number of voluntary liquidations and consolidations during the difficult years of 1929–33. Ibid. at 58.
The historical trend in England toward limited liability can likewise help to inform the debate. The turning point was opposition to unlimited liability by wealthy investors, who feared they would become the ultimate deep pockets for their firms’ liabilities.59 It is unclear, though, whether fully limiting liability was the only way to remove the fear, thereby ensuring investment. Without totally sacrificing the monitoring benefits of shareholder liability, that fear could have been mitigated by partially limiting shareholder liability, such as to double liability or pro rata liability.60

B. The general academic debate
Next consider the academic debate on whether limited liability should generally be the norm. To that end, the discussion below reviews scholarly challenges to limited liability made prior, and then in response, to the “global financial crisis” (GFC),61 and then compares those challenges with scholarly justifications for limited liability.

---

59 See supra note 49 and accompanying text.
60 Professor Gerding also argues that the historical trend in England toward limited liability followed a pattern – an expansion of limited liability and liberalization of corporate law rules, followed by a financial boom, then followed by a financial bust – that proceeded like clockwork every ten years from 1825 through the nineteenth century. Gerding, E. F., 2013, Law, bubbles, and financial regulation 74-77.
61 By “global financial crisis,” this article means the financial crisis of 2007-08.
1. Scholarly challenges prior to the global financial crisis

Scholarly challenges to limited liability initially focused on harm caused by corporate torts.\(^{62}\) The concern was that limited liability created moral hazard:\(^{63}\) shareholders whose liability is capped at the value of their equity contribution would want their firms to engage in riskier – and therefore more tort-prone – projects than is socially optimal.\(^{64}\)

To mitigate this suboptimal risk-taking, some scholars advocated pro rata unlimited shareholder liability;\(^{65}\) others advocated a control-based liability regime.\(^{66}\) Under the former, shareholders would be liable for unpaid tort judgments in proportion to their equity ownership of the firm.\(^{67}\) Under the latter regime, only shareholders with a “capacity to control”\(^{68}\) their firms would be liable for tort judgments.\(^{69}\)

---

\(^{62}\) See Hansmann & Kraakman, supra note 31, at 1879–80 (briefly introducing the existing scholarship proposing to curtail limited liability).

\(^{63}\) Ibid. at 1882.

\(^{64}\) Mendelson, supra note 16, at 1232-33. I have previously considered an analogous problem: “marginalizing” risk in debt markets. Investors follow the conventional wisdom of risk dispersion – they diversify their investment portfolio to reduce risk. Risk dispersion, however, can lead investors to both underestimate and under-protect against risk due to market failures including information failure, model failure, human processing failure and collective action failure. This “marginalization” of risk can result in harm to the actual market participants and harm extending beyond the market participants, such as financial crisis or systemic collapse. See Steven L. Schwarcz, S. L., 2012, “Marginalizing risk,” 89 Washington University Law Review, 487.

\(^{65}\) Hansmann & Kraakman, supra note 31, at 1892-1909.

\(^{66}\) Mendelson, supra note 16, at 1271-1301.

\(^{67}\) Hansmann & Kraakman, supra note 31, at 1892; see Ibid. at 1906-09 (arguing that pro rata liability would marginally increase shareholders’ incentives to monitor and help avoid their firm’s risky activities).

\(^{68}\) Mendelson, supra note 16, at 1271-1301.

\(^{69}\) A shareholder with the “capacity to control” is a shareholder with a controlling ownership stake in the corporation. Mendelson, supra note 16, at 1271 n.281. These shareholders are uniquely positioned with low-cost access to corporate information and are better able to influence corporate decisions. Ibid. at 1249-52. Corporations that have a shareholder with the “capacity to control” are more likely to engage in excessively risky activities. See Ibid. at 1258 (explaining that shareholders with the “capacity to control” are more attracted to hazardous activities with small risk of large loss and better able to influence the corporation to engage in these activities). Consequently, “under limited liability, there is a set of risky activities that a company with a dispersed share ownership would not select, but a company with a controlling shareholder would.” Ibid.

\(^{69}\) Ibid. at 1271; see Ibid. at 1272-79 (detailing the logistics of a control-based liability regime).
Scholars also challenged the limited liability of bank shareholders, arguing for a return to double liability.\textsuperscript{70} Professors Macey and Miller used historical data on bank failures\textsuperscript{71} to argue that banks whose shareholders are subject to double liability may fail less frequently than banks whose shareholders have limited liability.\textsuperscript{72} Professor Grossman, an economist, also used historical data to conclude that double liability generally had the effect of reducing risk-taking.\textsuperscript{73} These studies countered the dubious narrative that because double liability failed to guarantee bank stability during the Great Depression, it does not reduce risk-taking.\textsuperscript{74}

\textsuperscript{70} See Macey & Miller, supra note 32 and accompanying text. Recall that double liability meant shareholder liability for obligations in an amount equal to the par value of their shares.

\textsuperscript{71} Macey and Miller scrutinized the recovery rate of national bank assessments, the liquidation of national banks, the average annual losses from national bank failures, and bank capital ratios. Ibid. at 55–61.

\textsuperscript{72} Ibid. at 32. Immediately following the Great Depression, conventional wisdom was that double liability failed to adequately protect bank creditors and therefore the system needed to be replaced by deposit insurance to better protect bank creditors. See supra notes 36–41 and accompanying text. Macey and Miller, however, concluded that double liability was a successful mechanism for protecting bank creditors and promoting sound banking practices because recoveries from shareholders were good, the recovery rate during this period was approximately 51%; banks were encouraged to privately transfer assets before failure; banks were encouraged to more cautiously manage assets to avoid shareholder assessment; and banks were allowed to operate with lower capital ratios. Macey & Miller, supra note 32, at 56, 61–62.

\textsuperscript{73} Grossman, R. S., 2001, “Double liability and bank risk taking,” 33 Journal of Money, Credit & Banking 143, 157 (“The effects of double liability can be seen through lower failure rates, higher capital ratios, and higher liquidity ratios among state banks in multiple-liability states.”). Professor Grossman examined state-level failure data on state and national banks, state-level balance sheet data for state and national banks, and data on individual state and national banks in Alabama and Mississippi prior to and following the introduction of double liability in Mississippi. Ibid. at 147.

\textsuperscript{74} Ibid. at 157–58. Grossman also examined empirical data to understand the motivations for legislating double liability or limited liability. Grossman, R. S., 2007, “Fear and greed: the evolution of double liability in American banking, 1865–1930,” 44 Explorations in Economic History 59. States legislating double liability were motivated by fear; they wanted to deter high levels of bank failures by motivating sounder banking practices. Ibid. at 71–75. States legislating limited liability were motivated by greed; they wanted to encourage growth through increased bank investment and risk-taking. Ibid.
2. Scholarly challenges after the global financial crisis
In response to the GFC, scholars have revisited the limited liability debate, focusing on moral hazard as a cause of excessive risk-taking in the financial industry. Professor White, an economist, posits that the crisis resulted from the “steady erosion of incentives that induced management to control risk,” including investment banks’ shift from partnerships with unlimited liability to corporations with limited liability.75 Law professors Hill and Painter have independently made and developed similar arguments.76 Other scholars, such as Peter Conti-Brown, have echoed these arguments, focusing on the impact of the risk-taking: shareholders protected by limited liability can “pocket the benefits generated by [their firms’] risky activities,” while the costs of those risky activities is passed on to the government through bailouts and ultimately onto taxpayers.77

75  supra note 7, at 2–3. White suggests that double liability for banks reflected a “superior alignment of liabilities and incentives for shareholders and managers so that the risk-taking temptations of managers were controlled.” Ibid. at 4–6 (supporting this conclusion with numerical examples to demonstrate that the system of double liability incentivized shareholders, senior bank management, and directors to control and reduce risk-taking).

76  See infra notes 80–84 and accompanying text.

These arguments effectively continue the pre-crisis observation that limited liability theoretically incentivizes risky action. The investment banking example of the shift from unlimited to limited liability also helps to concretize this observation.

Prior to 1970, investment banks were organized as partnerships, with partners facing personal liability for the debts of their firm if it fails. This restrained risk-taking. Following a change to the New York Stock Exchange rules allowing brokerage firms to have publicly owned shares, investment banks switched to the corporate form.

---

78 Admati, A. R., and P. Pfleiderer, 2010, “Increased-liability equity: a proposal to improve capital regulation of large financial institutions,” 4 (Rock Center for Corporate Governance, Working Paper No. 68, 2010; Stanford University Graduate School of Business, Research Paper No. 2043). Generally, a firm’s directors owe a fiduciary duty to shareholders and not to creditors. In a limited context, however, a firm’s directors may owe a fiduciary duty to creditors: when the firm is in the vicinity of insolvency. There is, however, no hard and fast rule on which a director can rely. Therefore, it is unclear how a director should balance between protecting creditors and encouraging a corporation to innovate and take appropriate business risks. In practice, corporations take more risks to fulfill their duty to shareholders and ultimately shift the risk on to creditors. See Credit Lyonnais Bank Nederland, N.V. v. Pathe Commc’ns Corp., Civ. A. No. 12150, 1991 WL 277613 (Del. Ch. Dec. 30, 1991); Schwarcz, S. L., Structured finance, a guide to the principles of asset securitization § 3:2.1 (3d ed. 2002 & Supp. 2006).

79 Notwithstanding this observation, some scholars have argued that debt creates incentives that impose discipline on a firm’s managers. See Admati et al., 2012, “Liability holding companies,” 59 UCLA Law Review 852, 857-58 (discussing the argument that debt funding helps to reduce agency costs by imposing discipline of corporate managers). Covenants are a central mechanism used by creditors for limiting a firm’s ability to take excessive risks and monitoring a firm’s management. See generally Jensen, M. C., and W. H. Meckling, 1976, “Theory of the firm: managerial behavior, agency costs and ownership structure,” 3 Journal of Financial Economics 305, 337-39; Triantis, G. G., and R. J. Daniels, 1995, “The role of debt in interactive corporate governance,” 83 California Law Review, 1073, 1074. Other scholars have noted, however, that this is inapplicable to banks. Admati et al., supra, at 858. Bank debt, backed by deposit insurance and implicit government guarantees, fails to incentivize creditors to monitor bank management because creditors expect to recover even if the bank fails. Ibid. This exacerbates the moral hazard problem because the government and ultimately taxpayers bear the downside risks. Admati & Pfleiderer, supra note 78, at 4.


81 Ibid. at 1179.

82 Ibid. at 1181.
Because investment bankers were now protected by limited liability, “frenetic risk-taking [allegedly] became the norm.” Professors Hill and Painter liken this to gambling with “house money,” in which the investment bankers have no personal stake.

3. Scholarly justifications for limited liability
Next compare these scholarly challenges to limited liability with the scholarly justifications for limited liability. The traditional scholarly justification for limited liability is the encouragement of equity-capital investment. This so-called “efficiency justification” holds that limited liability creates appropriate incentives for widespread investor participation in equity ownership, especially in large corporations.
Besides addressing investor risk aversion, limited liability reduces monitoring costs: an equity investor need not overly worry about monitoring the firm’s risky actions (for which the investor could ultimately become liable), nor worry about monitoring the wealth of other shareholders (for whose shareholder liability the investor could ultimately bear under a joint-and-several shareholder liability regime; this could be especially tricky given the free transferability of equity shares). Additionally, by promoting the free transfer of shares, limited liability serves to create incentives for managers to act efficiently: shareholders can transfer their shares if managers perform poorly, and the new shareholders can appoint new managers.

---

88 See Mendelson, supra note 16, at 1218-19 (discussing arguments as to how limited liability corrects excessive investor risk aversion). Scholars have argued that limited liability allows shareholders to diversify their investments and consequently encourages shareholder risk-taking. Under limited liability, no one investment will expose all of the shareholder’s assets to risk. Therefore, a shareholder can invest in a diversified portfolio of varying risk where each investment’s ultimate risk is limited to the amount of capital invested. Blumberg, supra note 16, at 616; Easterbrook, F. H., and D. R. Fischel, 1985, “Limited liability and the corporation,” 52 University of Chicago Law Review 89, 96-97 (Easterbrook & Fischel); Ribstein, L. E., 1991, “Limited liability and theories of the corporation,” 50 Maryland Law Review 80, 102. In the absence of these diversification benefits, some projects with positive net present values would be rejected as “too risky.” Easterbrook & Fischel, supra, at 97.

89 Blumberg, supra note 16, at 614; Easterbrook & Fischel, supra note 88, at 94-95; Ribstein, supra note 88, at 103.

90 Blumberg, supra note 16, at 614-15; Easterbrook & Fischel, supra note 88, at 95. Limited liability also contributes to uniformity in share price. Without limited liability, investors would attach different values to shares, depending on that investor’s wealth. Therefore, limited liability, by removing a shareholder’s individual wealth from contemplation of liability, results in homogenous shares with a single market price. See Easterbrook & Fischel, supra note 88, at 92; Halpern et al., 1980, “An economic analysis of limited liability in corporation law,” 30 University of Toronto Law Journal 117, 147.

91 Easterbrook & Fischel, supra note 88, at 95. Limited liability promotes the free transfer of shares because, as discussed supra note 90 and accompanying text, the costs of monitoring other shareholders are reduced, simultaneously reducing the cost of transferring shares to a new shareholder.

92 Easterbrook & Fischel, supra note 88, at 95. Easterbrook and Fischel also argue that legislative-imposed limited liability is more efficient than legislative-permitted limited liability because contracting around liability would be cost prohibitive, thereby discouraging capital investment. Ibid. at 93. Accordingly, limited liability effectively “creates an efficient contract term applicable to all transactions.” Blumberg, supra note 16, at 615-16.
4. How existing scholarship informs the limited liability debate

The existing scholarship recognizes that limited liability can create moral hazard, leading to excessive corporate risk-taking. On balance, though, limited liability encourages equity-capital investment by addressing investor risk aversion and by reducing investor monitoring costs. It may also create incentives for managers to act efficiently.

The existing scholarship does not examine limited liability in the context of the emerging shadow banking system. Even the scholarship that uses investment banks as an example does not purport to analyze how investment banks – which can be viewed as a type of shadow bank – might or might not be generally representative of shadow banking firms. This article next analyzes limited liability in the shadow banking context.

II. Analysis

Should limited liability be the norm for the governance structure of shadow banking? To answer that question, this article begins by deriving a normative framework for rethinking the corporate governance assumptions about limited liability. Thereafter, the article applies that framework to shadow banking.

---

93 See supra text accompanying notes 5-6.
94 Hill and Painter focus on investment banking but nonetheless define the firms they discuss as “most firms that are federally insured banks or bank holding companies; firms that originate, buy, or sell mortgages; firms registered as broker-dealers or investment advisors under the Securities Exchange Act; and at least the larger hedge funds.” Hill & Painter, supra note 80, at 1190. Although this definition would, in part, cover the shadow banking system, their article lacks analysis as to whether investment banks might or might not be generally representative of shadow banking firms.
A. Deriving a normative framework

Any attempt to rethink assumptions relating to corporate governance should situate that inquiry within a normative framework. Because corporate governance rules are a subset of financial regulation,[95] that inquiry should be situated within a financial regulatory framework. The central purpose of financial regulation is correcting market failures.[96] Of the five general categories of market failures, three are potentially relevant to corporate governance: information failure, agency failure and externalities (externalizing harm onto third parties).[97]

---

95 Cf. “Corporate governance and the new financial regulation: complements or substitutes?” Research Newsletter (European Corporate Governance Institute), Spring 2011 (discussing the relationship between corporate governance and financial regulation).

96 See, e.g., Gowland, D., 1990, The regulation of financial markets in the 1990s, 21 (1990) (“[R]egulation can be beneficial if and only if there would otherwise be market failure.”). Welfare economists argue that regulation should also include the goal of maximizing social welfare. See, e.g., Wolf, C., 1979, “A theory of nonmarket failure: framework for implementation analysis,” 22 Journal of Law and Economics 107, 110-11 (noting that, from the perspective of welfare economists, “it is theoretically correct to consider distributional inequity as an example of market failure”).

97 The other two traditional market-failure categories — monopolies and other types of non-competitive markets, and the public goods problem (a form of collective action problem describing the inability of markets to provide goods that, like clean air, are non-excludable and non-rivalrous, since some parties will want to free ride on public goods when such goods are (inevitably) purchased by others) — do not appear to be relevant to corporate governance, much less to limited liability of shareholders.
In the shadow banking system, the third market failure – externalities – becomes much more important. That is because the paramount concern posed by the shadow banking system is that it “can, if left unregulated, pose systemic risks to the financial system.”98 Systemic risks, in turn, can cause massive harm to the real economy.99

In discussing the third market failure, I have argued, however, that “externalities” is a misleading term because it conflates cause and effect. Externalities are not the cause but, instead, the consequences of a failure.100

---

98 Schwarcz, supra note 3, at 625. See also Bakk-Simon et al., 2012, “Shadow banking in the Euro Area, European Central Bank Occasional Paper No. 133, Apr. 2012, at 3 (observing that disintermediation is “one of the main sources of financial stability concerns”).

99 Schwarcz, S. L., Systemic Risk, 97 Georgetown Law Journal 193, 207, 235 (2008) (attempting to estimate the costs of a systemic failure of the financial system, which could go beyond direct economic costs and include indirect “social costs in the form of widespread poverty and unemployment”); cf. Anabtawi, I., and S. L. Schwarcz, 2011, “Regulating systemic risk: towards an analytical framework,” 86 Notre Dame Law Review 1349, 1376 n.109 (arguing that because financial market participants are able to externalize significant social costs associated with their risk-taking, it is in their interest to take on excessive risks and oppose regulatory efforts to curtail, or increase the costs of, their ability to do so).

100 Economists often recognize, for example, that a market failure has occurred if the production of goods or services results in externalities. See, e.g., Bator, F. M., “The anatomy of market failure,” 72 Quarterly Journal of Economics 351, 351 (1958) (defining market failure as “the failure of a more or less idealized system of price-market institutions to sustain ‘desirable’ activities or to estop ‘undesirable’ activities”). The cause of the market failure is not externalities per se, however; rather, it is the problem with the production of goods and services that resulted in the externalities. The externalities merely signal that a market failure has occurred. Cf. Sunshine, M., 2009, “How did economists blow it (Part 2)? - They missed the negative externalities of America's limited liability society,” Sunshine Report (Sept. 8), http://www.thesunshinereport.net/marksunshine/?p=402 (arguing that although economic “theories about efficient markets and logical behavior are pretty good, the fundamental application of these theories stinks”).
When discussing the causes of market failures, I have proposed that we substitute for “externalities” the term “responsibility failure” – referring to responsibility for a firm’s ability to externalize a significant portion of the costs of taking a risky action.¹⁰¹ This article will use that more precise terminology.

This article therefore next analyzes limited liability from the standpoint of three market failures: agency failure, information failure and responsibility failure. To further clarify the terminology, references below to “limited liability” shall mean limited liability of equity investors; references below to “equity investors” shall mean shareholders and other persons, whether or not they own shares of stock, who share in the firm’s profits;¹⁰² and references below to “investors” shall mean equity investors.

¹⁰¹ Schwarcz, supra note 1, 1799-1800. Referring to externalities as a type of market failure is also misleading because externalities cannot be considered a truly distinct type of market failure; all types of market failures can result in externalities. Zerbe, R. O. and H. E. McCurdy, 1999, “The failure of market failure,” 18 Journal of Policy Analysis & Management 558, 561; cf. Andreas A. Papandreou, A., 1994, Externality and institutions 167-69 (“[I]f externality is simply another word for market failure, or institutional failure . . . the notion of externality becomes redundant.”). To avoid this circularity, some economists have even questioned whether “externalities” should denote a separate market-failure category. See, e.g., Papandreou, supra, at 99-100 (arguing that the “non-existence of markets” is the actual market failure referred to as “externalities” and that it is “not useful to treat externalities as a subset of market failure, nor for that matter as a cause of market failure”); Zerbe and McCurdy, supra at 562 (arguing that externalities should not be defined as market failures); cf. Ibid. at 564 (arguing that “a close examination of the market failure concept gives rise to all sorts of definitional problems” related to externalities).

¹⁰² This might include, for example, a manager who has a bonus-compensation scheme based on the firm’s profits.
B. Applying the framework to shadow banking

1. Limited liability and agency failure
Agency failure refers to conflicts of interest between principals and their agents — in our case, between a firm and its investors on the one hand and the firm's managers on the other hand. Because it is not a principal-agent conflict, agency failure does not include the conflict in which investor-managers have strong incentives to take risks that could generate out-size profits, even if that greatly increases systemic risk.

Examples would include conflicts between investors of a firm and the firm's officers and directors. This market failure does not appear to be directly relevant to analyzing the limited liability of investors.

2. Limited liability and information failure
Information failure refers to asymmetric information, or even mutual misinformation, between parties. An example of the former would be the insufficiency of disclosure to completely inform investors of the merits and risks of a highly complex securities investment. Information failure can arise from various causes, including the potential for transaction costs relating to information acquisition to diminish the incentive to acquire such information.

---

103 Because it is not a principal-agent conflict, agency failure does not include the conflict in which investor-managers have strong incentives to take risks that could generate out-size profits, even if that greatly increases systemic risk.

104 See Schwarcz, supra note 8. This market failure may be indirectly relevant to limited liability. For example, the greater the investor liability, the greater the incentive of investors to align managers to their interests.


This market failure is directly relevant to limited liability. If investors stand only to lose their equity investments in a firm, they will have limited incentives to incur costs monitoring the firm. I next argue, however, that monitoring incentives will be even more limited for investors of shadow banking firms. Among other reasons, that is because those investors may not bear all of the adverse consequences of such limited monitoring. Increasing investor liability can help to increase those monitoring incentives.

That does not fully address, however, the design of the monitoring incentives. Ideally, increasing investor liability should encourage not only monitoring of the firm, qua firm; it should also encourage monitoring of the firm’s potential to trigger systemic risk, thereby externalizing risk. I next address that concern.

3. Limited liability and responsibility failure
In the shadow banking system, limited liability is an important source of responsibility failure that can lead to externalities – and, even more significantly, to systemic externalities. If a shadow banking firm cannot pay all of the externalized harm for which it becomes liable, the firm’s investors will not (qua investors) be financially responsible beyond the capital they have invested. To that extent, limited liability creates moral hazard.

107 See infra notes 121-31 and accompanying text (explaining why investors of shadow banking firms will not bear all of the adverse consequences of a systemic collapse).


109 See supra notes 63-64 and accompanying text (discussing limited liability and moral hazard).
Admittedly, limited liability creates moral hazard even outside the shadow banking system. It is widely understood, for example, that “most of a corporate structure’s externalities result from the limited-liability rule of corporation law.”\textsuperscript{110} Moral hazard in shadow banking, however, is more likely to have systemic consequences. There are two reasons why: decentralization makes managers of limited liability shadow banking firms more likely to take risks than managers of other limited liability firms, thereby making shadow banking firms more likely to fail; and disintermediation makes the consequences of a shadow banking firm’s failure more likely to be systemic than the consequences of an ordinary firm’s failure.

First consider why decentralization makes managers of limited liability shadow banking firms more likely to take risks than managers of other limited liability firms. The relatively small firms, such as hedge funds, that operate in the shadow banking system are often managed directly by their primary investors.\textsuperscript{111}

\textsuperscript{110} Schwarcz, S. L., 2004, “Collapsing corporate structures: resolving the tension between form and substance,” 60 Business Lawyer 109, 144.

Because such investor-managers typically are entitled to a significant share of their firm's profits, they have strong incentives to take risks that could generate large profits. Some risks might even potentially generate such outsized profits that investor-managers would gain lifetime financial security. Yet if a risky action exposes their firm to significant liability for externalized harm, limited liability protects those investor-managers from losing more than their invested capital.

---


This is radically unlike the management incentives in non-shadow banking firms. In those firms (such as ordinary business corporations and even traditional banks), senior managers tend to share only indirectly in profits, such as through stock options. Most profits ordinarily are paid to non-manager investors. Furthermore, managers are often invested in maintaining their jobs. They are, therefore, much less motivated to take actions that risk the firm, such as exposing the firm to significant liability for externalized harm.

Managers of limited liability shadow banking firms are thus more likely to take risks than managers of other limited liability firms, thereby making shadow banking firms more likely to fail.

---


117 Cf. Hu, H. T. C., “Risk, time, and fiduciary principles in corporate investment,” 38 UCLA Law Review 277, 320 (1990) (observing with respect to a traditional firm that “[t]he manager cannot take as cavalier an attitude toward the diversifiable risks of his corporation as the stockholder can. If a corporation does badly because a new investment project fails, a manager must rely primarily on other projects undertaken by the same corporation to balance against it. His salary and prospects and his value to a potential new employer would be hurt by poor firm performance”).

118 Shadow banking firms are also more likely to fail than traditional banks because they are not subject to the type of solvency/prudential regulation imposed on traditional banks, see infra note 124, and also because shadow banking firms are not limited in their activities to financial intermediation so they can also take other types of business risks. See Crockett, A., and B. H. Cohen, “Financial markets and systemic risk in an era of innovation,” 4 International Finance 127, 137-38 (2001) (discussing that the expanded range of activity by financial institutions has increased “the exposure of institutions to [market] shocks” and also observing that the wider range of markets in which financial institutions are engaged has increased the “number of potential channels of propagation of shocks”).
That likelihood of failure is further increased by the fact that shadow banking financial intermediation, like traditional bank financial intermediation, often involves the short-term funding of long-term projects.\textsuperscript{119} This creates a liquidity risk that the short-term debt cannot be rolled over. A failure to roll over (i.e., refinance) short-term debt could result in the firm’s default.\textsuperscript{120}

The consequences of a shadow banking firm’s failure are also more likely to be systemic than the consequences of an ordinary firm’s failure. Like traditional banks, shadow banking firms engage in financial intermediation on which the real economy is dependent. Because all financial intermediaries – including shadow banking firms and traditional banks – tend to be highly interconnected, the failure of a shadow banking firm could trigger the failures of other financial intermediaries.\textsuperscript{121} Such a chain of failures would be the epitome of a systemic event, especially if it materially reduces the availability of financial intermediation.\textsuperscript{122}

Additionally, the aforesaid short-term funding of long-term projects not only increases the likelihood of a shadow banking firm’s failure\textsuperscript{123} but can increase the systemic consequences of that failure. Among other reasons, an event that prevents one shadow banking firm from refinancing can also prevent, or can be correlated to other events that also prevent, other shadow banking firms from refinancing.

\textsuperscript{119} SIVs, for example, issue short-term commercial paper to fund long-term projects. See Schwarcz, supra note 1, at 1805. Money-market mutual funds also provide short-term loans to fund long-term projects. Ibid. Admittedly, the traditional business of banking also uses short-term borrowing from depositors to finance long-term loans to bank customers. Ibid. at 1806. However, the potential systemic externalities are offset by prudential regulation and deposit insurance. See infra note 124.

\textsuperscript{120} See Schwarcz, supra note 1, at 1806.

\textsuperscript{121} See Anabtawi & Schwarcz, supra note 99, at 1355 (“The financial system is comprised of institutions that are highly interrelated. In this sense, it is a ‘network.’” (quoting Brunnermeier, M. K. 2009, “Deciphering the liquidity and credit crunch 2007-2008, 23 Journal of Economic Perspectives 77, 96)); see also Ibid. at 1356-61 (explaining how inter-institution correlation, or interconnectedness, helps to explain why certain failures become systemic and others, like Enron’s failure, do not).

\textsuperscript{122} See Schwarcz, supra note 99, at 204.

\textsuperscript{123} See supra notes 119-120 and accompanying text.
Because shadow banking does not (at least currently) require the type of solvency/prudential regulation imposed on traditional banks, such a correlated failure would be highly systemically risky. Economists have identified the failure of shadow banking firms to roll over short-term debt as a contributing factor to the GFC.

In summary, although the limited liability of firms can always cause externalities, the limited liability of shadow banking firms is much more likely than the limited liability of non-shadow banking firms to cause systemic externalities. This is certainly a quantitative distinction: systemic externalities can cause much more harm than non-systemic externalities. I next argue that this is also a qualitative distinction.

An important function of law, and in particular tort law, is to operate as a mechanism for internalizing a firm’s externalities. By empowering injured third parties to sue a firm for harm, tort law helps to internalize a firm’s externalities because the firm will either pay the cost of the externalities ex-post through a judgment or will not engage in the activity ex-ante to avoid the potential costs of an adverse judgment.

---

124 In bank-intermediated finance, systemic externalities, specifically bank runs, are mitigated by prudential regulation and deposit insurance. See Schwarcz, supra note 1, at 1810-11. In contrast, prudential regulation does not or cannot be applied to many shadow banking firms. Ibid. at 1811-15. Meanwhile, disintermediation by spurring the short-term funding of long-term projects can mimic the effects of a bank run. Ibid. at 1807-08. Section 113 of the Dodd-Frank Wall Street Reform and Consumer Protection Act empowers the Financial Stability Oversight Council (FSOC) to designate nonbank financial institutions (SIFIs) and therefore subject them to consolidated supervision by the Federal Reserve and enhanced prudential standards. Notably, as of this writing, only three firms have been designated as a SIFI by the FSOC. “Designations,” Financial Stability Oversight Council, http://www.treasury.gov/initiatives/fsoc/designations/Pages/default.aspx (last visited Sept. 28, 2014) (listing designated SIFIs as only American International Group, Inc., General Electric Capital Corporation, Inc., and Prudential Financial, Inc.).

125 See Schwarcz, supra note 1 at 1807-08 (discussing the findings of economists Gary Gorton and Andrew Metrick, Federal Reserve Board economists, and the European Central Bank).

126 See supra note 99 and accompanying text.
Still, limited liability has practical consequences for traditional banking firms: third parties injured by even non-systemic harm are not always able to recover damages that perfectly internalizes the harm because the firm causing the harm may have insufficient capital to pay the damages.127

The consequences of limited liability for shadow banking firms, however, are qualitatively different: third parties injured by systemic harm will have virtually no chance to recover damages that internalizes the harm. Although this difference arises partly because shadow banking firms are unlikely to have sufficient capital to pay for the massive and widespread harm caused by a systemic collapse,128 the conceptual reason for the difference is legal: as explained below, existing law does not — and adaptations to traditional law are unlikely to — effectively internalize that harm.

As mentioned, tort law operates as a mechanism for internalizing a firm’s externalities. To win their lawsuits, injured third parties normally must show their harm to be a causal129 and foreseeable130 consequence of the firm’s actions. Third parties injured by systemic harm caused by the firm’s actions would unlikely be able to show that, however. Systemic harm can affect a wide range of third parties in unpredictable ways, such as an individual who is forced to close her family-owned restaurant during a systemically caused recession.

127 Cf. supra notes 62–64 and accompanying text (discussing scholarly arguments that limited liability generally encourages firms to engage in riskier, and therefore more tort-prone, projects than is socially optimal). Third parties injured by non-systemic harm, specifically tort victims, might not always be able to recover damages. Cf. Hansmann & Kraakman, supra note 31, at 1881 (“Already, strong empirical evidence indicates that increasing exposure to tort liability has led to the widespread reorganization of business firms to exploit limited liability to evade damage claims.”).

128 Cf. Schwarcz, supra note 99, at 206 (discussing the extent and magnitude of that harm).

129 Under the “but-for” test of causation, “causation exists only when the result would not have occurred without the party’s conduct.” Black’s Law Dictionary, supra note 2, at 228.

Nor is it likely that changing the causation-and-foreseeability standard to enable third parties to win those lawsuits would be an efficient solution. At the very least, courts would face a line-drawing problem and would be forced to make complicated decisions as to what systemic externalities should (or should not) be considered compensable.

The analysis has shown that limited liability investors of shadow banking firms lack sufficient monitoring incentives to avoid engaging in systemically risky actions, and that adaptations to traditional legal remedies are unlikely to adequately internalize systemic externalities. I next examine whether redesigning limited liability could be a more efficient solution to this problem.

III. Redesigning limited liability
Because it increases responsibility failure in the shadow banking system, limited liability should be redesigned if the benefits of the redesign outweigh the costs. This part examines how that might be done.

A. Setting goals for redesigning limited liability
Any redesign of limited liability should at least take into account its traditional justifications. The goal of the “efficiency” justification is to create incentives for widespread investor participation in equity ownership.

131 In theory, civil damages can be imposed under existing law for all harms, regardless of foreseeability, pursuant to the allocation-of-resources justification of enterprise liability. See Calabresi, G., 1961, “Some thoughts on risk distribution and the law of torts,” 70 Yale Law Journal 499, 529. The rationale for enterprise liability is that prices should reflect the “actual costs” of goods so as to allow purchasers to make informed decisions. Ibid. at 502. Therefore, “the cost of injuries should be borne by the activities which caused them,” regardless of fault, because injuries represent a “real cost” of those activities. Ibid. at 505. Foreseeability is irrelevant under enterprise liability because unforeseeable harms are “just as truly costs” of doing business as foreseeable harms. Ibid. at 529.

132 Other possible non-traditional legal solutions, such as regulation directed at ensuring shadow banking firm solvency, are beyond the scope of this article. Cf. Posner, supra note 9, at 413-14 (stating that “it is arguable that the government should take a more active role in assuring the solvency of corporations” to mitigate externalities caused by corporate limited liability, and observing that “the regime in banking . . . and in European corporate law” uses “continuous regulatory scrutiny of the corporation by an administrative agency”).

133 See supra notes 85-87 and accompanying text.
Investors have risk aversion;\textsuperscript{134} the more risk they are exposed to, the less incentive they will have to invest. Even fully rational investors will refuse to invest if their risk is unlimited, because the expected value of their losses might well exceed the expected value of their gains.\textsuperscript{135}

Therefore, any redesign of limited liability should attempt not only to minimize investor risk aversion but also to make investors comfortable that the expected value of their potential gains should exceed (by a sufficient margin to encourage investment) the expected value of their potential losses.\textsuperscript{136} The most effective way to accomplish that would be to set some type of cap, or limit – such as double liability, which a recent study suggests is a “superior alignment of liabilities and incentives”\textsuperscript{137} – on the potential liability.

Another traditional justification for limited liability is that it reduces monitoring costs because an investor need not overly worry about monitoring the firm's risky actions, for which the investor could ultimately become liable.\textsuperscript{138} In the shadow banking system, this justification should be given less weight because the potential for shadow banking firms to trigger systemic risk may well justify, if not necessitate, increased monitoring.

A related monitoring-cost justification for limited liability is that an investor need not worry about monitoring the wealth of other investors, whose liability the investor could ultimately bear under a joint-and-several liability regime.\textsuperscript{139}

\begin{footnotes}
\item[135] But cf. Conti-Brown, supra note 77, at 414, 446 (arguing that the successful operation of investment banks as partnerships until late in the 20th century suggests that limited.\textsuperscript{136} A related justification for limited liability was that as corporations increase in size, there are more associated risks, and the risks are consequently of much larger scale, which can serve to discourage investment by all but the wealthiest of investors. Blumberg, supra note 16, at 612-13. Shadow banking firms, however, tend to be relatively smaller.
\item[137] See supra note 75, at 4.
\item[138] See supra note 89 and accompanying text.
\item[139] See supra note 90 and accompanying text.
\end{footnotes}
This risk — including the complications it entails due to the free transferability of equity shares — could be avoided by a redesign, such as double liability, in which any given investor’s liability would be independent of the liability of other investors.

Yet another traditional justification for limited liability is that by promoting the free transfer of shares, it serves to create incentives for managers to act efficiently: investors can transfer their shares if managers perform poorly, and the new investors can appoint new managers.140 This justification does not appear to be compelling.141 Even existing investors could choose to vote out existing managers.

The final traditional justification for limited liability turns on fairness: that imposing additional liability on investors who did not contribute to management decisions created a perception of unfairness.142

140 See supra notes 91-92 and accompanying text.
141 This justification may make even less sense for shadow banking firms, like hedge funds, that are closely held. Some argue that limited liability is not efficient for small, closely held corporations. Halpern et al., supra note 90, at 148. Their rationale includes that limited liability then introduces more aggregate costs than in large, publicly traded corporations and also incentivizes shifting risk to parties less capable of bearing it, such as involuntary and trade creditors. See Freedman, J., 2000, “Limited liability: large company theory and small firms,” 63 Modern Law Review 317, 332-35.
142 See supra notes 56-57 and accompanying text (discussing this liability in the context of double liability of bank shareholders). It may well be appropriate to consider fairness as a normative goal. See, e.g., Kaplow, L., and S. Shavell, 2001, “Fairness versus welfare,” 114 Harvard Law Review 961, 1015 (arguing that it is appropriate to consider distributional fairness in making a policy decision).
If, however, such liability is imposed only on investor-managers, which is this article’s shadow banking focus – and perhaps only on investor-managers with the power to “control” the firm – then it should be more consistent with the fairness justification. Moreover, imposing such liability on only those investors would be efficient insofar as it ties the increase in liability to control over risk-taking decisions.

In sum, any redesign of limited liability in the shadow banking system should have at least the following goals: (i) it should increase such liability in a way that increases investor incentives to monitor (and guard against) the firm’s potential to trigger systemic risk; (ii) it should minimize investor risk aversion and encourage investment by setting a cap on liability sufficient to make investors comfortable that the expected value of their potential gains should exceed, by a sufficient margin, the expected value of their potential losses;

Investor-manager liability may also have indirect benefits. For example, limited liability might undermine – and thus increasing liability might support – the business judgment rule and the policies behind the rule, such as keeping courts from judging business decisions through hindsight-biased lenses, and allowing managers to take the risks necessary for maximizing investor value. See, e.g., In re Citigroup Inc. Shareholder Derivative Litigation, 964 A.2d 106, 124 (Del. Ch. 2009) (stating that doctrines such as the business judgment rule “properly focus on the decision-making process rather than on the substantive evaluation of the merits of the decision” to counteract the hindsight bias that some courts exhibit); Kaal, W. A., and R. W. Painter, 2010, “Initial reflections on an evolving standard: constraints on risk taking by directors and officers in Germany and the United States,” 40 Seton Hall Law Review 1433, 1440–41, 1449 n.56 (2010) (stating that “managers often must take reasonable risks” in the ordinary course of business and arguing that hindsight bias can have a significant influence on whether risks taken are viewed as reasonable or excessive); cf. Armour, J., and J. N. Gordon, 2014, “Systemic harms and shareholder value,” Journal of Legal Analysis (forthcoming 2014), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2307959 (arguing that the business judgment rule protection, which makes sense for officers and directors of a non-financial firm, leads to excessive risk taking in a systemically important financial firm; and proposing officer and director liability rules to counter that risk taking).

(iii) it should discourage cross-investor monitoring by ensuring that any given investor’s liability is independent of the liability of other investors; and (iv) to ensure fairness and maintain efficiency, it should increase liability only for investor-managers (because it is that dual nature that creates the real risk). Because a number of these goals are in tension, no possible redesign of limited liability could ever perfectly achieve all of these goals. For example, the greater the investor liability, the more investment would be discouraged.  

Any redesign would also have to confront the collective action problem of cross-border capital flight. I have already discussed how increasing investor liability might trigger capital flight between states. Domestically, that problem could be solved by increasing investor-manager liability under federal, as opposed to state, law. The collective action problem could also arise internationally, however. If one or more nations increase investor liability, investors may decide to move their money to firms in other nations, creating an international collective action problem. Cross-border cooperation would be needed to help mitigate that problem.

---

145 Cf. White, supra note 7, at 8 (observing a “potential trade-off” insofar as modifying limited liability might “create[] an incentive that induce[s] [bank] shareholders to act more cautiously, reining in managers’ efforts to expand loans that contributed to the growth of the economy”). Another potential concern is that increasing investor liability could make equity investments even more relatively expensive than debt, thereby increasing leverage. Although that is a theoretical concern, so many other considerations go into the determination of a firm’s capital structure that I do not believe it is likely to become a practical concern (and if it ever becomes a practical concern, that concern could be addressed through the issuance of non-voting equity).

146 See supra note 55 and accompanying text.

147 Federal law constitutionally allows piercing of the corporate veil to impose control-person liability in furtherance of federal statutory policies. Cf. Pension Benefit Guar. Corp. v. Ouiemt Corp., 630 F.2d 4, 12 (1st Cir. 1980) (holding that the Employee Retirement Income Security Act’s (ERISA) imposition of plan-termination liability jointly and severally on the plan sponsor and all members of its control group is not an unconstitutional denial of due process).

Finally, in furthering the first of these goals – increasing investor incentives to monitor and guard against systemic risk – this article does not claim that systemic risk could be eliminated.¹⁴⁹ Nonetheless, that goal is important because systemic externalities are the externalities most likely to cause widespread and serious harm.¹⁵⁰

B. Testing redesign proposals against the goals
This article does not purport to test all possible proposals to redesign limited liability against the foregoing goals. Consider in the context of the shadow banking system, however, the generalized proposals referenced earlier.¹⁵¹ Professors Hill and Painter propose using compensation as the test, imposing personal liability on investors that earn over a threshold number (such as U.S.$3 million annually).¹⁵² A purely compensation-based test has the drawback, however, that it does not target only investor-managers. Additionally, a compensation-based test might be able to be manipulated by creatively structuring compensation in ways that avoid the test.¹⁵³ Professors Hansmann and Kraakman propose the use of pro rata unlimited investor liability, under which investors would be liable in proportion to their investment.¹⁵⁴

¹⁴⁹ For analysis of the limits of law to control systemic risk, see Schwarcz, S. L., 2012, “Controlling financial chaos: the power and limits of law,” 2012 Wisconsin Law Review 815, 829–33 (discussing requiring banks and systemically important nonbank to pay for a systemic risk protection fund in order to help to internalize the most harmful externalities of the shadow banking system while stabilizing systemically important firms and markets) and Acharya et al., 2010, “A tax on systemic risk,” Feb. 3, at 2–4, http://vlab.stern.nyu.edu/public/static/SR-tax-nber.pdf (arguing that each institution should be required to pay a “tax” that is calculated according to the extent to which that institution is likely to contribute to systemic risk). Indeed, the law generally does not require that all externalities be internalized. See, e.g., Trebilcock, M. J., 1993, The limits of freedom of contract 20 (asking what types of externalities the law should require to be internalized).

¹⁵⁰ Schwarcz, supra note 99, at 206.

¹⁵¹ Those proposals are generalized insofar as they are not designed specifically for the shadow banking system. They nonetheless are worth testing because they reflect the serious thoughts of respected scholars on limited liability generally. See e.g., Hill & Painter, supra note 80.

¹⁵² See Hill & Painter, supra note 80, at 1188.

¹⁵³ It also is unclear how Hill and Painter’s proposal addresses risk aversion; they do not specify whether investor liability should be subject to any cap. See Hill & Painter, supra note 80.

¹⁵⁴ See supra note 65 and accompanying text.
Their proposal has two drawbacks: it may discourage investment because it does not set a cap on liability, and it does not target only investor-managers. Professor Mendelson proposes the use of a control-based liability regime under which only investors with the “capacity to control” would be liable. This proposal should be superior to pro rata unlimited investor liability because it better aligns the costs of risky activities with control. In a pro rata regime a controlling investor-manager’s liability is limited by his share of corporate equity and not his actual influence on risky activities. However, Mendelson’s proposal may discourage investment because it does not set a cap on liability.

Furthermore, a drawback to all of these referenced proposals is that they fail to effectively address systemic risk. Any redesign of limited liability must resolve the dilemma that investor-managers would have relatively little incentive to monitor and guard against their firm’s potential to trigger systemic risk if, as argued, tort law bars third parties injured by systemic harm from recovering damages.

---

155 See supra note 66 and accompanying text.
156 By the same logic that investors with the “capacity to control” can use their unique position to induce the corporation to engage in excessively risky activities, these investors can also use their unique position to limit firm risk. See Mendelson, supra note 16, at 1280-85.
157 Indeed, it is the inherently systemically risky nature of shadow banking that limits this article’s recommendations to the shadow banking sector. I am not suggesting, for example, that limited liability be changed for operating companies that do not engage in financial intermediation, even if such companies are run by dominant shareholders. The consequences of such a company’s failure are not likely to be systemic, and — all things being equal — the tort system is more likely to enable injured third parties to recover from the company, thereby helping to internalize the cost of the failure.
158 See supra notes 127-130 and accompanying text. Because there is always some risk of losing a lawsuit, investor-managers should have some incentive, albeit limited, to monitor and guard against their firm’s potential to trigger systemic risk.
A possible solution to this dilemma would be to couple any increase in limited liability with a privatized systemic risk fund — which would be used to mitigate systemic harm\(^\text{159}\) — into which systemically risky shadow banking firms would be required to contribute.\(^\text{160}\) Privatizing the funding would help to reduce a shadow banking firm’s incentive to create systemic externalities by engaging in financially risky activities;\(^\text{161}\) indeed, the likelihood that firms will have to make additional contributions to the fund to replenish bailout monies should motivate them to monitor each other and help control each other’s risky behavior.\(^\text{162}\) Making investor-managers personally liable, subject to a cap, for shortfalls in their firm’s fund contributions would likewise motivate them to monitor and help control their firm’s systemically risky behavior.\(^\text{163}\)

\(^{159}\) See, e.g., Schwarcz, supra note 149, at 829–30.

\(^{160}\) Ibid. (discussing, among other things, how a privatized systemic risk fund could assess which firms are systemically risky and how much each such systemically risky firm should be required contribute).

\(^{161}\) Ibid. at 830.

\(^{162}\) Ibid. at 830–31. Because their own funds would be at risk, for example, fund contributors would have incentives to inform regulators when other firms take unwise risks. Gordon, J. N., and C. Muller, 2011, “Confronting financial crisis: Dodd-Frank’s dangers and the case for a systemic emergency insurance fund,” 28 Yale Journal on Regulation 151, 155–56 (2011) (calling for a systemic emergency insurance fund that is funded by the financial industry). If the required contributions to the fund were risk-adjusted, fund contributors would also have incentives to report firms that are underpaying. Schwarcz, supra note 149, at 831.

\(^{163}\) Any redesign should attempt to take into account, however, the fairness of making investor-managers of a firm personally liable for that firm’s additional fund contributions necessitated by systemic harm caused by other firms.
IV. Conclusion

We tend not to focus on liability limitation at the firm level, simply accepting it as a fact of life. Limited liability is not always optimal, however, for firms that make up the shadow banking system. It motivates investor-managers of those firms to take risks that could generate outsize personal profits, even if that greatly increases systemic risk.

The law does not effectively mitigate these systemic externalities. Tort law, for example, traditionally helps to mitigate non-systemic externalities resulting from limited liability by empowering injured third parties to sue for harm that is a causal and foreseeable consequence of the tortfeasor’s actions. Systemic harm, however, affects a wide range of third parties in unpredictable ways; it is neither directly causal, nor clearly foreseeable.

To mitigate systemic externalities, limited liability should be redesigned, as explained in this article, for investor-managers of shadow banking firms. Any such redesign must balance the need to increase liability sufficiently to reduce systemic risk with not discouraging investment. The redesign should also minimize costs by discouraging the need to engage in cross-investor monitoring. These competing goals may well be achievable by restricting the increased liability to a capped multiple of the original investment, such as double liability.

Regardless of how limited liability is redesigned, it faces the dilemma that investor-managers would have relatively little incentive to monitor and guard against their firm’s potential to trigger systemic risk if, as indicated, tort law bars injured parties from recovering damages. A possible solution to this dilemma would be to couple the redesigned limited liability with a privatized systemic risk fund – which would be used to mitigate systemic harm – into which systemically risky shadow banking firms would be required to contribute. If their firm has insufficient capital to make these contributions, investor-managers would become personally liable for at least a portion of the insufficiency, thereby motivating them to monitor and help control their firm’s systemically risky behavior.

An ultimate question for any redesign of limited liability is empirical: will its benefits exceed its costs? The answer to that question will depend on the actual mechanics of the redesign, and their real-world impact on risk-taking and investment. Annex 1 below sets forth practical considerations that are likely to be relevant to any such cost-benefit determination.
Annex 1 - Guidance on redesigning limited liability
This annex provides practical guidance on redesigning limited liability, including guidance for assessing costs and benefits. Such guidance is qualified by, and intended to be interpreted by reference to, the more complete analysis in the article.

1. Limited liability should be increased only for investor-managers of firms that operate in the shadow banking system. Because the parameters of shadow banking are not yet well defined, the redesign should take into account how to delimit those firms. Because the primary goal of increasing limited liability is to reduce systemic risk, shadow banking firms designated as systemically important financial institutions (SIFIs) under the Dodd–Frank Act should at least be included in the redesign of limited liability. Other shadow banking firms should be considered for inclusion, however. Being relatively small, many shadow banking firms might avoid SIFI designation; nonetheless, they could be systemically important if, for example, their solvency or liquidity is highly correlated with that of other shadow banking firms.

2. The term “investor-managers” means equity investors who also have significant power to control the firm's actions. To minimize discouraging investment, the redesign should apply only to the subset of those equity investors who are entitled to a significant share of their firm’s profits, since they are the ones who have strong incentives to take risks with their firms.

3. Limited liability should be increased sufficiently to motivate investor-managers to monitor and guard against systemic risk. This could be done by increasing such liability to a multiple of the original investment, discussed overleaf.

---

164 See supra notes 3-5 and accompanying text.
165 See supra notes 97-99, 108-126 and accompanying text.
166 See supra note 124.
167 See supra notes 142-143 and accompanying text.
168 See supra notes 112-114 and accompanying text.
4. Any such increase in limited liability should be constrained so as not to unduly discourage investment.\textsuperscript{169} This could be done by setting a cap on liability, such as restricting liability to a small multiple of the original investment. Recent scholarship suggests that double liability might represent a good balance.\textsuperscript{170} What actually represents a good balance – e.g., a multiple of two (double liability), or less than two (e.g., 1-1/2 liability), or more than two – will ultimately be an empirical question.

5. Any increase in limited liability should minimize costs by discouraging the need to engage in cross-investor monitoring.\textsuperscript{171} Increasing such liability to a multiple of the original investment (such as double liability) would satisfy this requirement because any given investor-manager’s liability would be independent of the liability of other investors.

6. Because the law does not currently have an effective mechanism to internalize systemic externalities,\textsuperscript{172} an increase in limited liability would not, by itself, necessarily expose investor-managers to systemic costs for which they should be responsible. To mitigate systemic risk-taking, any increase in limited liability should be coupled with a mechanism to internalize those costs. One such mechanism could be a privatized systemic risk fund into which systemically risky shadow banking firms would be required to contribute. Investor-managers would thus be liable, up to their liability cap, for their firm’s insufficiency in making contributions.\textsuperscript{173} Creating and maintaining such a fund would have its own costs and political considerations.

7. To solve the collective action problem of cross-border capital flight (i.e., that increasing liability in any given jurisdiction could drive investors to firms in other jurisdictions) – or at least the fear of capital flight\textsuperscript{174} – federal law, rather than state law, could be used to increase limited liability in the U.S..\textsuperscript{175} Some form of cross-border cooperation would be needed, however, to help mitigate the problem of international capital flight.\textsuperscript{176}

\begin{footnotes}
\item[169] See supra notes 134-137 and accompanying text.
\item[170] See supra note 75 and accompanying text.
\item[171] See supra notes 90 & 139 and accompanying text.
\item[172] See supra notes 126-32 and accompanying text.
\item[173] See supra notes 157-62 and accompanying text.
\item[174] See supra notes 55, 145-148 and accompanying text.
\item[175] See supra note 147 and accompanying text.
\item[176] See supra note 148 and accompanying text.
\end{footnotes}
Abstract
Traditionally, the corporate treasury function in a bank (often simply referred to as “treasury”) is responsible for managing interest rate risk, maturity mismatch, and capital and funding strategy for the bank. The recent global financial crisis (GFC) highlighted deficiencies in the treasury operating model and functional capabilities to effectively manage capital, funding, liquidity, leverage and investments in a crisis scenario. Further, risk management and internal audit primarily focused on the Pillar 1 risks, and provided limited oversight on treasury risks (market risk in the banking book, counterparty credit risk, and funding and liquidity risk). In the aftermath of the GFC, corporate treasury is faced with a challenging market and regulatory environment characterized by low interest rates and margin compression, increasing divergence between lending and deposit growth rates, decline in availability of intraday credit from custody and clearing banks, increase in cash and high quality liquid asset (HQLA) holdings, increase in collateral requirements for derivatives and securities financing, constraints on short-term wholesale funding (STWF) dependency, and heightened prudential standards. The centralized treasury funding model is being challenged with increasing constraints on the mobility of capital, funding, liquidity and collateral across legal entities and jurisdictions. In response, the size of the corporate treasury function at global banks has increased dramatically, and banks are making significant investments in corporate treasury data and IT infrastructure. To ensure that the target state corporate treasury is effective and efficient, and that the investments are channeled toward a well-defined and coherent strategic vision, global banks will need to rethink the treasury operating model. This article highlights the key challenges facing corporate treasury, and the key considerations in redesigning a “best in class” treasury operating model.
1. Role of corporate treasury

To put it simply, the key objective of the corporate treasury function in a bank is to manage the balance sheet of the bank. This description, however, is not particularly helpful, and one needs to define the meaning of “managing the balance sheet of the bank.”

The balance sheet of a bank comprises assets, liabilities and off-balance sheet exposures, segregated between banking and trading book, and supported by scarce financial resources, namely, capital, funding and liquidity. Hence, “managing the balance sheet of the bank,” may be defined as management and allocation of financial resources of the bank (capital, funding and liquidity) to meet its strategic objectives within a range of market, and regulatory constraints, in normal and stress scenarios.

Firstly, to manage the balance sheet of the bank, corporate treasury focuses on managing the sensitivity of its balance sheet to market risk, primarily in the form of interest rate and foreign currency risk in the banking book. Market risk is managed by matching the re-pricing profile of assets, liabilities and off-balance sheet exposures, and hedging any residual market risk based on the bank's risk appetite and hedging strategy.

For banks with significant capital markets activities, the segregation between banking and trading book is a regulatory concept, and all market risk exposures are identified, monitored and managed based on the market risk framework. While market risk in the trading book is managed by the trading desk, and market risk management, corporate treasury is the primary owner (first line) of market risk in the banking book.

Secondly, one of the key economic functions of a bank is maturity transformation – a key source of funding liquidity risk. The failure of Lehman Brothers, and the liquidity crunch faced by many institutions during the GFC, reinforced the importance of liquidity management, and the need to reduce dependency on short-term wholesale funding.

Funding and liquidity risk is managed by matching the maturity profile of assets and liabilities (structural), and by holding a pool of cash and marketable securities that can be monetized in a crisis. Corporate treasury is responsible for managing funding liquidity risk, market liquidity risk and contingent liquidity risk arising from all assets, liabilities and off-balance sheet exposures; as well as managing the liquid asset buffer (pool of high quality liquid assets).
In addition, treasury is also responsible for managing the bank’s short-term funding and long-term debt issuance across all material currencies, managing the capital policy, plan and issuance, and acting as internal “market makers” for funding across all material currencies and tenors, thereby providing an economically competitive, risk-sensitive and transparent cost of funding.

2. Treasury operating model
Before the GFC, corporate treasury primarily focused on managing interest rate risk in the banking book, investments, capital and funding, working in individual silos, and often not aligned with finance and risk. The crisis highlighted the strong inter-dependencies between solvency and liquidity, and across different risk types, and the need to manage the balance sheet of the bank in an integrated manner. The corporate treasury organization and operating model needs to be redesigned based on the lessons learnt from the crisis, and to respond effectively to the new market and regulatory environment. In rethinking the treasury operating model, the bank should reassess governance, organization structure, roles and responsibilities, inter-dependencies, policies, processes, data and IT capabilities to meet the strategic objectives.

2.1 Governance
The corporate treasury function is typically a cost center, led by the treasurer, who reports to the chief financial officer (CFO). The asset liability management committee (ALCO) is the main executive oversight body, with representation from finance, risk, treasury and lines of business (LoB). The Board Risk Committee provides independent oversight on all material risks, including market risk, funding and liquidity risk, and investment risk. In some banks, the role of ALCO has been expanded and re-branded as the balance sheet management committee (BSCO), which is the key senior executive committee to make strategic and tactical decisions related to capital, leverage, funding, liquidity, hedging and investments.

2.2 Independent oversight
Prior to the GFC, risk and internal audit predominantly focused on the Basel Pillar 1 risks, and had limited oversight on corporate treasury. Risk management and internal audit had limited resources, as well as lacking in the appropriate technical skills and knowledge, to provide an effective oversight on risks within corporate treasury. Market risk in the banking book, funding and liquidity risk, investment risk and intraday liquidity risk are material risks
for most banks, and institutions will need to enhance the risk management framework and internal audit capabilities to provide an effective oversight and control environment.¹

The regulatory reform proposals call for significant enhancements to the risk management framework, including heightened expectations of the risk management committee and the role of the chief risk officer (CRO) in enterprise-wide risk management oversight, policies and framework. In addition, there is an increasing focus on managing model risk within corporate treasury, and ensuring that all key models are catalogued and subjected to model development and validation standards.

More specifically for liquidity risk, the risk management function is expected to review the adequacy and effectiveness of the liquidity risk management process; review and approve the risk appetite statement; provide independent validation of stress testing – scenario, assumptions, methodology and results; and approve the contingency funding plan. Further, any assumptions within treasury that maybe classified as a model should be subject to model development standards, and independent validation in line with the model governance framework. For funding and liquidity risk, banks have struggled to clearly define what assumptions are classified as a model. For example, operational deposit methodology for wholesale deposit is classified as a model for calculating the LCR (Liquidity Coverage Ratio), whereas LCR calculation as a whole is often not classified as a model.

2.3 Global banking, key considerations
For global banks, the operating model should consider the right balance between centralization versus decentralization, and the role of regional, country or legal entity treasury versus corporate treasury. The operating model should define the roles and responsibilities between corporate treasury, LoB and legal entities (LE). Typically, banks are structured by LoB as the primary management layer, and region or country as the secondary layer. However, with increasing regulatory constraints based on legal entity and jurisdiction, corporate treasury will need to rethink its LoB-centric view.

¹ IIF (2007): “Principles of liquidity risk management” recommendation number 7: firms should ensure that funding and liquidity risk management practices are incorporated in a firm-wide, integrated risk-management framework that also includes market, credit, operational and other appropriate risks, International Institute of Finance (IIF), March.
2.4 Impact of structural reform
The structural reform agenda, which includes increasing separation of retail and commercial banking from investment banking and subsidiarization of banking (operating as a subsidiary rather than a branch), will have a significant impact on the treasury operating model. The structural reform agenda will result in a fragmented global banking landscape, resulting in an adverse capital, funding and liquidity impact, wherein a global bank will be required to hold “locked” resources at a legal entity and jurisdiction level, rather than at a consolidated group level. The bank will also face heightened regulatory expectations on risk management and governance at a LE level, and more “heavy-handed” regulatory supervision in the host country.

2.5 Role of legal entity/country treasury
To deal with this new regulatory environment, banks will be required to build their treasury capabilities at an LE level for each significant jurisdiction, led by the LE or country treasurer. The LE treasurer will be accountable to the LE board and LE management team, and should be able to operate independently of the corporate treasury function, especially in a stress scenario. For example, the host country regulator would expect that the LE treasurer has control over the high quality liquid assets (HQLAs) maintained at a LE level, and will prioritize the interests of the LE (versus parent company) in a stress scenario. Further, the host country regulator will expect the LE treasurer to assess the LE capital adequacy requirements before repatriation of profits to the parent company in a foreign jurisdiction. In addition to meeting host country regulatory requirements, the LE treasurer will be accountable for the implementation of corporate treasury and risk management framework and policies within the LE.

2.6 Role of LoB treasury
The LoB treasury should have dual reporting to corporate treasury and LoB management, and be responsible for identification and measurement of market risk in the banking book, and liquidity risk within the LoB, as well as coordinating with corporate treasury on management and mitigation of the risk.
In addition, LoB treasury should have a strong understanding of the contractual and behavioral features of customers, products and financial instruments, and will be well-suited to “own” the key risk metrics – for example, prepayment curve on loans or rollover risk on term deposits. It is important to establish clarity on risk ownership for market risk in the banking book and liquidity risk between corporate treasury and LoBs, and align the funds transfer pricing (FTP) framework with risk ownership.

2.7 Think globally: one treasury
In spite of the increasing balkanization of global banking, global coordination of treasury activities will be critical for ensuring consistent standards and practices across global operations, and efficient use of financial resources (capital and liquidity). For example, funding and interest rate risk hedging for all major currencies should be centrally coordinated by corporate treasury to leverage internal netting, and eliminate internal arbitrage opportunities. Similarly, a global FTP framework is critical for consistent customer experience and pricing for a global conglomerate that has many touch points and interactions with the bank.

3. Functional pillars
Asset-liability management (ALM), FTP, funding and liquidity management, capital management and investment management are the key pillars of a corporate treasury function in a bank. In addition, with the increasing demand on collateral, corporate treasury will have a pivotal role in optimizing collateral across the bank.

Before the GFC, most banks managed the functional pillars on a “silo” basis and failed to capture the inter-dependencies – for example, impact of credit risk on prepayment risk. The CCAR/ICAAP and enterprise stress testing requirements are starting to capture some of cross-functional dependencies across treasury, finance and risk. Nevertheless, banks have a long journey ahead, and the “silo” treasury, risk and finance data and IT architecture continue to be a key impediment to integrated balance sheet management.

---

2 Comprehensive capital analysis and review (CCAR)/internal capital adequacy assessment process (ICAAP).
3.1 ALM and FTP
While ALM is relatively matured in banking, FTP is still an evolving discipline, especially for banks with significant trading book operations. One of the key challenges facing ALM is the potential impact of increase in interest rates, now projected to be at some point in 2015. In the last few years, banks have positioned their balance sheets to benefit from an increase in rates, mainly by repositioning their investment portfolios. In terms of emerging issues, the potential development of Pillar 1 capital charges for interest rate and credit spread risk in the banking book is likely to be one of the most significant changes in ALM.

As banks seek to improve their liquidity risk management data and IT capabilities, the ALM solution has been the primary source for contractual and behavioral cash flows for the banking book. Further, the ALM solution is a core component of the CCAR submission, and is the likely source for pre-provision net revenue (PPNR) projections in normal and stress scenarios.

FTP is the process of allocating cost of funding and liquidity, and transferring interest rate risk from the lines of business to corporate treasury at an appropriate level of granularity. Significant improvements are required in the FTP framework, especially for institutions with a significant trading book. The FTP framework should include a granular allocation of the contingent liquidity cost (CLC) based on the results of liquidity stress testing. Liquidity coverage ratio and net stable funding ratio (NSFR) will result in additional constraints on securities financing and prime brokerage business – the FTP framework will need to capture the cost of contingent collateral requirements. Further, the FTP framework should be capable of granular allocation across a range of dimensions, including transactions, positions, products, LoBs, customers and LEs.

3.2 Capital management
The regulatory reform agenda has been successful in strengthening the capital position at global systemically important banks (G-SIBs) by narrowing the list of instruments permissible within the definition of capital, increasing the capital requirements to an effective minimum of 7% and tightening the rules for calculation of RWAs, especially for trading book and counterparty credit risk, and introduction of the leverage ratio as a backstop to risk-weighted capital requirements.
The leverage ratio, due to be effective in 2018, is expected to be a binding constraint on banks with significant securities financing, derivatives and other off-balance sheet activities.

With the increase in regulatory capital (RC) requirements, the importance of economic capital (EC) as a tool for the measurement of overall level of risk or capital has declined.

Regulators have a strong influence on the dividend payout and share buy-back proposals based on the results of annual stress testing (CCAR/ICAAP), and have not been shy of exercising their influence. Although RC is now the key binding constraint for most banks, EC is still used as an internal measurement of risk for capital allocation and risk-based performance management.

Corporate treasury will lead capital planning and allocation, leverage ratio calculation and bank levy surcharges in coordination with lines of business, risk management, finance and tax.

3.3 Funding and liquidity management
G-SIBs have made significant progress in managing the funding and liquidity risk, as evidenced by a significant increase in the holding of HQLAs — estimated to be in excess of U.S.$12.2 trillion across all banks, of which more than 88% is in level 1 assets. Further, banks have significantly reduced their dependency on short-term wholesale funding, by changing the funding mix toward stable deposits, increasing the tenor profile and reducing dependency on funding based on illiquid collateral.

The LCR and the NSFR have been the key regulatory drivers in reducing vulnerabilities to funding and liquidity risk. In addition, banks have also experienced significant changes in the funding markets, such as decline in the size of the tri-party repo markets, lack of intraday credit from tri-party clearing banks (tri-party repo reform) and decline in availability of funding for illiquid securities.

---

Corporate treasury has started to make significant investments in talent, policies, process and infrastructure to deal with the changing regulatory and market environment.

To ensure independence, banks are in the process of segregating the short-term and long-term funding desk (execution) from the liquidity risk management function. The liquidity risk management will be responsible for the definition of liquidity risk appetite, strategic funding plan, stress testing and contingency funding plan, whereas the funding desk will be responsible for executing the funding strategy of the bank based on funding plan and market conditions.

3.4 Collateral management
Today, banks are under unprecedented pressure to manage and optimize collateral across their derivatives, secured financing, prime services and corporate treasury activities. The amount of collateral held and posted by financial institutions (FIs) is expected to increase sharply, driven by regulatory and market forces. In addition to the increase in collateral requirements, institutions are faced with increasing operational, legal and regulatory challenges in the movement of collateral across LEs, jurisdictions and LoB. Historically, collateral management has operated in “silos” across business lines, legal entities and jurisdictions, with limited single-ownership for collateral across the firm. FIs are in the process of redesigning the operating model for collateral management to deal with the changing regulatory and market environment, and shifting focus to think of collateral management as a strategic initiative rather than a series of operational issues. Some firms are in the process of setting up a “global collateral management function” to coordinate management, trading and optimization of collateral across the firm. Others are establishing “collateral pricing desks” to facilitate coordinated pricing of collateral into transaction and business activity. Institutions agree on the need for a more holistic and consistent view of collateral across the entire trade life cycle, risk management and operations. In addition, there has been ongoing industry dialogue with the custody and clearing banks on the need to provide an automated, real-time view of the cash and collateral position.

The funding desk within corporate treasury, in coordination with operations, will have a pivotal role in optimizing collateral across the bank.
3.5 Investment management
Most universal global banks take in more deposits than they make loans, and more recently the gap between deposit and loan growth rates have widened, resulting in significant surplus cash available for investments. Cash assets, including cash as well as balances due from central banks and other depository institutions, increased by 55.2% to U.S. $2.5 trillion in 2013 across commercial banks in U.S., and the trend is expected to continue in 2015.4

Figure 1

The chief investment office (CIO) is responsible for investing the short-term excess cash and providing a reasonable risk-adjusted return to the bank within the constraints of the “investment policy.” Typically, investments are in high-quality fixed income securities and derivatives to hedge balance sheet risk. In addition, corporate treasury is responsible for managing the pool of HQLAs to mitigate liquidity risk. Coordination between the CIO and corporate treasury function is important to avoid any potential conflict of interest between liquidity risk management and profit motives.

4 Federal Reserve Board (FRB) - H.8. Assets and liabilities of commercial banks in the U.S.
Rethinking the treasury operating model

In rethinking the mandate for corporate treasury, the bank should consider the potential conflict of interest between the liquidity risk management and investment management function within corporate treasury.

Leading to the GFC, we have observed cases where liquidity risk management was subjugated by the profit motive within corporate treasury, resulting in investments in high-yielding illiquid securities that could not be monetized in a crisis. This potential risk has been partially addressed by the strict eligibility criteria for HQLAs and regulatory calibration of the size of HQLAs (LCR). However, as the average return of equity (RoE) trends lower, the bank may be tempted to leverage the investment portfolio in corporate treasury as a source of incremental income.

Investment management is faced with a challenging regulatory and market environment, and will need close coordination with ALM, liquidity and capital to manage the overall balance sheet of the bank. Historically low interest rates and quantitative easing has led to low-yield on HQLAs, and increasing negative carry. The LCR rules have imposed additional constraints on the size and composition of the investment portfolio, for example, by limiting holdings of U.S. agency securities that have higher yields than U.S. treasury securities (agency securities are only eligible as a Level 2A security and subject to a 40% cap in LCR). Further, the U.S. Basel III rules have changed the treatment of unrealized gains and losses for the available for sale (AFS) portfolio, resulting in additional capital and liquidity considerations in investment decisions.

---

5 Negative carry is defined as the difference between the cost of funding and the yield earned on a security. In some cases, negative carry may also be defined in terms of opportunity cost, i.e., average return on asset less yield earned on the investment (cost of funding).
Rethinking the treasury operating model

4. Center of excellence (COE)
In the last few years, the size of the corporate treasury function has increased dramatically, driven by an increasing focus on capital, funding, liquidity and leverage, by both internal and external stakeholders. Further, before the GFC, investments in treasury resources, data and technology capabilities had lagged behind the balance sheet growth and complexity, and geographical expansion.

The size of the corporate treasury function at a large global bank has increased to 250-500 full-time employees, depending on the size of the balance sheet and business profile, and centralized-versus-federated organizational structure. Global banks with extensive international operations have significantly more treasury resources spread across a number of locations to access local currency funding and hedging, thereby making cost and resource rationalization more challenging.

The regulatory reform agenda has been a key driver in increasing the size of the treasury organization, including daily liquidity reporting and stress testing requirements; capital stress testing, planning and allocation; recovery and resolution planning; increasing focus on risk data aggregation and reporting capabilities; and increasing home and host regulatory interaction. Further, heightened board oversight and management reporting requirements have also contributed to an increase in the size of corporate treasury.

With the stabilization of the prudential regulatory environment and increasing maturity of some key treasury processes, global banks are now exploring opportunities to rationalize costs and resources in corporate treasury.

One of the key opportunities to realize cost and resource efficiencies is the migration of mature treasury processes to global or regional COEs, typically in a low-cost center. Differences in operating costs (e.g., human resource, real estate and infrastructure) between key treasury centers (New York, London, Hong Kong, Singapore and Tokyo) and low-cost centers (India, Scotland, Poland and Philippines) are estimated to be in the region of 30%. Further, right-shoring requires standardization, documentation and automation of processes leading to efficiencies, and a more robust internal control environment.
Rethinking the treasury operating model

The treasury right-shoring strategy should be aligned to finance, risk and technology right-shoring strategies, including co-location of key treasury processes with finance, risk and technology. For example, the daily FTP allocation and reporting process should be co-located with the finance management reporting process, and liquidity reporting co-located with regulatory reporting.

The key processes that may be suitable for right-shoring typically include, FTP allocation and reporting, liquidity reporting and risk-weighted asset reporting. In addition, treasury data and IT processes should be aligned with the broader data and IT right-shoring strategy (e.g., extract, transform and load (ETL) and data quality (DQ) for corporate treasury).

In developing a treasury COE, the banks should take into account risks, and critical success factors, including, availability of skilled resources in the target location and turnover; training of resources, extended parallel run, communication strategy, as well as well-defined service line agreements to manage operational risk and internal control failures.

5. Data and technology
For many years, investments in corporate treasury data and technology (IT) capabilities have lagged behind balance sheet growth and expansion of business operations. The GFC highlighted a number of deficiencies in treasury IT capabilities, including lack of appropriate granularity and frequency of data and reporting, inability to aggregate data by key dimensions (legal entity), inability to report available cash and collateral position in a timely manner, and misalignment of data and reporting across treasury, risk and finance. Banks lack a strategic solution for funding and liquidity risk, intraday cash and collateral management, enterprise stress testing and capital management.

5.1 Risk data aggregation and reporting
The Basel Committee issued the Principles for Effective Risk Data Aggregation and Risk Reporting (RDAR), which require significant enhancements to governance, infrastructure, data aggregation capabilities and reporting capabilities for all material risks. The treasury data and IT strategy should ensure alignment with RDAR principles; for example, the treasury data dictionary should be aligned with the enterprise data dictionary.
In the last five years, primarily driven by heightened expectations from the regulators, banks have undertaken significant investments, and re-engineering of treasury IT capabilities, in some cases committing more than U.S.$100 million. Banks have mobilized a dedicated treasury IT function, with a dual reporting line to the corporate treasurer and the chief information officer, to design, build, test and operate the treasury technology infrastructure. The treasury IT function is responsible for ensuring that corporate treasury adheres to enterprise data and technology standards.

Funding and liquidity risk management has been one of the most significant areas of IT investment in recent years. The objective of the liquidity solution is to build a single repository of contractual and behavioral cash flows, and collateral across all on and off-balance sheet exposures, legal entities, LoBs, currencies and maturities – refreshed on a daily basis. In addition, the solution will provide analytical and reporting capabilities, such as risk appetite and limit monitoring, stress testing and scenario analysis, management and regulatory reporting. The underlying data should be reconciled to the general ledger on a daily basis with well-defined materiality and exception thresholds, and adjustment process.

Although significant progress has been made in improving treasury IT capabilities across the industry (mostly across the GSIBs), much more needs to be done to address the deficiencies from the crisis. Banks have been forced to deliver solutions, often tactical and manual, to meet aggressive regulatory timelines, such as daily liquidity reporting in the U.K. and U.S. As a result, banks today lack a strategic vision and architectural blueprint for treasury IT that will support the business in the long run.

The definition of this strategic treasury IT architecture should start with the clear definition of business objectives and target operating model, key functional and analytical capabilities, reporting requirements, and data requirements, including, granularity, frequency and timeliness. In addition, the bank should ensure alignment of the treasury IT architecture to the broader risk and finance architecture, and enterprise data management and quality framework.
6. Conclusion
The corporate treasury function in a bank is at the epicenter of a challenging market and regulatory environment. The traditional treasury operational model will need to evolve, and corporate treasury will need to play a more strategic role in managing the forward-looking balance sheet of the bank, and optimize the use of scarce financial resources across lines of business, legal entities, products and customers. Although the key building blocks of the market and regulatory framework are now well-defined, a number of uncertainties and emerging risks still remain. Further, the lack of harmonization in structural reform agenda across U.S., U.K. and Europe, may lead to additional challenges for corporate treasury, and inefficiencies for banks.

6.1 Key emerging issues
Intraday liquidity risk management framework is weak across the industry, and although the Basel Committee has published a consultative paper on this topic, regulatory expectations on this topic are not well-defined and leading practices are still evolving. Intraday liquidity risk is a key risk for G-SIBs, more so for the financial market infrastructure organizations (FMIs) such as central clearing counterparties. Traditionally, intraday liquidity risk has been managed by operations, and the role of corporate treasury and risk management is not well defined.

Significant progress has been made in reducing the funding liquidity risk and fire-sale risk in the STWF markets, including tri-party repo infrastructure reform, regulatory constraints on securities financing activities (e.g., LCR, NSFR and leverage ratio), improvement in the collateral quality and increase in funding tenor, especially for illiquid collateral. Nevertheless, fire-sale risk in the STWF markets is an ongoing regulatory concern, especially with the increasing prominence of shadow banking. In recent discussions, regulators have explored the option of additional capital surcharges based on a bank’s reliance on STWF. 

---

6.2 Treasury, going forward
Finally, going forward, corporate treasury will be required to play a strategic role, with a seat at the table on strategic issues, such as strategic business planning, customer strategy, pricing, new product approval, acquisitions and performance management. Treasury will be expected to take a leading role in managing the balance sheet of the bank, dynamically, in an integrated manner, in coordination with LoBs, LEs, corporate finance and risk management. Hence, a well-defined treasury operating model will be crucial to the strategy of a bank in this new environment.
Abstract
We examine the relationship between bank holding company (BHC) size and components of non-interest expense (NIE), in order to shed light on the sources of scale economies in banking. Drawing on detailed expense information provided by U.S. banking firms in the memoranda of their regulatory filings, we find a robust negative relationship between size and normalized measures of NIE. The relationship is strongest for employee compensation expenses and components of “other” NIE, such as information technology and corporate overhead expenses. In addition, the authors find no evidence that the inverse relationship between banking firm size and NIE ratios disappears above a given size threshold. In dollar terms, their estimates imply that for a BHC of mean size, an additional U.S.$1 billion in assets reduces NIE by U.S.$1 million to U.S.$2 million per year, relative to a base case where operating cost ratios are unrelated to size.

1 This article is reprinted with the permission of the Federal Reserve Bank of New York, and will appear in the Federal Reserve Bank of New York Economic Policy Review, vol. 20, no. 2, Forthcoming, available at http://www.newyorkfed.org/research/epr/2014/1403kovn.html. Tables 4 and 7 from the original article, as well as Appendix A, have been omitted from this reprinting because of space constraints. Please see the original article for these materials. Appendix B is an online-only appendix available at: http://www.newyorkfed.org/research/epr/2014/1403kovn_appendixB.pdf.

2 Lily Zhou was a senior research analyst at the Federal Reserve Bank of New York at the time this article was prepared. We would like to thank Peter Olson for outstanding research assistance and Gara Afonso, Jan Groen, Joe Hughes, Don Morgan, an anonymous referee and workshop participants at the Federal Reserve Bank of New York for helpful comments and suggestions. The views expressed are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.
Introduction
The largest U.S. banking firms have grown significantly over time, their expansion driven by a combination of merger activity and organic growth. In 1991, the four largest U.S. bank holding companies (BHCs) held combined assets equivalent to 9% of gross domestic product (GDP). Today, the four largest firms’ assets represent 50% of GDP, and six BHCs control assets exceeding 4% of GDP. Despite recent financial reforms, there is still widespread concern that large banking firms remain “too big to fail,” i.e., policymakers would be reluctant to permit the failure of one or more of the largest firms because of fears about contagion or damage to the broader economy [see, for example, Bernanke (2013)].

A growing number of market observers advocate shrinking the size of the largest banking firms in order to limit the problem of too-big-to-fail. The most direct approach would be to simply impose a firm cap on the size of assets or liabilities; for example, Johnson and Kwak (2010) propose a size limit of 4% of nominal GDP. An alternative would be to impose levies or progressively higher capital requirements on large banking firms to encourage them to shed assets.

Would such policies impose any real costs on the economy? A number of recent academic papers suggest that the answer may be “yes” because of the presence of economies of scale in banking. Scale economies imply that the cost of producing an additional unit of output (for example, a loan) falls as the quantity of production increases. A number of papers find evidence of scale economies even among the largest banking firms [Hughes and Mester (2013); Wheelock and Wilson (2012); Feng and Serletis (2010)]. Taken at face value, this research implies that the introduction of limits on bank size would impose deadweight economic costs by increasing the cost of providing banking services.

We contribute to this line of research by studying the relationship between size and components of NIE, with the goal of shedding light on the sources of scale economies in banking. NIE includes a wide variety of operating costs incurred by banking firms: examples include employee compensation and benefits, information technology, legal fees, consulting, postage and stationery, directors’ fees, and expenses associated with buildings and other fixed assets.
Our hypothesis is that lower operating costs may be a source of scale economies for large BHCs, because large firms can spread overhead such as information technology, accounting, advertising and management over a larger asset or revenue base. Our analysis, therefore, tests for an inverse relationship between BHC size and scaled measures of different components of NIE.

One novel contribution of this paper is to make use of detailed NIE information provided by U.S. banking firms in the memoranda of their quarterly regulatory FR Y-9C filings. The Y-9C reports contain detailed consolidated financial statements and other data for U.S. BHCs (see Section 3 for details). Since 2001, about 35% of total NIE is classified in the Y-9C as part of a broad “other NIE” category. For the period 2008–12, we disaggregate this line item into nine author-defined categories, using memoranda information from Schedule HI of the Y-9C. In part, this involved manually classifying about 5,500 individual “write-in” text fields reported by individual BHCs. To our knowledge, ours is the first paper to make use of these data.

We start by estimating the relationship between BHC size (measured by the natural logarithm of total assets) and total NIE scaled by net operating revenue, assets or risk-weighted assets (RWA). We find a statistically and economically significant negative relationship between BHC size and these NIE ratios, robust to the expense measure or set of controls used. Quantitatively, a 10% increase in assets is associated with a 0.3% to 0.6% decline in NIE scaled by income or assets, depending on the specification. In dollar terms, our estimates imply that for a BHC of mean size, an additional U.S.$1 billion in assets reduces NIE by U.S.$1 million to U.S.$2 million per year, relative to a base case in which operating cost ratios are unrelated to size.3

These results hold across the size distribution of banking firms, and over different parts of our sample period. We find no evidence that these lower operating costs flatten out above some particular size threshold. The point estimate of the slope of the relationship steepens, if anything, although the statistical uncertainty associated with the estimate becomes larger owing to the small sample.

3 For details of this calculation, see Appendix B, available as a separate file at http://www.newyorkfed.org/research/epr/2014/1403kovn_appendixB.pdf. The appendix was omitted from the original article because of space constraints.
The relationship between size and the NIE ratio is negative for each of the three main components of NIE reported in BHC regulatory filings: employee compensation, premises and fixed asset expenses, and other NIE. Using our novel by-hand classification of other NIE into nine subcomponents, however, we find significant variation in the size-expense relationship among the subcomponents. The inverse relationship between size and expense is particularly pronounced for corporate overhead (for example, accounting, printing and postage), information technology (IT) and data processing, legal fees, other financial services, and directors' fees and other compensation. In contrast, large BHCs spend proportionately more on consulting and advisory services than do smaller firms, relative to revenue or assets. Large BHCs also incur proportionately higher expenses relating to amortization and impairment of goodwill and other intangible assets.

Overall, our results are consistent with the presence of scale economies in banking, as found in recent academic literature [for example, Wheelock and Wilson (2012); Hughes and Mester (2013); Feng and Serletis (2010)] and industry research [Clearing House Association (2011)]. In particular, our findings suggest that these scale economies stem in part from an operating cost advantage of large BHCs in areas such as employee compensation, information technology and corporate overhead expenses.

We emphasize that a number of caveats apply to our results. First, our estimates represent reduced-form statistical correlations; caution should be exercised in drawing a causal interpretation from them. Although our regressions control for a wide range of BHC characteristics, firm size may still be correlated with omitted variables that are also associated with lower expenses, such as the quality of management. This caveat also seems to apply more generally to the existing literature on scale economies in banking.

Second, our results may also reflect factors other than scale economies. One possibility, closely related to scale economies but conceptually distinct, is that large firms operate closer to their production frontier on average; that is, they have greater X-efficiency (see Section 2 for a discussion).  

---

Our analysis does not attempt to separate the effects of X-efficiency from those of scale economies. We note, however, that Hughes et al. (2001) and Hughes and Mester (2013) find that estimated scale economies are larger for more efficient banks than for less efficient ones, controlling for size.
Another possibility is that large banking firms have greater bargaining power vis-à-vis their suppliers and employees. If cost differences are due only to bargaining power effects, then limiting the size of the largest BHCs would not necessarily generate deadweight economic costs, although it might instead reallocate rents to employees or suppliers. An additional possibility is that our results are influenced by too-big-to-fail subsidies for large BHCs. Our prior is that such subsidies would be more likely to be manifested as a lower cost of funds for large firms, or a more leveraged capital structure, than as lower operating costs. However, it is still possible that a too-big-to-fail banking firm could respond by reducing expenditures on functions such as information technology or risk management; these would show up as part of NIE.

These caveats aside, our results and those of related research suggest that imposing size limits on banking firms is unlikely to be a free lunch. For example, taking our estimates at face value, a back-of-the-envelope calculation implies that limiting BHC size to no more than 4% of GDP would increase total industry NIE by U.S.$2 billion to U.S.$4 billion per quarter. Limiting the size of banking firms could still be an appropriate policy goal, but only if the benefits of doing so exceeded the attendant reductions in scale efficiencies.

A second contribution of this article is to present new evidence on other determinants of BHC operating costs. In particular, we find that proxies for organizational complexity (for example, the number of distinct legal entities controlled by the BHC), as well as measures of the diversity of business activities, are robustly correlated with higher expense ratios. This result appears consistent with prior research on the diversification discount in banking (for example, Goetz et al. (2013)). A third contribution is to present new stylized facts about the composition of NIE, based on our data collection efforts. For example, we document the large share of NIE that is composed of corporate overhead, investment technology and data processing, consulting and advisory services, and legal expenses.

Details of this calculation are presented in Appendix B, http://www.newyorkfed.org/research/epr/2014/1403kovn_appendixB.pdf.
The remainder of the article proceeds as follows: Section 2 presents background and reviews the literature on economies of scale in banking. Section 3 describes the data, discusses our method for classifying other NIE and presents descriptive statistics. Section 4 presents multivariate analysis of the relationship between size and NIE ratios. Section 5 studies components of NIE. Section 6 summarizes our findings.

2. Background and literature
Our analysis is closely related to academic literature on scale economies and organizational efficiency in banking. In a microeconomic production model, the cost function traces out the relationship between output and the minimum total cost required to produce that output, for a given set of input prices. A firm exhibits economies of scale if minimum cost increases less than proportionately with output – for example, if the firm could double its output by less than doubling its costs, holding input prices fixed.

A large literature empirically estimates the cost function for banks and/or BHCs, and tests for the presence of scale economies by measuring whether the elasticity of total costs with respect to output is greater than, equal to, or less than unity (indicating diseconomies of scale, constant returns to scale or economies of scale, respectively).

The earliest studies of scale economies in banking [for example, Benston (1972)], estimated during an era when U.S. banking organizations were on average much smaller than today, found evidence of modest economies of scale. Subsequent research, using more flexible cost functions, found that these scale economies were limited to small banks [for example, Benston et al. (1982) and Peristiani (1997); see also Berger and Humphrey [1994] for a survey].

More recent research, however, has found evidence of scale economies even among the class of large banks and bank holding companies. Examples include Wheelock and Wilson (2012), Hughes and Mester (2013), Feng and Serletis (2010) and Hughes et al. (2001). This departure from earlier findings reflects greater statistical power, attributable to the use of larger datasets with many more observations for large banking firms, as well as the evolution of empirical techniques.
Do big banks have lower operating costs?

For example, Wheelock and Wilson (2012) estimate a non-parametric cost function rather than the typical parametric translog function estimated in earlier literature, while Hughes and Mester (2013) and Hughes et al. (2001) endogenize bank risk and capital structure decisions. The difference in time periods may also play a role (for example, the greater use of information technology may have changed the extent to which scale economies are present).

The theoretical derivation of the cost function assumes that the bank maximizes profits, or equivalently, minimizes costs for any given level of output. A related body of literature on bank efficiency, however, finds evidence of surprisingly large cost differences between otherwise similar banks. These differences are viewed as evidence of X-inefficiencies, that is, firms operating inside their production possibilities frontier because of agency conflicts, management problems or other inefficiencies [DeYoung (1998); Berger et al. (1993); Berger and Humphrey (1991)].

Rather than analyzing total scale economies or X-efficiency, this paper presents disaggregated evidence on the relationship between firm size and detailed components of NIE. We have in mind the idea that operational and technological efficiencies related to size are likely to show up in the data in the form of lower operating costs in areas such as information technology and corporate overhead (for example, accounting and human resources) because large BHCs are able to spread the fixed component of these costs over a broader revenue or asset base. Our goal is to shed additional light on the mechanisms driving differences in efficiency between small and large firms. We note that our empirical finding that large BHCs have lower average operating costs could be driven by the presence of scale economies in the production of banking services, higher average X-efficiency for large firms, or both. For some categories of NIE, it could also be possible that lower costs for larger banking firms not only reflect technological efficiencies, but also greater bargaining power relative to suppliers, customers or employees.

Our analysis is related to recent research by the Clearing House Association (2011) that uses proprietary management information systems data from a number of large banks to estimate product-specific scale curves in seven areas: online bill payment, debit cards, credit cards, wire transfers, automated clearing house, check processing and trade processing.
Do big banks have lower operating costs?

The Clearing House finds that in each of these areas, unit costs are decreasing in production volume, a conclusion that suggests the presence of fixed costs or other technological benefits of size. The economies of scale associated with these seven services are estimated to total U.S.$10 billion to U.S.$25 billion per year.

Although our approach is similar in some respects to the analysis by the Clearing House, we make use of data from audited regulatory filings, rather than internal management information system data, and study components that together sum up to total NIE, rather than just a subset of NIE (the seven items studied by the Clearing House together cover only 7% to 10% of NIE). We also study the entire cross section of BHCs, while the Clearing House sample consists of only six firms.

Our approach is related to the literature on banking mergers that uses accounting variables to estimate the effects of mergers on operating performance. Kwan and Wilcox (2002) find evidence that bank mergers reduced operating costs, although more so for the early 1990s than the late 1980s. Cornett et al. (2006) examine different measures of efficiency improvements for large mergers, and find evidence for cost-efficiency improvements in addition to other revenue improvements. Hannan and Pilloff (2006) show that cost-efficient banks tend to acquire relatively inefficient targets. Using German banking data, Niepmann (2013) finds a negative correlation between size and scaled operating costs—a result consistent with our findings for U.S. firms.

Davies and Tracey (2014) argue that standard estimates of scale economies for large banks are influenced by too-big-to-fail (TBTF) subsidies, and that scale economies are no longer present after controlling for TBTF factors. Hughes and Mester (2013) dispute this conclusion, arguing that the cost function used by Davies and Tracey is misspecified. One potential advantage of our focus on NIE is that operating costs (for example, information technology, printing, postage and advertising) may be relatively more likely to reflect technological features of the firm’s production process than any distortions due to TBTF. Instead, TBTF seems most likely to affect the firm’s funding costs and capital structure. It seems difficult, however, to rule out the possibility that TBTF subsidies may affect our results or those of previous literature.
3. Data and descriptive statistics
Our analysis is based on quarterly FR Y-9C regulatory data filed by U.S. BHCs. The Y-9C filings include detailed balance sheet and income data, as well as information about loan performance, derivatives, off-balance-sheet activities and other aspects of BHC operations. Data are reported on a consolidated basis, incorporating both bank and nonbank subsidiaries controlled by the BHC [see Avraham et al. (2012) for more details]. Our analysis considers only “top-tier” BHCs – that is, the ultimate parent U.S. entity. Our sample includes top-tier U.S. BHCs with a foreign parent, although it excludes “stand-alone” commercial banks that are not owned by a BHC, and BHCs that are too small to file the Y-9C (the Y-9C reporting threshold varies over time, but is currently U.S.$500 million). Our sample excludes investment banks, thrifts and other types of financial institutions, unless those firms are owned by a commercial BHC.

NIE is reported in the consolidated Y-9C income statement (Schedule HI), broken down into five categories. Note that NIE does not include loan losses due to defaults, trading losses, gains and losses on owned securities or taxes; these are recorded in other parts of the income statement. Our analysis focuses on NIE because it is the most likely area in which firms would realize operating cost advantages from size.

We compute several normalized measures of NIE. The first measure, widely used by practitioners and industry analysts, is the “efficiency ratio,” defined as the ratio of NIE to “net operating revenue,” the sum of net interest income and non-interest income:

\[
\text{Efficiency ratio} = \frac{\text{Non-interest expense}}{\text{Net interest income} + \text{Non-interest income}}
\]

---

6 BHC net income in Schedule HI is calculated as follows: net income = net interest income + noninterest income − NIE − provision for loan and lease losses + realized securities gains (losses) − taxes + extraordinary items and other adjustments − net income attributable to noncontrolling interests. See Copeland (2012) for descriptive information on how the main components of BHC income have evolved over time.
A higher efficiency ratio indicates higher expenses, or equivalently, lower efficiency. Effectively, this ratio measures the operating cost incurred to earn each dollar of revenue. Efficiency ratios vary widely across BHCs, as we document below, but typical values range from 50% to 80%. Efficiency ratios are sometimes computed excluding certain noncash items from NIE, such as amortization of intangible assets. We refer to such measures as “cash” efficiency ratios.

One limitation of the efficiency ratio is that it is sensitive to quarter-to-quarter movements in net operating revenue. For example, ratios spiked for many BHCs during the financial crisis, because of trading losses and other non-interest losses. (In rare cases, the efficiency ratio even flips sign, because the sum of net interest and non-interest income is negative.) To provide an alternative normalization that is less sensitive to these concerns, we also present results based on scaling NIE by total assets or RWA, rather than net operating revenue:

\[
\text{Expense asset ratio} = \frac{\text{Non-interest expense}}{\text{Total assets (or risk-weighted assets)}}
\]

These normalizations can be computed for total NIE, or for NIE sub-components such as compensation.

3.1 Descriptive statistics
Table 1 presents descriptive statistics for NIE over the period from first-quarter 2001 to fourth-quarter 2012. We selected this period to take advantage of additional detail on non-interest income expense that was added to the Y-9C in 2001, thereby allowing us to separate non-interest income (which we use as a control) into components such as investment banking fees, income from insurance fees, deposit fees and servicing fees. Note that the sample period for our regression analysis in Section 4 begins in first-quarter 2002 because we incorporate lagged income variables from the previous four quarters. A total of 2,810 BHCs are present in the sample for at least one quarter.
Do big banks have lower operating costs?

Panel A of the table reports summary statistics for four normalized measures of NIE: the efficiency ratio, the cash efficiency ratio (which excludes goodwill impairment and amortization from NIE), NIE scaled by total assets and NIE scaled by RWA. The industry efficiency ratio averages 66.3% over 2001–12, although it is somewhat higher (71.7%) in 2012. The standard efficiency ratio and the cash efficiency ratio differ little on average, reflecting the fact that goodwill impairment and amortization expense generally represent a small total of total NIE.

The distribution of the expense ratios is skewed to the right. For example, the difference between the 5th percentile of the efficiency ratio and its median is 19.5%, significantly smaller than the difference of 28.0% between the median and the 95th percentile value. Furthermore, the right tail includes some extremely high values (for example, the 99.5th percentile is 198.4%), likely driven by one-time spikes in revenue. To reduce the influence of outliers, our regression analysis winsorizes the top and bottom 0.5% of observations for each NIE ratio (all data below and above the bottom and top 0.5th percentiles, respectively, are set equal to the 0.5th and 99.5th percentiles).

We examine the components of NIE in Panel B of the table, based on the five NIE categories reported on Schedule H1.7

- **Compensation** (49.4% of industry total over the sample time period, reported on FR Y-9C as “salaries and employee benefits”) includes wages and salaries, bonus compensation, contributions to social security, retirement plans, health insurance, employee dining rooms and other components of employee compensation.

- **Premises and fixed assets** (11.6% of total, reported on Y-9C as “expenses of premises and fixed assets net of rental income”) includes depreciation, lease payments, repairs, insurance and taxes on premises, equipment, furniture and fixtures. It excludes mortgage interest on corporate real estate.

---

Do big banks have lower operating costs?

- **Goodwill impairment** (1.8% of total, reported on Y-9C as “goodwill impairment losses”) represents losses incurred when goodwill exceeds implied fair value and is revalued downwards. This item is reported separately from “other NIE” from 2002 onward.
- **Amortization expense** (1.9% of total, reported on Y-9C as “amortization expense and impairment losses for other intangible assets”) includes amortization of goodwill and other intangible assets owned by the BHC, as well as impairment losses for intangible assets other than goodwill. This item is also available from 2002 onward.
- **Other** (35.0% of total) includes a broad range of other operating costs, such as telecommunication and information technology costs, legal fees, deposit insurance, advertising, printing, postage and so on. Additional information on these expenses is provided in the memoranda to Schedule HI, as we explain in detail below.

**Table 1: NIE summary statistics**

<table>
<thead>
<tr>
<th></th>
<th>Industry</th>
<th>Individual observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full sample</td>
<td>2012</td>
</tr>
<tr>
<td>Panel A: Efficiency measures, in percentage: 2001-12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency ratio</td>
<td>66.32</td>
<td>71.68</td>
</tr>
<tr>
<td>Cash efficiency ratio</td>
<td>63.29</td>
<td>70.39</td>
</tr>
<tr>
<td>Expense-to-asset ratio</td>
<td>0.82</td>
<td>0.82</td>
</tr>
<tr>
<td>Expense-to-RWA ratio</td>
<td>1.22</td>
<td>1.35</td>
</tr>
<tr>
<td>Panel B: Components of NIE, as a percentage of total: 2001-12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compensation</td>
<td>49.36</td>
<td>48.68</td>
</tr>
<tr>
<td>Premises and fixed assets</td>
<td>11.63</td>
<td>9.64</td>
</tr>
<tr>
<td>Goodwill impairment</td>
<td>1.75</td>
<td>0.02</td>
</tr>
<tr>
<td>Amortization expense</td>
<td>1.93</td>
<td>1.78</td>
</tr>
<tr>
<td>Other</td>
<td>34.95</td>
<td>39.88</td>
</tr>
</tbody>
</table>


Notes: The table reports summary statistics for 2,810 unique BHCs from 2001:Q1 to 2012:Q4, a total of 58,217 firm-quarter observations. The column labeled “industry” reports the average industry efficiency ratio, calculated by summing across all bank holding companies each quarter, taking the ratio, and then taking the time-series mean, either over the 2001:Q1-2012:Q4 sample period or over calendar year 2012. The denotation “p” refers to percentiles of individual observations (for example, “p50” is the median). Variables are defined in Appendix A of the original article. RWA is risk-weighted assets.
Figure 1 plots the time series evolution of the four normalized measures of total industry NIE. Each expense measure declined between 2001 and mid-2007, a period when the revenues and assets of the banking system grew rapidly. For example, the industry efficiency ratio fell from 65.4% in first-quarter 2001 to 58.8% in second-quarter 2007, while the expense asset ratio declined from 0.96% to 0.72% over the same period. This downward trend was reversed during the 2007–09 financial crisis. Since the efficiency ratio is mechanically inversely related to net operating revenue, the reversal for that NIE measure is perhaps unsurprising. However, the expense asset ratio also increased, whether normalized by total assets or RWA. In recent years, NIE ratios have stabilized at levels higher than those prevailing prior to the onset of the crisis. The rise in the efficiency ratio in part simply reflects the decline in net operating revenue and measures of profitability for the banking industry, owing to compression of net interest margins and lower non-interest income.

Appendix B also plots the evolution of the relative shares of the five NIE sub-categories. Goodwill impairment expenses are almost entirely concentrated in 2008, with negligible levels for this expense category before and after 2008. Other NIE makes up a progressively larger fraction of total NIE over the past five years. (In 2012, this category represented 39.9% of total NIE, a share similar to that held by compensation expenses).

As a first look at the relationship between firm size and normalized NIE, the main focus of this paper, we present scatter plots of BHC size and the efficiency ratio (Figure 2). The plots are based on year-to-date 2012 expense data and assets as of the end of 2012. A striking feature of the chart is the variability in NIE across firms, particularly among smaller BHCs. This finding is also borne out in our multivariate analysis in Section 4. The variability points to the importance of adding controls for those observable differences in BHCs’ activities that are associated with different types of expenses. These controls are described in Section 3.3.

---

Do big banks have lower operating costs?

Figure 1: NIE ratios over time

Notes: Income data are quarterly and are not annualized. Ratios are reported in percentages. NIE is NIE; RWA is risk-weighted assets.
Do big banks have lower operating costs?

Figure 2: Scatter plots of operating cost ratios and BHC Size

Notes: Scatter plots are based on average quarterly NIEs over 2012 and total BHC assets as of the end of 2012. BHC is bank holding company.
3.2 Classifying other NIE
The category “other NIE” represents more than one-third of industry NIEs since 2001. To shed light on these costs, we examine data from the memoranda to Schedule HI. Since 2008, Schedule HI has allowed BHCs to classify other NIE into eleven standardized sub-categories; in addition, space is provided for BHCs to report additional “write-in” expense items that were not captured by the standardized fields. For the 11 standardized sub-categories, BHCs are instructed to record items for amounts greater than $25,000 that also exceed 3% of total other NIE.

Write-in items bear the additional requirement that the expense item exceed 10% of total other NIE. Since 2008, amounts in the 11 standardized categories have made up 38% of total other NIE, while the write-in fields have constituted another 28% of other NIE. The remaining 34% of other NIE is not reported in the Schedule HI memoranda, presumably because it does not meet the reporting thresholds described above.

It is particularly challenging to classify and analyze items recorded in the write-in expense fields, because these amounts are reported using non-standardized language by each BHC. For example, NIEs related to foreclosures and to properties that are “other real estate owned” are variously written in as “reo,” “ore,” “R.E.O,” “oreo,” “foreclose,” and so on, as well as various misspelled text strings, such as “oero” and “forclosuer” (sic). Overall, more than 30,000 text strings are written in by the BHCs in our sample between 2008 and 2012. Approximately, 5,500 of these strings are unique. Individual BHCs often tend to use the same text field from one quarter to the next when referring to a given data item, a practice that reduces the total number of fields to be classified.

9 The eleven standardized memoranda categories are (a) data processing expenses, (b) advertising and marketing expenses, (c) directors’ fees, (d) printing, stationery, and supplies, (e) postage, (f) legal fees and expenses, (g) FDIC insurance assessments, (h) accounting and auditing expenses, (i) consulting and advisory expenses, (j) automated teller machine (ATM) and interchange expenses and (k) telecommunications expenses. See FR Y-9C Schedule HI Memorandum Item 7.

10 “Other real estate owned” refers to real estate owned by a bank as a result of the foreclosure of a mortgage loan.
We classify each unique text string into broad categories, proceeding in two steps. First, we classify each string into 1 of 90 sub-categories, such as “card rewards,” “custodian fees,” “affordable/low-income housing,” “servicing,” “dues/ memberships/subscriptions,” and “lockbox fee.” We chose these subcategories by grouping together apparently similar items, employing our institutional knowledge where possible, as well as internet searches and our best judgment. A list of these sub-categories, along with the percentage of non-missing values, is presented in Appendix B of the original paper, not available here. This classification was in part done by hand, and in part via Stata code that conducted Boolean searches for keywords within each text string. The sub-categories include four separate “miscellaneous/other” categories, one for text strings that are well-defined but do not fit into any obvious category (for example, “cattle feed,” “livestock,” and “image processing”), one for items that we did not understand (for example, “tops expense”), one for items that are vague or otherwise unclassifiable (for example, “sundry loss”), and one for text strings that combine multiple items with values listed.

Since most of the sub-categories are fairly sparsely populated we then aggregate them into nine categories that are better suited to statistical analysis. We also assign each of the 11 standardized memoranda items to one of the same nine author-defined categories. By doing this, we are able to classify 66.2% of other NIE into the nine high-level categories, which are listed below:

- **Corporate overhead** (18.6% of other NIE) is intended to measure general corporate expenses and includes four standardized Y-9C items: “accounting and auditing,” “printing, stationery, and supplies,” “postage,” and “advertising and marketing.” It also includes write-in expenses related to corporate overhead costs, such as travel, business development, recruitment, professional memberships and subscriptions, and charitable contributions.

- **Information technology and data processing** (12.6% of other NIE) covers the standardized Y-9C item “data processing expenses,” as well as write-in expenses related to information technology, software, and internet banking.

- **Consulting and advisory** (11.1% of other NIE) is the standardized Y-9C item “consulting and advisory expenses.” It does not include any write-in expenses.
Do big banks have lower operating costs?

Legal (6.7% of other NIE) includes the standardized Y-9C item “legal fees and expenses,” as well as write-in line items related to “litigation,” “settlement,” “records retention,” “legal reserve,” and similar items.\(^\text{11}\)

Retail banking (6.4% of other NIE) is intended to reflect operating costs related to lending and deposit-taking and includes the standardized NIE category “ATM and interchange expenses,” as well as write-in items related to loans, retail banking or credit cards (for example, costs related to real estate owned properties, credit reports, credit card rewards, branch closing costs, lockbox fees, check fraud, and so on).

Federal Deposit Insurance Corporation (FDIC) assessments and other government-related expenses (5.8% of other NIE) includes the standardized Y-9C item “FDIC deposit insurance assessments” and write-in expenses related to the Community Reinvestment Act, compliance with regulation and other items. In practice, deposit insurance fees make up the bulk of these expenses.

Other financial services (3.0% of other NIE) embraces written-in expense items for financial activities other than traditional lending and depository services – in particular, asset management, insurance, and miscellaneous derivatives- and trading-related expenses.

Directors’ fees and other compensation (0.3% of other NIE) includes the standardized Y-9C category “directors’ fees,” as well as write-in fields related to director compensation or other compensation costs.

Miscellaneous (1.8% of other NIE) reflects the four types of miscellaneous categories described above – that is, items that cannot be easily classified or are not understood by us based on the content of the write-in field.

In a small minority of cases, the write-in field content suggests an expense item that may have been classified as other NIE by mistake (for example, costs related to employee compensation). We did not attempt to reclassify these expenses, given the limited context and detail in the write-in fields.

\(^\text{11}\) The standardized “legal fees and expenses” other NIE category includes fees and retainers paid for legal services obtained, but excludes legal settlements and legal expenses associated with owned real estate. Legal settlements and legal reserves established against expected future settlements are recorded in the write-in text fields, if separately reported.
Do big banks have lower operating costs?

Descriptive statistics for these nine author-defined categories of other NIE are presented in Panel B of Table 2. Note that the individual percentiles and standard deviations reported in the table are based on annual expenses, rather than quarterly values. We adopt this approach because of the significant number of zero values reported for even these nine aggregated categories. Our analysis of the other NIE subcategories is based on these year-end cumulative expenses.

The variation across BHCs in the relative size of different components of other NIE is striking. For example, the category “other financial services,” which includes NIE related to insurance and other nonbanking financial services, has a median value of zero, but at the 99.5th percentile, it is 15.9% of total other NIE. This varied distribution of expenses is consistent with the dispersion in products and services offered by BHCs.
Do big banks have lower operating costs?

Table 2: Components of other NIE

Panel A: FR Y-9C classification of other NIE: 2008-12

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage of total other NIE, industry</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Y-9C</td>
<td>37.99</td>
</tr>
<tr>
<td>Text classified</td>
<td>28.21</td>
</tr>
<tr>
<td>Unclassified</td>
<td>33.80</td>
</tr>
<tr>
<td>Total</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Panel B: Components of other NIE, as a percentage of total other NIE: 2008-12

<table>
<thead>
<tr>
<th>Individual observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component (author-defined)</td>
</tr>
<tr>
<td>Industry p0.5 p5 p25 p50 p75 p95 p99.5 Mean Standard deviation</td>
</tr>
<tr>
<td>Corporate overhead 18.63 0.00 2.43 10.29 16.26 22.70 34.58 50.95 17.07 10.07</td>
</tr>
<tr>
<td>Information technology and data processing 12.63 0.00 0.64 8.21 13.84 19.81 29.91 45.01 14.54 8.69</td>
</tr>
<tr>
<td>Consulting and advisory 11.07 0.00 0.00 0.00 2.31 5.78 12.73 29.97 3.74 5.23</td>
</tr>
<tr>
<td>Legal 6.68 0.00 0.00 0.00 3.53 6.19 12.43 24.71 4.16 4.71</td>
</tr>
<tr>
<td>Retail banking 6.35 0.00 0.00 0.00 6.41 13.48 29.64 55.24 9.24 10.55</td>
</tr>
<tr>
<td>FDIC assessments and other government 5.81 0.00 0.00 6.80 11.53 16.95 25.54 37.34 12.26 7.58</td>
</tr>
<tr>
<td>Other financial services 3.01 0.00 0.00 0.00 0.00 4.00 15.85 0.56 2.72</td>
</tr>
<tr>
<td>Directors’ fees and other compensation 0.25 0.00 0.00 0.00 0.00 3.45 6.99 14.60 1.91 2.85</td>
</tr>
<tr>
<td>Miscellaneous 1.76 0.00 0.00 0.00 0.00 0.00 5.75 24.91 0.84 3.98</td>
</tr>
<tr>
<td>Total classified 66.20 4.02 35.11 55.83 66.87 75.05 85.72 95.35 64.32 15.73</td>
</tr>
<tr>
<td>Unclassified 33.80</td>
</tr>
</tbody>
</table>


Notes: The table reports summary statistics for 2,810 unique BHCs from 2008 to 2012. Annual data are as of year-end, for a total of 4,999 firm-year observations. Panel A summarizes information on the following types of NIE: (i) FR Y-9C line items: eleven standardized other NIE items reported in FR Y-9C Schedule HI: Memoranda, (ii) text classified: other NIE items reported in Schedule HI: Memoranda as text fields, and (iii) unclassified: other NIE items not classified in Schedule HI (for example, because the amounts do not exceed the reporting threshold). Panel B includes summary statistics for the nine author-defined other NIE categories, which are constructed from the FR Y-9C line items and the text fields. These data are described in Section 3.2. FDIC is Federal Deposit Insurance Corporation.
Do big banks have lower operating costs?

Table 3: Summary statistics for control variables

<table>
<thead>
<tr>
<th>Asset shares (percentage of total assets)</th>
<th>Industry, by size cohort</th>
<th>Individual observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 1%</td>
<td>95%-99%</td>
<td>75%-95%</td>
</tr>
<tr>
<td>Total loans</td>
<td>42.08</td>
<td>59.58</td>
</tr>
<tr>
<td>Residential real estate loans</td>
<td>14.94</td>
<td>16.63</td>
</tr>
<tr>
<td>Commercial real estate loans</td>
<td>4.26</td>
<td>15.65</td>
</tr>
<tr>
<td>Commercial and industrial loans</td>
<td>12.54</td>
<td>15.65</td>
</tr>
<tr>
<td>Credit card loans</td>
<td>3.53</td>
<td>2.33</td>
</tr>
<tr>
<td>Other consumer loans</td>
<td>4.68</td>
<td>6.11</td>
</tr>
<tr>
<td>All other loans</td>
<td>6.03</td>
<td>6.32</td>
</tr>
<tr>
<td>Trading assets</td>
<td>15.52</td>
<td>1.45</td>
</tr>
<tr>
<td>Federal funds and repurchase agreements</td>
<td>13.67</td>
<td>2.20</td>
</tr>
<tr>
<td>Cash</td>
<td>5.49</td>
<td>5.76</td>
</tr>
<tr>
<td>Investment securities</td>
<td>12.65</td>
<td>20.60</td>
</tr>
<tr>
<td>Other real estate owned</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Fixed assets</td>
<td>0.70</td>
<td>1.24</td>
</tr>
<tr>
<td>Investments in unconsolidated subsidiaries</td>
<td>0.33</td>
<td>0.18</td>
</tr>
<tr>
<td>Investments in real estate ventures</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>Intangible and other assets</td>
<td>8.02</td>
<td>6.77</td>
</tr>
</tbody>
</table>

Risk

| Risk-weighted assets (percentage of total assets) | 63.85 | 75.08 | 71.72 | 72.95 | 71.82 | 67.04 | 71.68 | 11.89 |
| Nonperforming loans (percentage of total loans) | 2.94 | 1.85 | 2.05 | 1.83 | 1.95 | 2.51 | 1.65 | 2.65 |
| HHI income | 0.53 | 0.56 | 0.59 | 0.64 | 0.67 | 0.53 | 0.69 | 0.17 |
Do big banks have lower operating costs?

<table>
<thead>
<tr>
<th>Revenue composition (percentage of net operating revenue)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest income</td>
</tr>
<tr>
<td>Trading income</td>
</tr>
<tr>
<td>Non-interest non-trading income</td>
</tr>
<tr>
<td>Fiduciary income</td>
</tr>
<tr>
<td>Investment banking fees</td>
</tr>
<tr>
<td>Service charges on deposits</td>
</tr>
<tr>
<td>Net servicing fees</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Funding structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deposits/assets (percent)</td>
</tr>
<tr>
<td>Publicly traded (percentage of sample)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Business concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHI assets</td>
</tr>
<tr>
<td>HHI income</td>
</tr>
</tbody>
</table>

Notes: The table reports summary statistics for 2,810 unique BHCs from 2001:Q1 to 2012:Q4, a total of 58,217 firm-quarter observations. The first six columns are industry ratios (computed by first summing numerator and denominator across all firms in the relevant size class), or are statistics weighted by firm size, except for the two indicator variables “publicly traded” and “BHC is foreign-owned.” Size cohorts are recalculated in each quarter. The last two columns are unweighted statistics across all firms. Note that the sample period for the regression analysis begins in 2002:Q1, not 2001:Q1, because specifications include lagged income variables from the previous four quarters. See Appendix A of the original article for variable definitions. HHI is Herfindahl-Hirschman Index; BHC is bank holding company.

3.3 Controls
Operating costs are likely to vary significantly across BHCs engaged in different business activities. While the decision to enter different businesses is endogenous, and may be related to size, we are primarily interested in understanding how size is related to operating expenses on an apples-to-apples basis. For this reason, our regression analysis controls for a variety of BHC characteristics reported in the FR Y-9C. Summary statistics for these controls are presented in Table 3.
In order to show how these controls are related to bank size, we also present industry averages for the following size cohorts: largest 1%, 95% to 99%, 75% to 95%, 50% to 75%, and smallest 50%. Differences in BHC characteristics by size are clear from differences in sample means within the cohorts. However, there is substantial variation in business models apparent within size cohorts as well.

The controls in Table 3 are grouped into six categories, as follows:

- **Asset shares**: our asset composition control variables measure the fraction of balance sheet assets held in various types of loans and other assets (for example, trading assets, securities, cash and fixed assets). As shown in Table 3, small firms hold a higher fraction of total assets in the form of loans, while trading assets are a significantly higher share of total assets for the largest BHCs than for any other group.

- **Risk**: we control for two additional measures of asset risk — RWA as a percentage of total assets and nonperforming loans (NPLs) as a percentage of total loans. The relationship between firm size and risk is non-monotonic for both risk measures, although we note that the largest firms have significantly higher nonperforming loan ratios than other BHCs.

- **Revenue composition**: this variable refers to the percentage of net operating revenue (the sum of interest and non-interest income) that is earned from different sources: (i) interest income, (ii) trading income and (iii) five different components of non-interest nontrading income. Since these components can be volatile, we include these variables in the regressions, in the form of a four-quarter rolling average lagged value. (The standard deviation reported in the table is based on this four-quarter rolling average.) It is notable that large BHCs earn a significantly higher percentage of revenue from non-interest income.

---

12 To compute the industry average for the asset and income ratios, we sum the numerator and denominator of the ratio across all firms in the size cohort, and then take the ratio of the two sums. In contrast, the mean and standard deviation reported in the first two columns represent the unweighted mean and standard deviation of the individual observations in the sample. Of course, the mean of the individual observations may differ substantially from the industry mean if the ratio in question is correlated with firm size.
Funding structure: in some specifications, we include two controls for funding structure, the ratio of deposits to assets and a dummy for whether the BHC is a publicly traded company (firms with foreign parents are coded as private, regardless of whether their ultimate parent is public). Large firms fund less of their assets with deposits, on average.

Business concentration: research in organizational economics has found that diversified firms tend to be less efficient and less profitable than focused firms. In studies that are most relevant to our analysis, Goetz et al. (2013) find that geographically diversified commercial banks have lower valuations, while Laeven and Levine (2007) find a diversification discount (based on the firm's activity mix) in an international cross section of banks. In the spirit of these studies, we include Herfindahl-Hirschman Index (HHI)-style measures of asset and income concentration, computed as the sum of squared asset weights and income weights, respectively, based on the categories presented in Table 3. Higher values of these measures indicate greater concentration. As Table 3 shows, large firms have more diversified assets and activities (lower HHI), reflecting their greater reliance on financial activities outside of traditional lending and deposit-taking.

Organizational complexity: organizationally complex firms may also have higher operating costs, because of various internal inefficiencies (for example, duplication of efforts across different subsidiaries or divisions within the same firm). It is important to attempt to disentangle the effects of size and structure, given that large firms are likely to be organizationally complex. Our analysis includes three measures of organizational structure, the log number of subsidiaries [following Avraham et al. (2012)], the percentage of subsidiaries domiciled overseas and a dummy for whether the BHC has a foreign parent. As shown by the sample means across size cohorts, large firms have more complex organizational structures than small firms on each of these dimensions. The differences are striking: the largest BHCs (those in the top 1% of the size distribution) have 962 subsidiaries on average, 22.7% of which are domiciled overseas. BHCs below the sample median in size, however, have only four subsidiaries on average, and only 4.8% of these subsidiaries are domiciled outside the U.S.
4. Analysis
In this section, we study the relationship between BHC size and measures of total NIE scaled by revenue or assets, examining how this relationship is affected by controlling for differences in firms' business models and by the normalization of NIE used. Our analysis progressively adds controls for a wide range of measures of the composition of BHC assets and sources of income, on the presumption that some types of assets or activities are likely to be more complex and time-consuming to manage than others. For example, a BHC with a large portfolio of other real estate owned assets will likely incur significant property maintenance and management expenses associated with these assets, compared with an otherwise similar banking firm that has liquidated such properties in return for cash, government securities or other simple assets. Similarly, a portfolio of consumer loans is likely to have different screening and monitoring costs than a portfolio of commercial loans. Including these controls seems particularly important given that asset composition varies significantly by firm size, as documented in Section 3.

4.1 Total NIE
We undertook an ordinary least squares estimate of the relationship between the efficiency ratio and BHC size measured by the log of total assets.\textsuperscript{13} We find a statistically and economically significant inverse relationship between size and the efficiency ratio in each regression specification. That is, NIEs per dollar of net operating revenue are lower for large BHCs.

We began our analysis with a regression that controls only for time-series variation in the efficiency ratio, through the inclusion of quarter fixed effects. Each subsequent regression specification successively adds more explanatory variables associated with differences in BHCs' business activities. We begin with simple controls for the composition of BHC assets and add more detailed measures of the risk of those assets, the composition of revenue, funding structure, business concentration, organizational complexity and geography.

\textsuperscript{13} The findings are provided in Table 4 of the original article. The table has been removed here due to space constraints.
Do big banks have lower operating costs?

We find that the inclusion of additional controls tends to steepen the inverse relationship between BHC size and the efficiency ratio. Including controls for BHC asset composition (for example, the percentage of assets in fixed assets, residential real estate loans, trading assets, and so on) increases the magnitude of the coefficient on bank size by 54% (from -1.32 in specification 1 to -1.96 in specification 3), and increases the explanatory power of the model by 13 percentage points. Controlling for the percentage of income generated by different activities (for example, trading, investment banking and deposit service charges) shifts the coefficient to -2.63 (specification 6). The inclusion of controls for organizational complexity further steepens the association between BHC size and the efficiency ratio; the coefficient increases in magnitude from -2.98 to -4.13.

For the model including all controls but excluding firm fixed effects, the coefficient on size of -4.151 implies that a 10% increase in size is associated with a 42 basis point decrease in the efficiency ratio, equivalent to 0.6% of the sample average efficiency ratio. In dollar terms, the coefficient implies that for a BHC at the mean of the data (U.S.$9.1 billion in assets), an increase in size of U.S.$1 billion is associated with a reduction in operating expenses of U.S.$437,000 per quarter, relative to a counterfactual in which the efficiency ratio is not associated with size. The corresponding calculation for the smaller coefficient from column 2 implies a reduction in operating expenses of U.S.$199,000 per quarter. The final specification includes BHC fixed effects, and thus examines only changes in size within bank holding companies. This within-firm analysis includes both size changes from organic growth and size changes from mergers. While still statistically significant, this coefficient is somewhat smaller in magnitude than that which includes all controls but firm fixed effects (-2.47 compared with -4.15). There is some evidence that NIEs after mergers are inflated by one-time merger related costs [Kwan and Wilcox (2002)], which may account for this difference. The standard error of the size coefficient estimate from final specification is much larger than in the other specifications; in other words, the coefficients are estimated with lower power, owing to the smaller residual variation in the efficiency ratio not absorbed or accounted for by the fixed effects and other controls.
As expected, observable differences among BHCs explain a significant fraction of the variation in NIEs. Simple asset controls alone more than double the adjusted R2 of the initial specification. However, even the fixed effects specification in final regression has an R2 of only 54.9%, implying a large amount of residual variation in operating costs. Furthermore, the inclusion of BHC fixed effects nearly doubles the R2 relative to the specification that excludes firm fixed effects but includes all other controls, a result suggestive of large persistent differences in operating costs across observably similar firms. This finding seems consistent with prior literature on X-inefficiency, which shows that many banking firms operate significantly inside the efficient production frontier [see, for example, Berger et al. (1993)]. It is worth noting that BHC size alone explains only a very small fraction (less than 1%) of the total variation in NIE in the data, as illustrated graphically in Figure 2.

In sum, we find consistent evidence that large BHCs have lower operating costs as measured by the efficiency ratio, although the strength of the relationship is sensitive to the set of controls used. Instead of taking a strong stance on the “appropriate” set of controls, throughout the paper we present results for specifications using controls from 3 of the regressions used in the original table, namely columns 1, 2, and 10 from Table 4 in the original article. A comparison of the results across these specifications enables the reader to observe how the relationship between NIEs and size is influenced by the inclusion or exclusion of controls for the mix of BHC assets and business activities.

Although our main focus is on the relationship between operating costs and firm size, estimates for several of the controls examined are also of independent interest. In particular, BHC organizational complexity, measured by the log number of subsidiaries, is associated with higher NIE ratios. BHCs with a foreign parent also have higher expenses. Proxies for greater organizational focus are associated with lower NIE: BHCs that have more concentrated asset portfolios and more concentrated sources of non-interest income have lower expenses, all else equal, although the marginal explanatory power of additional concentration is relatively low. Each of these relationships is robust to the inclusion of BHC fixed effects. Although not presented here, these relationships are also robust to specification changes that allow for a more flexible linkage between size and the efficiency ratio. This finding suggests that our results are not likely to be driven only by the largest BHCs.
Caution should be exercised in applying a causal interpretation to these associations, given that we do not have a convincing econometric instrument for organizational complexity or focus. But taken at face value, each of these estimates implies that complex, diversified firms have higher operating expenses than focused or organizationally simple firms, consistent with the conclusions of prior literature on the diversification discount in banking [Goetz et al. (2013); Laeven and Levine (2007)].

4.2 Other functional forms
The specifications so far assume a log-linear relationship between BHC size and the efficiency ratio. Next, we allow for a more flexible functional form by estimating fractional polynomial specifications that permit the data to determine the shape of the relationship between size and the NIE ratio. An alternative to regular polynomials, fractional polynomials provide flexible parameterization for continuous variables. We use the Stata function fracpoly to determine an optimal polynomial specification (optimal polynomial) and also estimate a specification with exponents ranging from -2 to 2 – that is, log assets raised to the -2, -1, 0, 1, and 2 power (flex polynomial). These best-fit polynomials are shown in Figure 3 along with the ordinary least squares line of best fit.

Overall, the log-linear functional form assumed in the analysis appears to be a good approximation, although we note that, based on point estimates, the point-estimated relationship between log assets and the efficiency ratio is somewhat concave at the tails. Specifically, the relationship between BHC size and the NIE ratio is relatively flat among small BHCs (those with assets below U.S.$150 million), while the relationship is steeper among the largest BHCs (those with assets above U.S.$750 billion). For the vast range of asset sizes, the relationship between log size and efficiency ratio is close to linear, and the 95% confidence interval of the alternative forms is very similar. Thus, we use a log-linear specification for the remainder of the analysis.

In addition to investigating flexible polynomial specifications, we separate the sample into different size cohorts, re-sorted in each quarter, and estimate separate specifications for each cohort. This approach allows the relationship between NIE and control variables, as well as size, to vary by BHC size class. (In the fractional polynomial approach, the coefficients on explanatory variables other than size are unrelated to size.)
Do big banks have lower operating costs?

**Figure 3: Efficiency ratio and BHC size, flexible functional forms**

Source: Authors’ calculations, based on statistical analysis of FR Y-9C data.
Note: Functional forms are partial predictions based on varying log of assets ($000s), holding other covariates fixed at their sample means. The efficiency ratio is normalized to be equal to zero for a BHC with $10 billion in assets.

Each column of Table 4 represents specifications 1, 2, and 10 of Table 4 in the original article estimated on a subset of the BHCs sorted by size in each year. The first column replicates the results on the entire sample, for comparison. Without including controls for BHC asset mix, it appears that much of the coefficient on size is driven by BHCs below the median asset size (column 6).
Do big banks have lower operating costs?

Table 4: Coefficient on log assets, by size cohort

<table>
<thead>
<tr>
<th>Specification (1)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Top 1%</td>
<td>95%-99%</td>
<td>75%-95%</td>
<td>50%-75%</td>
<td>Bottom 50%</td>
</tr>
<tr>
<td></td>
<td>-1.320^a</td>
<td>1.860</td>
<td>1.273</td>
<td>-1.790^b</td>
<td>-0.768</td>
<td>-6.140^b</td>
</tr>
<tr>
<td></td>
<td>(0.235)</td>
<td>(1.647)</td>
<td>(1.164)</td>
<td>(0.687)</td>
<td>(1.509)</td>
<td>(1.633)</td>
</tr>
<tr>
<td></td>
<td>-1.892^a</td>
<td>-2.864</td>
<td>-0.379</td>
<td>-1.888^b</td>
<td>-1.914</td>
<td>-3.195^c</td>
</tr>
<tr>
<td></td>
<td>(0.228)</td>
<td>(2.020)</td>
<td>(1.278)</td>
<td>(0.674)</td>
<td>(1.352)</td>
<td>(1.334)</td>
</tr>
<tr>
<td></td>
<td>-4.151^a</td>
<td>-8.018^c</td>
<td>-5.138^a</td>
<td>-4.132^a</td>
<td>-4.238^a</td>
<td>-5.055^a</td>
</tr>
<tr>
<td></td>
<td>(0.326)</td>
<td>(3.931)</td>
<td>(1.442)</td>
<td>(0.696)</td>
<td>(1.204)</td>
<td>(1.311)</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>58,217</td>
<td>604</td>
<td>2,405</td>
<td>12,197</td>
<td>15,181</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations.
Notes: The table presents an analysis of the relationship between size, measured by the log of total assets (lagged by one quarter), and efficiency ratio, defined as total NIE as a percentage of net operating revenue. Each row represents the coefficient on size for specifications (1), (2), and (10) of Table 4 in the original article, estimated on a subset of bank holding companies sorted by size in each quarter. Specification (1) includes time fixed effects. Specification (2) includes time fixed effects as well as controls for the percentage of assets in each broad category (asset shares). Specification (10) includes the controls from specification (2) as well as controls for types of loans, revenue composition, funding structure, business concentration, organizational complexity and headquarters state fixed effects. Robust standard errors reported in parentheses are clustered by BHC and quarter. a: p<0.01; b: p<0.05 and c: p<0.1

As additional controls are included, economies of scale become apparent in many of the size cohorts. In the specification including all controls, the estimated coefficient on size is negative in all cohorts and statistically significant. As suggested by the flexible polynomial specifications, the point estimate coefficient on size is largest in the top 1% of the sample. What do these findings imply for the policy debate around size limits for the largest BHCs? We find no evidence that the inverse relationship between size and operating costs disappears above any particular size threshold; indeed, our point estimates suggest that, if anything, the relationship is steeper among the largest firms. This result is consistent with scale economies from sources other than bargaining power to the extent that we believe that differences in bargaining power may be small within the top 1% of BHCs. The statistical precision of our estimates is limited, however, given the small number of observations for the largest BHCs.
Do big banks have lower operating costs?

### Table 5: Alternative measures of operating costs

<table>
<thead>
<tr>
<th></th>
<th>NIE/risk-weighted assets</th>
<th>NIE/assets</th>
<th>Cash NIE/net revenue (cash efficiency ratio)</th>
<th>Log NIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original specification:</td>
<td>-1</td>
<td>-2</td>
<td>-10</td>
<td>-1</td>
</tr>
<tr>
<td>Log assets</td>
<td>-0.044*</td>
<td>-0.115*</td>
<td>-0.083*</td>
<td>-1.686*</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.0013)</td>
<td>(0.006)</td>
<td>(0.231)</td>
</tr>
<tr>
<td>Asset share controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>All controls</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>R²</td>
<td>0.016</td>
<td>0.231</td>
<td>0.487</td>
<td>0.968</td>
</tr>
<tr>
<td>N</td>
<td>58,217</td>
<td>58,217</td>
<td>58,217</td>
<td>58,192</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations.

Note: The table presents an analysis of the relationship between size, measured by the log of total assets (lagged by one quarter), and different measures of efficiency. The dependent variables in the first three specifications are cash efficiency ratio, defined as total NIE less goodwill impairment and amortization expense over net operating revenue; in the next three specifications, NIE/assets ratio, defined as total NIE over total assets; and in the final three specifications, NIE/RWA ratio, defined as total NIE over total RWA. For each alternative measure of efficiency ratio, specifications (1), (2) and (10) of Table 4 in the original article are presented. Specification (1) includes controls for quarter fixed effects. Specification (2) includes the controls from specification (1) as well as controls for the percentage of assets in each broad category. Specification (10) includes the controls from specification (2) as well as controls for types of loans, revenue composition, funding structure, business concentration, organizational complexity, and headquarters state fixed effects. Models are estimated with robust standard errors and two-way clustering by firm and quarter. a: p<0.01; b: p<0.05 and c: p<0.1

### 4.3 Alternative measures of operating costs

The efficiency ratio may be distorted in periods when net operating income is temporarily low. Next, we test the sensitivity of our results to other normalizations of NIE: the expense asset ratio discussed in Section 3 (NIE/total assets), NIE/RWA, and a “cash” efficiency ratio, which excludes noncash expenses such as goodwill amortization in the numerator.

---

14 During the 2007-08 financial crisis, trading losses and other losses brought net operating income close to zero for several large BHCs.
Do big banks have lower operating costs?

We do this because noncash expenses are often associated with one-time costs relating to mergers and acquisitions that are not likely to persist, and may be associated with size. We also estimate a specification using the log of NIE as an alternative measure of operating costs.

As before, for each normalization of NIE, we re-estimate specifications with the set of right-hand-side variables from columns 1, 2, and 10 of Table 4 in the original article and present the coefficient on asset size. Results are presented in Table 5. Regardless of the normalization used, the coefficient on size is negative and statistically significant once BHC controls are included. In the specification including all controls, the estimated coefficient on size is approximately 7% to 10% of the average expense ratio.

For the specifications using the log of NIE as the dependent variable, the coefficient on log assets can be directly interpreted as the elasticity of operating costs with respect to size. In line with our other results, this elasticity is less than unity – in other words, a 10% change in BHC size is associated with a less than 10% change in NIE operating costs, a finding consistent with the presence of scale economies in operating costs. For the specification including all controls, the operating cost elasticity is 0.899, much smaller than one, although it is significantly closer to one for the specification just including asset controls (0.979). Both these estimates are statistically significantly smaller than unity.

5. Decomposition of NIE
This section examines the relationship between BHC size and components of NIE. First, we consider the five major components of NIE reported in the Y-9C income statement. Probing more deeply, we then analyze the nine sub-components of “other NIE,” using our manual classification of these expenses as described in Section 3.

One goal of this disaggregated analysis is to shed additional light on the sources of the lower operating costs enjoyed by large BHCs. Although these lower costs could be due to scale economies or other efficiency benefits of size, they could also reflect implicit government guarantees for large BHCs, or the greater bargaining power of these firms. For example, large banks may endogenously select riskier activities, but invest less in risk management because of implicit insurance associated with being “too big to fail.”
Do big banks have lower operating costs?

Alternatively, large banks may simply take advantage of greater bargaining power to reduce expenses. These different explanations have very different normative welfare implications. Efficiency benefits of size imply that limiting size would impose deadweight economic costs, while explanations relating to bargaining power and TBTF primarily relate to the allocation of economic rents. Although the breakdown of expenses in the Y-9C does not allow us to fully disentangle these different explanations, we are able to draw some suggestive conclusions.

5.1 Major components of NIE

We begin by studying the five expense categories reported on Schedule HI: compensation (49.4% of NIE), premises and fixed assets expense (11.6%), goodwill impairment (1.8%), amortization (1.9%) and other (35.0%).\textsuperscript{15} As before, we normalize each expense by net operating revenue, and for parsimony, focus on the coefficient on log assets for specifications 1, 2, and 10 from Table 4 in the original article.

Each of the three largest categories of NIE declines as a percentage of net revenue as size increases, all else equal, with or without the inclusion of controls for BHC characteristics. Focusing on the specifications including these controls (either for asset composition alone or for all controls), we find that the inverse relationship between BHC size and scaled NIE is steepest for compensation, followed by other NIE, based on this calculated elasticity. For the specifications including all controls, a 10% increase in size is associated with a 0.735% decline in compensation scaled by net operating revenue and a 0.683% decline in the corresponding ratio for other NIE. The result for employee compensation is perhaps surprising, given that large BHCs have more employees in highly compensated roles such as investment banking and trading. However, the higher productivity and additional revenue earned by these employees (the denominator of the efficiency ratio) appears to offset this higher compensation.

Expenses related to premises and fixed assets may represent a category of operating costs for which scale efficiencies are lower (for example, building lease costs are roughly proportionate to the size of the leased space, at least within a specific geographic area).

\textsuperscript{15} The results are presented in Table 7 of the original article. The table has been removed here due to space constraints.
Given this, it is perhaps unsurprising that estimated economies of scale are smaller for premises and fixed assets expense: for this category, our point estimate implies that a 10% increase in size is associated with a 0.478% decline in expenses scaled by operating revenue.

Significantly, expenses related to the impairment and amortization of goodwill and other intangible assets are actually proportionately higher for large firms – a fact that distinguishes these expenses from the other categories. We estimate a positive, statistically significant (in most specifications) coefficient on these expenses. The likely key reason for this finding is that large BHCs often have grown by way of acquisitions, which will sometimes result in goodwill when the acquisition purchase price exceeds the tangible book value of assets purchased. Consequently, these firms report higher expenses related to the amortization or impairment of these assets. Although the positive slope for these two expense categories is economically significant, the two categories together make up only a relatively small proportion (3.7%) of total industry NIE.

5.2 Sub-components of other NIE
In this section, we examine the nine sub-components of “other NIE” identified in Section 3.2. (Recall that these categories reflect both standardized memoranda items reported on the Y-9C since 2008 and “write-in” text strings classified by us.) Previous work estimating scale curves for these disaggregated categories has been based on case studies or has had limited sample size (for example, Clearing House Association [2012]).

Overall, we find evidence that scaled expense falls with size for most, but not all, components of other NIE, especially after including controls for BHC asset and income composition. When controls for the composition of assets and income sources are included in the specification, large BHCs exhibit lower expenses in categories in which a fixed cost can be spread across an expanded scale of operations, such as corporate overhead, information technology and data processing.
When we look at the results for the other NIE components, we find that corporate overhead is the largest component of other NIE, and a component for which we estimate significant scale efficiencies (a high estimated coefficient on size relative to mean level of expense). Corporate overhead includes expenses, such as accounting and auditing, advertising and marketing, treasury expenses, travel and business development, charitable donations, insurance and utilities. These expenses appear to have significant operational leverage; the estimated coefficient on size is −0.33, approximately 7% of the mean level of corporate overhead expenses. Similar scale economies are observed for expenses associated with information technology and data processing, with an estimated coefficient on size that is −6.6% of mean IT expense. This finding is consistent with the view that spreading overhead expenses associated with technology may be one source of cost advantage for large banking firms.

In contrast to these two categories, we find that expenses associated with consulting and advisory services are proportionately higher for large BHCs. Prior to adding controls for BHC characteristics, our estimates show that the coefficient on size and consulting expenses is positive and statistically significant. This coefficient remains significant when asset composition controls are included, although once all controls are included the coefficient is positive but no longer statistically significant. This suggests that consulting and advisory services may be related to non-interest income, rather than to the composition of BHC assets. Despite recent publicity surrounding large BHCs’ legal issues and large-dollar-value settlements, over the 2008–12 period, legal expenses also increase less than proportionately with BHC size, particularly in the specification including the full set of controls [specification (10)]. This expense category includes both legal fees and retainers paid for legal services performed, as well as expenses associated with legal settlements and reserves, to the extent we can identify these expenses from the write-in text fields. Some part of this finding may reflect the fact that small banks may lack internal legal departments, for which expenses would be recorded as part of compensation, and thus have higher external legal fees.

The assignment of write-in fields to retail banking requires perhaps the most judgment on our part. This category includes collection expenses, credit reports, mortgage-related expenses such as appraisal and title fees, branch expenses, checks, lockboxes and robbery, among many others.
After including asset composition controls, the estimated coefficient remains negative although not statistically significant. This result may reflect the wide variation in the types of retail banking businesses that are not well captured by our BHC characteristics. Alternatively, economies of scale may be limited or not present for branch banking (at least among the set of expenses classified in this category), since many costs only scale until the next branch is opened.

Similarly, we find a negative but statistically insignificant relationship between size and normalized FDIC assessments and other government-related expenses after including the full complement of BHC characteristics. The majority of the expenses in this line item are due to deposit insurance, and thus it would be surprising to uncover economies of scale once we control for the amount of deposit financing. This coefficient would likely shrink further if our regression specification included a control for the fraction of insured deposits, rather than total deposits.

The category “other financial services” represents the sum of expenses associated with BHCs’ non-banking businesses, such as asset management, trust and custody services, and insurance. Given the likely differences in the NIEs of these businesses, it is not surprising that the estimated coefficient changes sign from positive to negative once we control for the composition of BHCs’ assets and non-interest income. Banking firms that earn a high percentage of income from fee income should naturally have higher expenses. But holding all else equal and controlling for income composition, we find that larger BHCs have lower scaled expenses in this category: we estimate a coefficient of 7.4% of the mean value. This result is consistent with cost economies of scale in non-compensation expenses associated with businesses such as insurance and asset management.

The component of other NIE for which scale economies are largest in percentage terms is directors’ fees and other compensation. For this category, the coefficient on size is almost three times as large as the sample mean. This makes intuitive sense; even though directors of large BHCs have higher compensation, board size does not increase dramatically with firm size. This coefficient is negative and significant regardless of the set of controls used.
Do big banks have lower operating costs?

Miscellaneous expenses include items as varied as expenditures for cattle feed and reducing gold to market. It also includes nonspecific write-in text fields such as “miscellaneous expense,” “miscellaneous fee” and “other expense.” Regardless of the controls for bank businesses used, we do not see economies of scale in these varied expenses, although some economies may exist in the residual category “other expenses,” which includes all NIEs not otherwise classified.

6. Conclusion
We find a robust inverse relationship between the size of BHCs and scaled measures of operating costs. Quantitatively, a 10% increase in assets is associated with a 0.3% to 0.6% decline in NIE scaled by income or assets, depending on the specification. In dollar terms, our estimates imply that for a BHC of mean size in our sample, an additional U.S.$1 billion in assets reduces NIE by U.S.$1 million to U.S.$2 million per year, relative to a base case where operating cost ratios are unrelated to size. This inverse relationship is robust to various changes in model specification, although the magnitude of the relationship is sensitive to the set of controls used.

Unpacking our results, we find that while size is associated with lower scaled operating costs for most components of NIE, the largest contributions in dollar terms come from employee compensation, premises and fixed assets, corporate overhead, and information technology and data processing. While not a large component of total NIE, directors’ fees and other compensation account for the largest proportionate savings, presumably a reflection of the fact that corporate boards do not expand with firm size, even if their members are better paid on average.

Our results likely reflect a combination of three factors: first, large BHCs benefit from “operational leverage” or economies of scale, whereby they effectively spread costs over a higher revenue or asset base. Second, “X-efficiency” – a factor closely related to operational leverage – may be higher for large BHCs; that is, these firms may operate closer to the production frontier on average. Third, large BHCs may have greater bargaining power than smaller firms with suppliers or employees. We are not able to pin down with confidence the relative contribution of these three factors.
Do big banks have lower operating costs?

We emphasize, however, that the inverse relationship between BHC size and scaled measures of NIE is not limited to particular components of expense or particular segments of the BHC size distribution.

Consistent with recent research that identifies the presence of scale economies in banking, our results suggest that imposing size limits on banking firms would be likely to involve real economic costs. Although the limitations of our econometric methodology must be borne in mind, a back-of-the-envelope calculation applied to our estimates implies that limiting BHC size to be no larger than 4% of GDP would increase total NIE by U.S.$2 billion to U.S.$4 billion per quarter. These costs should be weighed against the potential benefits of size limits as policymakers address the “too-big-to-fail” problem.
Do big banks have lower operating costs?

References
Benston, G., 1972, “Economies of scale of financial institutions,” Journal of Money, Credit, and Banking 4, 312-41
Do big banks have lower operating costs?


Niepmann, F., 2013, “Banking across borders with heterogeneous banks,” Federal Reserve Bank of New York Staff Reports, no. 609, April


The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

Arun Muralidhar
Adjunct Professor of Finance, George Washington University, Founder, Mcube Investment Technologies and AlphaEngine Global Investment Solutions

Kazuhiko Ohashi
Professor of Finance, Graduate School of International Corporate Strategy (ICS), Hitotsubashi University, Tokyo, Japan

Sunghwan Shin
Professor of Finance, Hongik University, Seoul, Korea

Abstract
The Capital Asset Pricing Model (CAPM) has been the backbone of asset market finance, even though many academic studies have revealed its limitations, both theoretical and empirical. This paper argues that including liability or benchmark considerations in investment decisions may provide a credible explanation for some of the limitations. In effect, the CAPM is an “absolute wealth”-centric model, driven by the assumption that investors derive utility from absolute wealth. In reality, investors first make investment decisions to ensure sufficient assets to meet obligations, such as future pension payment or future consumption (i.e., liabilities), and focus on relative wealth. Thereafter, many investors hire external managers and evaluate them on performance relative to assigned benchmarks. Interestingly, existing robust literature on, and product offerings for, “liability driven investing,” have not led to a shift in academic theories of asset pricing to reflect the liability perspective. After all, if asset owners choose investments to service liabilities (proxied by benchmarks), it is expected that liabilities or these benchmark proxies will impact asset prices.

1 These are the personal views of the authors and do not reflect the views of any of the organizations to which they belong. We would like to thank Lester Seigel for his unique insight on RAPM and to Prof. Michael Brennan for his feedback on a related paper. Thanks also to Robert Savickas, Mehmet Gerz, Jean Luc Vila, Robert Merton, John Cochrane, John Campbell, Claude Erb, Dimitri Vayanos, Yegin Chen and Alexandre Baptista for their comments on this general idea. All errors are ours.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

Therefore, a shift from an “absolute wealth”-centric to a “liability-centric” or “relative wealth” perspective will create a Relative Asset Pricing Model (RAPM), with the CAPM as a very specialized case of this model. Hence, by examining how liabilities are proxied in practice, the many tests that found the need for additional factors to explain asset behavior might now be rationalized and explained. In this vein, even some behavioral criticisms leveled at CAPM can fit under this liability-centric CAPM-like paradigm. Finally, the asset allocation and portfolio rebalancing implications of this model may better explain some of the risky asset return characteristics such as momentum observed in the market. Therefore, if future research is directed to this compelling liability-centric facet embedded in practitioner behavior and explores this new dimension, RAPM may complement traditional CAPM, and help investors and academicians alike to develop better approaches to asset allocation, rebalancing as well as to asset pricing.

Introduction
A recent spat between Dempsey (2013), who argues that the Capital Asset Pricing Model (CAPM) is largely dead and incorrect as it does not test well, hence should not be used by investors, and Partington (2013) ², who suggests no better model exists and people use it for all its flaws, hence CAPM reigns supreme, encapsulates the two extreme factions of the past 50 years in academic asset market finance. The lack of consensus on the appropriate model to price and allocate assets poses problems for global chief investment officers (CIOs) who have to invest trillions of dollars and also complicates what we teach the next generation of finance professionals [Shojai and Feiger (2009)]. If investment products and advice are not rooted in robust theory, it also complicates the development and marketing efforts of executives at investment consulting firms, asset management companies, and investment banks.

² Levy (2011) makes a similar argument.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

Could the two views represented by Dempsey (2013) and Partington (2013) be consistent with each other? Could it be that CAPM proponents, and those who propose alternatives – like Black (1972), who argues for a Zero-Beta CAPM, albeit without clarity on the exact specifics of the zero-beta portfolio; Fama-French (2004) or Chen et al. (1986), who argue that additional factors beyond “market” beta are needed to explain asset returns, minus a model to explain what the factors might be;3 or even Kahneman-Tversky (1979), who argue that Modern Portfolio Theory (MPT) is wrong because it lacks a reference point and investors appear to exhibit loss aversion and utility from changes in wealth, yet do not articulate a clear reference point – provide meaningful and reconcilable insights through their models? The simple resolution may lie in developing a theory that is rooted in how institutions and individuals make investment decisions; namely, wealth is valuable not for its own sake but in helping us meet some liability or obligation.4 This paper makes the very bold claim that the seemingly contradictory theories could be consistent with each other by shifting the view of asset pricing and asset allocation, from an “absolute wealth”-centric view to a “liability” or “benchmark”-centric view, focusing on “relative wealth.” The reconciliation we provide in this paper may explain empirical anomalies, in turn justifying specific investment approaches.

The paper is structured as follows: Section 1 makes the case for the “liability” or “benchmark”-centric approach based on existing investment practice, and proceeds to present a simple Relative Asset Pricing Model (RAPM) to show the impact of this liability-centric view of asset pricing. It very summarily discusses how this approach may help provide a basis for Black’s Zero-Beta CAPM. Section 2 presents two possible variants to incorporating liabilities into such a model, and for simplicity this paper focuses on funded status (asset-liability ratio) rather than surplus maximization (assets minus liabilities), though the latter provides more tractable models and has been the staple of previous research. Section 3 ties this work to the key critique of Prospect Theory that MPT is missing a reference point and shows how the liability is a relevant reference point in investments.

---

3 We only reference Fama-French, but clearly this encompasses the many factors that Dempsey (2013) highlights have been found to be academically relevant.

4 Cochrane and Culp (2003) make a similar statement; namely, “[I]nvestors do not value money directly. The theory adopts a more sophisticated approach and recognizes that the pleasure or ‘utility’ of consumption that money can buy is really what matters.”
Section 4 demonstrates how RAPM provides a theoretical perspective to Fama–French findings, but also potentially other liability-related factors. Section 5 briefly discusses the flexibility of this generic model in accommodating other relative models of agency, home/firm bias investing or background risk, and Section 6 highlights the behavior of liability-aware investors to demonstrate that these investors adopt Three Fund Separation, forecasted by such a model. This section also discusses some opportunities for CIOs of funds to improve investment processes, and CEOs of financial institutions to create new products. Section 7 concludes by suggesting how this approach provides interesting and useful avenues for future research.

1. The Relative Asset Pricing Model – the role of liabilities

Most finance professionals are familiar with the basic asset pricing formula from CAPM, whereby the expected return on an asset depends on its relationship to a “market” portfolio of all risky assets, with the return defined as \( r(M) \), as shown in equation 1, and where the “beta” is defined in equation (2) – calculated as the covariance of the returns of an asset to the returns of the market portfolio divided by the variance of the returns of the market portfolio. The variable \( F \) is a risk-free asset, with zero volatility and zero correlation to other assets. \( E[\cdot] \) is the expectations operator. One of the key assumptions of this model is that investors maximize their utility, which is an increasing function of expected return, and decreasing function of return variance.

\[
E[r(i)] - r(F) = \beta_i \cdot E[r(M) - r(F)] \quad (1)
\]

\[
\beta_i = \frac{\text{cov}[r(M), r(i)]}{\text{var}[r(M)]} \quad (2)
\]

---

5 There is one additional aspect to the relative paradigm; namely, principal investors delegate to agents whom they hope are skillful. We will not address this issue here, but Muralidhar and Shin (2013) propose a simple model to show how this aspect impacts the demand for various assets.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

One very appealing and potentially useful recommendation derived from this model is how assets should be allocated. In a simple one period world, all investors allocate their assets between the risk-free asset and the market portfolio, with the only differences reflecting the investors' risk tolerance — a process termed Two Fund Separation. Another useful, and relevant in practice, outgrowth of this model is that one can derive measures of risk-adjusted performance with the Sharpe ratio being the most famous measure. Most other models do not embody all of these three useful facets (i.e., recommendations for asset pricing, asset allocation and risk-adjusted performance) for investors.

However, most institutional investors are not driven solely by the need to grow assets for the sake of increasing wealth, but make investment decisions to service specific liabilities or spending objectives. In a survey of 32 global CIOs of a range of institutional funds, 62.5% stated that maximizing funded status was their primary objective (with additionally one response for “Fulfill obligations,” another for “Provide for as much as possible inflation adjusted pensions” and one for “Minimize shortfall risk,” which one could argue are related concepts). For example, for pension funds the liability is the promise of a future pension; for insurance companies, it is future claims that need to be paid out if an event occurs; for endowments and foundations, it is faculty salaries, charitable works/scholarships or faculty research they seek to fund [Merton (1993)]. Even retail investors save and invest to achieve some future spending objective, that could broadly be termed a liability [Merton (2014)], or it could be segmented into multiple sub-liabilities if the investment is for different needs.

---

6 This result is from Tobin (1958). Merton (1973) offers an inter-temporal model with a bequest motive and results in Three Fund Separation with the third portfolio hedging the dynamic opportunity set.
7 Sharpe (1994). Interestingly, Modigliani-Modigliani (1997) provides a related measure which is more appealing in practice because it shows the tie between risk-adjusted performance and portfolio structuring.
8 Some of the earliest research on this specific topic of investing relative to liabilities can be traced to Leibowitz (1986) and Sharpe-Tint (1990).
9 Muralidhar (2014).
10 Das et al. (2010) refer to these as mental accounts. Janssen-Kramer and Boender (2013) demonstrate how these various goals of individuals (for expenses, one-time cash flows, capital targets or conditional cash flows) can be converted into cash flow projections, thereby allowing researchers to apply “private asset-liability management” techniques to determine optimal investment strategies. Step one in the process is converting goals into projected liabilities.
When a regulator imposes strict rules on what an entity must do to recognize and mark-to-market liabilities (even specify corrective actions, depending on whether they are funded or underfunded), these liabilities are explicitly proxied, modeled, measured and monitored.\footnote{Muralidhar and van Stuijvenberg (2005) demonstrate how Dutch pension liabilities can be proxied with a range of fixed-income swaps, but the concept is generic. Leibowitz (1986) shows how pension liabilities can be proxied with bonds. Merton (2014) suggests that deferred annuities would be the liability proxy for retail investors in defined contribution plans.} There are plenty of research papers and a whole industry solely focused on Liability Driven Investing (LDI), and this approach is a mainstream for a large class of investors.

For the rest of the investment community, the liability is implicit, and effectively nothing more than the Strategic Asset Allocation (SAA) to which to benchmark the portfolio (often a different proxy from the true liability). Depending on the type of regulation and investor, these explicit or implicit liabilities may be largely fixed-income instruments (pension funds or insurance companies with explicit liabilities) or some mix of stocks, bonds, commodities and illiquid assets (endowments, U.S. public pension plans, sovereign wealth funds, and retail investors with implicit liabilities). Through a unique accounting ruling, U.S. public pension funds have their expected target rate of return, say 7.5% per annum, as their effective but implicit liability, which is then translated back to an SAA of multiple assets;\footnote{There is an active debate in the U.S. about whether this is the right way to discount liabilities, as U.S. corporate pension funds are required to discount liabilities with a long-duration bond index with credit risk, and we discuss the implications of any changes to this approach in Section 8.} retail investors behave much the same.\footnote{See for example how www.wealthfront.com establishes an investment policy for a retail client.} Therefore, all these approaches can be seen as nothing more than creating a liability proxy for investors, against which investment performance is measured. Regardless of how these liabilities are treated by regulators or investors, if one takes a generic CAPM-type model and adjusts for the fact that investors derive utility from wealth relative to liabilities, it is possible to offer a “linear” model of the form suggested in equation (3A) and (3B), which we refer to as the RAPM.\footnote{Muralidhar et al. (2014a, b). The model is derived following Reisman-Lauterbach (2004) and replacing peer benchmarks with the liability portfolio. The model is also derived from a consumption CAPM approach.}

The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

---

\footnote{Muralidhar and van Stuijvenberg (2005) demonstrate how Dutch pension liabilities can be proxied with a range of fixed-income swaps, but the concept is generic. Leibowitz (1986) shows how pension liabilities can be proxied with bonds. Merton (2014) suggests that deferred annuities would be the liability proxy for retail investors in defined contribution plans.}

\footnote{There is an active debate in the U.S. about whether this is the right way to discount liabilities, as U.S. corporate pension funds are required to discount liabilities with a long-duration bond index with credit risk, and we discuss the implications of any changes to this approach in Section 8.}

\footnote{See for example how www.wealthfront.com establishes an investment policy for a retail client.}

\footnote{Muralidhar et al. (2014a, b). The model is derived following Reisman-Lauterbach (2004) and replacing peer benchmarks with the liability portfolio. The model is also derived from a consumption CAPM approach.}
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

In this model, L refers to the global Liability replicating proxy portfolio (with returns r(L) and non-zero volatility) – which would be a long-duration bond portfolio (with credit) if all investors behaved like U.S. corporate pension funds, or a mix of stocks, bonds, commodities and alternative assets (or an aggregate of all notional SAAs) if all investors behaved like U.S. public pension plans, endowments and foundations.

Define M* as the market portfolio of all assets including risky and risk-free assets, M as the market portfolio of all risky assets (as in MPT), F to be the traditional risk-free asset, and w to be the market value weight of M in M*, i.e., r(M*)=w*r(M)+(1-w)*r(F). Then (3A) can be restated as in (3B) below to make it compatible with CAPM.

\[ E[r(i)] - r(F) = \frac{\text{cov}[r(M^*), r(i)] - \text{cov}[r(L), r(i)]}{\text{var}[r(M^*) - r(L)]} \times E[r(M^*) - r(L)] = \Gamma_{M\cdot i} \cdot E[r(M^*) - r(L)], \quad (3A) \]

where \( \Gamma_{M\cdot i} = \frac{\text{cov}[w \cdot r(M) - r(L), r(i)]}{\text{var}[w \cdot r(M) - r(L)]} \) can be seen as a “relative beta.”

Define M* as the market portfolio of all assets including risky and risk-free assets, M as the market portfolio of all risky assets (as in MPT), F to be the traditional risk-free asset, and w to be the market value weight of M in M*, i.e., r(M*)=w*r(M)+(1-w)*r(F). Then (3A) can be restated as in (3B) below to make it compatible with CAPM.

\[ E[r(i)] - r(F) = \Gamma_{M\cdot i} \cdot E[w \cdot r(M) + (1 - w) \cdot r(F) - r(L)] \quad (3B) \]

One will quickly note the similarities between equations (1) and (3A). In a simplistic world, where liabilities do not matter or are deterministic, r(L) = r(F), the CAPM model in equation (1) is retrieved. A major difference between equations (1) and (3A) is that the “beta” is replaced in equations (3A) or (3B) with the “relative beta.” Note also that if the liability portfolio is a linear combination of multiple assets, as in a typical SAA, one might expand the “L” in the equation as needed to give additional terms. We discuss this aspect of the equation in Section 4, in the context of explaining Fama-French [or Chen et al. (1986)]. However, this adjustment of focusing not just on assets (or wealth), but rather the ratio of assets/liabilities (or funded status), reveals many of the reconciling aspects of RAPM.

---

15 Recall that relative to L, F is a risky asset, even though it has no volatility or correlation to other assets/securities.

16 In a deterministic world, the liability should grow at the risk-free rate to prevent arbitrage opportunities.
When early tests of CAPM suggested a higher intercept and a lower slope than indicated in the CAPM in equation 1 (suggesting low/high beta assets had a systematically higher/lower return than the model predicted), Black (1972) suggested a revised model where the traditional theoretical risk-free rate is replaced by a zero-beta asset, Z, which would have a higher return than F. The Z portfolio would also have non-zero volatility and correlations with other assets (but a zero correlation to the market portfolio, hence its zero-beta moniker) and Two Fund Separation would hold. However, the paper is silent on the exact details of this zero-beta asset or portfolio. Viewed from the broader RAPM paradigm, the liability proxy portfolio is a candidate for the zero-beta “asset,” and shows how Black (1972) is possibly correct, though it requires a strong assumption on the correlation between the liability and risky portfolios. In contrast, RAPM is a less constrained model.

Just like CAPM, this asset pricing model can be used to derive asset-allocation recommendations consistent with LDI practice, and there is prior research to demonstrate the appropriate risk-adjusted performance measure in this relative world.

2. Two ways to incorporate liabilities (funded status and surplus/deficits)

This paper has briefly made the case for incorporating liabilities into the core asset-pricing framework, as opposed to including assets only (as is the case in existing literature). However, there are two approaches through which liabilities can be incorporated into the model, which begs the question: should the focus be on surplus and deficits (assets minus liabilities) or funded status (asset-liability ratio or funded ratio), to derive RAPM above? A strong case can be made for both approaches depending on how investors behave and Merton (2014) argues for funded status.

---

17 Waring and Whitney (2009) make an argument for “L” being the risk-free portfolio in a CAPM setting, when investors worry about asset-liability considerations. However, their result is driven by brute force rather than an equilibrium or arbitrage argument.

18 Muralidhar (1999) extends Modigliani-Modigliani (1997) to demonstrate that the modified Sharpe or Information ratio might not be adequate in a relative paradigm if the investor is keen to establish that skill is inherent in the investment process. For brevity, we will not focus on risk-adjusted performance measures in this article.

19 Sharpe and Tint (1990). Interestingly, they did not extend their paper on asset allocation into examining the ramifications for asset pricing, which is a bit of a surprise since Prof. Sharpe was one of the key developers of the CAPM framework.
Note, however, that the RAPM model suggested in equation (3A) above is derived from maximizing the objective function, which focuses on maximizing funded status, but makes a simplifying assumption to solve the model in a linear form, that $r(\text{assets})/r(\text{liabilities})$ can be approximated by $r(\text{assets}) - r(\text{liabilities})$. The choice of funded status as the key objective is driven by our survey responses of real CIOs as noted earlier.

Even though Sharpe and Tint (1990) and others consider surplus maximization as the key objective in their models of asset allocation in the presence of liabilities, a convincing case can be made for changing the objective function to funded status maximization. Because the funded status perspective is under represented in the literature, we emphasize its impact here. All the regulations and reporting in most countries are on assets/liabilities (funded ratio) and not assets-liabilities (assets minus liabilities). With respect to regulations, in the Netherlands, once the funded ratio falls below 105%, asset owners undergo a lot of stress and headache with respect to regulator audits, having to write recovery plans, increase contributions, potentially lowering benefits and revising asset allocation strategies. U.S. regulations require corrective action, but much less strict and onerous than the Netherlands. Today, with more than 90% of U.S. (corporate and public) pension funds operating with less than 100% funding, and regulators requiring that corporate funds make up the funding gap in three years, there is still substantial allocation to non-liability assets (or risky behavior in a RAPM context). On the flip side, consumption-based asset pricing models are well established in finance, and one must question the value of a model that is based upon utility defined by a ratio. In this regard, the surplus may be a better target to optimize and we are agnostic to the choice. Our version of RAPM in equation (3A) is a linearized model of a probably more complex three-dimensional model, as described below in Section 4. Appendix 1 provides a contrast in the asset pricing and allocation implications between maximizing surplus versus funded status.

---

20 See also Elton and Gruber (1992) who focus on maximizing the return on net worth, where net worth is defined as assets - liabilities. Their paper has many prescient insights on investment strategies to follow for liability-aware investors.

21 Though we show in Muralidhar et al. (2014b) how RAPM can be derived using the Consumption CAPM with respect to Reference Levels literature by amending the reference level to be the liability proxy.

22 We thank Professor Michael Brennan for this observation.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

3. Prospect theory reconciliation
Kahneman-Tversky examined human behavior and found it hard to believe that an individual with U.S.$10,000 in wealth would be indifferent to gaining U.S.$500 (from starting with U.S.$9,500) or losing U.S.$500 (from U.S.$10,500). They suggested MPT was missing a reference point, and this paper highlights liabilities as the appropriate reference point, as wealth must be measured relative to liabilities. Kahneman-Tversky speculated that the reference point might be the “current asset position,” but they felt investors appear to derive value from changes in wealth, not levels of absolute wealth as assumed in MPT. Barberis et al. (2001), in an excellent paper, attempt to capture Prospect Theory in a utility theory framework. In addition to the utility of consumption, they first add a utility of changes in financial wealth to incorporate the missing Prospect Theory aspect. The utility of changes in financial wealth is based on the performance of risky assets, and on the assumption of loss aversion. More importantly, this loss aversion is based on some reference benchmark level of stocks (e.g., some past historical price) and that becomes the state/exogenous variable and reference point. Should the reference point be the stock market, as suggested by Barberis et al. (2001), or liabilities? Interestingly, while Kahneman and Tversky (1979) (KT) attempt to anchor their questions to some relevant notion of wealth in society (i.e., tying the questions to values meaningful to average Israeli monthly income), they may have unintentionally touched on the importance of liabilities without explicitly realizing it when they stated, “[F]or example, the utility function of an individual who needs U.S.$60,000 to purchase a house may reveal an exceptionally steep rise near the critical value.” In our parlance, this “critical value” of U.S.$60,000 is the liability!

A recent IMF analysis of institutional investors shows that the case may even be stronger for the liability-centric model. While arguably KT were examining the behavior of individuals, and our data pertains to institutions so there may be a disconnect unless we assume a representative investor, Figure 1 provides an interesting analysis of risk aversion of a select group of institutional investors, based on their funded status from 2001-11.

23 They note that “[T]he same level of wealth, for example, may imply abject poverty for one person and great riches for another—depending on their current assets,” which in effect is nothing more than a normalization for different liabilities.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

**Figure 1: Contrasting funded status and risk aversion (2001-11)**

Sources: Boston College Center for Retirement Research; and IMF staff estimates.
Note: Size of bubble represents allocation to alternative investments; 2011 is 25.5%.

This data from IMF (2013), based on allocation information of the weakest 10% of U.S. public pension plans, demonstrates a few interesting characteristics.

---

The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

First, as the funded status of public pension plans began to decline from 2001 onward, their allocation to risky assets generally increased in the following year. In finance parlance, risk aversion declined with funded status, with a marked increase in allocation to alternative assets as represented by the size of the bubbles. Second, clearly risk aversion is also dynamic in time, as suggested by Barberis et al. (2001). Third and most interesting, which potentially contradicts an assumption made in the utility function in Barberis et al. (2001), there was a generally upward trend in stock markets from 2003-07. Therefore, from their model, one would assume that as equity prices rise, risk aversion would also increase; but in this subset of funds, risk aversion actually declines, and substantially so. The reason may be because, over this period, the value of liabilities for most pensions worldwide rose more than the value of assets in which the funds were invested (driven in large part by central banks depressing interest rates and an additional phenomenon described below in Section 7). This opens the door to the alternative approach that links risk aversion to funded status, as opposed to equity levels, reinforcing the importance of a liability-centric approach and a better understanding of how investors make decisions. From the perspective of prospect theory, this observation could be heuristically explained by the investors’ behavior of loss aversion. This behavior, linking risk aversion to funded status, is also validated in a more recent study of 126 U.S. public pension plans covering U.S.$2.6 trillion in assets as of FY 2012 – we see the same trend in increased riskiness of portfolios as funded status declined (even in periods where asset values rose). A recent survey in the November 2014 issue of CIO magazine demonstrates the opposite effect; as funded status rose 8% over 2011-14 (coincident with rising stock markets) on a PPA basis, investors increased their allocation to long duration fixed income by 8% and narrowed the gap between liability and asset duration from 2.8 years to 2.6 years (i.e., risk averse behavior). This rotation in-and-out of risky assets has interesting implications for CEOs at investment firms as they can estimate demand for their products based on the trend in funded status.

---

25 http://www.publicfundsurvey.org/publicfundsurvey/summaryoffindings.html. Figures A and B show how the funded status (A/L) has changed over the years. Figure O shows the change in asset allocation over the same window and one can see how bonds declined and alternatives increased (i.e., risk aversion declined as funded status fell).

The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

In summary, incorporating liabilities as the reference point may resolve some of the Prospect Theory critiques of MPT, which proves the earlier point that both perspectives have merit when viewed through the appropriate lens. Also, we suggest a much more relevant and appropriate reference portfolio than just the stock market, which appears to be backed up by the simple evidence shown here. Clearly, a more detailed analysis needs to be conducted to finalize these findings [Muralidhar (2014)].

4. Fama-French and the search for more factors
CAPM has been heavily tested and many additional variables have been proposed for inclusion in the model (e.g., Harvey et al. (2014) test as many as 312 factors), though not always with a model to justify their inclusion. Equations (3A) and (3B) may add one possible explanation for additional factors to the model, ignoring the fact that the coefficient on the variables differs from traditional beta, based on how liabilities have been explicitly or implicitly proxied.

For example, as noted in Dempsey (2013) in reviewing Fama–French model, “the expected return on portfolio j is expressed as

\[ E[r(j) \cdot r(F)] = \beta_1 E[r(M) \cdot r(F)] + \beta_2 E[r(SMB)] + \beta_3 E[r(HML)] \]  

where SMB = portfolio of small capitalization (S) stocks minus large capitalization (B) and HML = portfolio of high-book-to-market (H) stocks minus low-book-to-market (L) stocks.” Now, it is clear that if the liability return in (3A) can be proxied by the combination of the risk-free rate, the size factor and the book-to-market factor, then the equation (3A) can be expanded to (4). That is, additional factors can be interpreted as proxies of the liability factor in RAPM.27 To see this point, one may revert to the evolution of investment approaches, noting that through the late 1990s, most U.S. funds proxied their liabilities via their SAA, with portfolios composed of equities (typically large cap and growth, and with allocations in the 60%+ range, accounting for 80%-90% of the portfolio’s volatility) and short duration, high quality credit bonds.

27 Brennan et al. (2012) make a similar argument, but in the context of agency. We would argue that the liability impact is more dramatic than the effect of agent benchmarks.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

So it is possible that the value and size effect reflect the fact that “liability” in our equation adds a premium to value and small cap stocks (because they are excluded assets), and all they were capturing is that the primary risk factor in proxy liabilities was large-growth stocks.

Liabilities may affect other factors, too. For example, post the tech bubble meltdown, and a severe decline in funded status among all funds globally, many pension funds began to adopt a more LDI approach under pressure from regulators, thereby providing an additional bid to long-term rates (beyond what the Federal Reserve was trying to achieve). This trend accelerated post-2008, as the second equity meltdown in a decade further damaged funded status.

In this sense, RAPM becomes even more interesting when broadening the fact that liabilities are proxied with non-equity assets, especially bonds (discussed in more detail in Section 7). One can make a similar case with the results of Chen et al. (1986) and Duarte (2013). Chen et al. (1986) include at least two variables that could be candidates for inclusion in the liability portfolio: $UTS = \text{long term bonds} - \text{short term bonds}$; $UPR = \text{long term corporate bonds} - \text{long term government bonds}$. Jagannathan and Wang (1996) offer at least one fixed income variable (yield spread between BAA- and AAA-rated bonds). Again, most liabilities, especially in pension funds, have a component tied to economic growth/industrial production, because pensions depend on wages (and most university endowments have a spending policy that assumes 5% growth); insurance portfolios are more often “bond heavy” reflecting their perception that the liability is bond-like. The fact that there are fixed income terms as additional factors in an asset pricing model again syncs up with the fact that most pension fund and insurance liabilities can be replicated by a long-duration bond, though many funds included short-duration high-quality bond benchmarks in their SAA for many years.

One interesting factor that requires an explanation outside RAPM is “momentum” [Carhart (1997)]. Our explanation derives from investor behavior, as most funds have rebalancing strategies around the liability proxy/SAA, to account for the relative risk caused by drift away from the SAA. In effect, rebalancing strategies are “mean-reverting” and counter-trend, in that it sells the winning asset and buys the losing asset in order to rebalance. The momentum effect probably captured this tendency of investors to rebalance and is validated in Sharpe (2010), as he demonstrates how it can be profitable to use a momentum strategy when everyone else is counter-trending.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

In summary, depending on how liabilities used to be, and are currently proxied, one may have a model-based hypothesis of what factors can and should be added to the model, and still preserve the risk-return trade-off from CAPM, as represented by the coefficients in equation (3A). However, since liability proxying practices have changed so much over time, the relative importance of these factors will be dynamic. As noted earlier, investors who went from using implicit proxies (or large, growth equity allocations in the SAA) to explicit proxies (long-duration bonds with credit risk) would have changed their investment behavior, thereby affecting asset prices over the test period.

5. Other models of asset pricing – home bias, benchmarking, agency, background risk
The broader RAPM paradigm also explains other related aspects of finance literature, including the home-market bias or firm-bias allocations in portfolios (derived from the “Keeping up with the Joneses” utility). Other literature focused on agency issues in investment management and the impact of benchmarking investors are also interpreted as merely simple variants of the notion of relative asset pricing, with each paper effectively changing RAPM’s “liability.” While it is still early, there are indications that this type of model might even be able to provide a basis for the equity risk premium puzzle and some of the findings of the theory of background risk.

29 Brennan (1993) is the path-breaking work in this area.
30 Gomez and Zapatero (2003) suggest a two factor CAPM in the presence of benchmarking, which can be seen as early work validating Fama-French with a model. The difference between this approach and RAPM or Reisman-Lauterbach is that in this model the agent holds a portfolio composed of all the securities in the market portfolio but just weighted differently from market weight. Our model assumes that the liability portfolio is carved out of the market portfolio.
31 Mehra and Prescott (1985). The puzzle relates to the fact that it is hard to explain why investors hold government bonds in as high a proportion as they do, plus the fact that “the historical U.S. equity premium is an order of magnitude greater than can be rationalized in the context of the standard neoclassical paradigm of financial economics.” In RAPM, the demand for bonds can be driven solely by their liability hedging benefit. The premium can also be rationalized by the fact that we offer a different model of the size of the premium under RAPM.
32 Background risk pertains to a theory that tries to explain the portfolio allocation decisions of individuals that do not fit the typical MPT recommendations. For example, those who have risky income or wealth characteristics (termed high “background risk”), tend to hold less stock (and also less diversified portfolios). In our model, L is modeled generically so it could even potentially be used to capture a background risk process.
6. Relationship with reality – asset allocation and asset pricing and implications for investors

6.1 Asset allocation

One issue with CAPM was that investors did not appear to engage in Two Fund Separation. In contrast, among LDI-aware investors and consultants advising such funds, there is an increasing tendency to split assets into three buckets: liability hedge (LHP), risk seeking or growth portfolio (RSP), and cash. This fits reasonably well with the RAPM framework, as one would expect a form of Three Fund Separation, with investments in L (LHP), M-L (RSP) and F (cash). The tougher issue is to reconcile the investment behavior of investors with implicit liability portfolios, a similar conundrum faced by CAPM.

An interesting trend that makes the liability-centric view even more important is that many institutions will gradually move to much more explicit liability proxies. For example, U.S. public pension plans are currently allowed to use a technique (for determining contributions) that permits discounting their projected liability cash flows with the expected target return on the asset portfolio, and hence the SAA tends to have substantial equity allocations as a high target return leads to lower liability valuations and hence higher funded status estimates. However, going forward, there is incredible pressure to use discount rates similar to those used by U.S. corporate funds, U.K. funds, Dutch funds and Danish funds (to establish their funded status for contributions), namely, a long-duration bond-based discount rate reflecting the risk of pension cash flows. Even if the technique for determining contributions does not change, some rating agencies use this latter technique to evaluate the level of state and county underfunding, thereby affecting state/county credit ratings. If these efforts reach a tipping point, allocations to long-duration bonds among U.S. public plans will increase (because higher borrowing costs will hurt public funds, if credit ratings decline), and lead to a dramatic shift in asset allocation and to how liabilities are proxied. This, in turn, will affect asset prices as assets in the current SAAs, especially stocks, will be less favored as well to reflect this rebalancing.

---

33 Amenc et al. (2010) document this facet of LDI investing.
6.2 Applicability to all assets
Most tests of MPT are conducted largely on stocks, even though MPT assumes that the market portfolio includes all assets.\(^\text{34}\) This is most evident in the Fama-French and related research. Focusing on only stocks (or treating stocks the same as bonds) risks ignoring the unique characteristics of a substantial part of capital markets. To convey orders of magnitude, equities have declined from 50% of the estimated global market to about 35% \([\text{Doeswijk et al. (2012)}]\) and bonds are increasingly the dominant asset in institutional portfolios. However, if the liability proxy portfolio strictly consists of bonds, then RAPM offers a different return driver for bonds from stocks.

RAPM highlights the tug-of-war in predicting returns between the reward required for pure asset risk and the compensation resulting from the liability hedge of a particular asset. In effect, assets with a strong correlation to liabilities, within the risky asset bucket, would earn a lower rate of return than those that do not have this attractive property. Large and growth stocks might be less rewarded than small and value stocks (as mentioned earlier), but one can extrapolate this exercise to all assets. This highlights the fact that the role of correlations in the asset pricing paradigm is more complex and simple mean-variance plots for assets are insufficient for a complete understanding of asset returns and that possibly a three-dimensional mean-variance-correlation analysis is required.

6.3 Asset pricing implications
RAPM provides a different model for asset pricing than CAPM, suggesting that investors might need to reassess assumptions made in their asset pricing and asset allocation models. Interestingly, this provides opportunities to sophisticated investors to capture risk premia and unique profits. For example, changing pension regulation will modify the required structure of the liability portfolio from one that includes stocks and short-duration bonds to one solely replicated by long-duration bonds (with a hint of credit risk in the U.S.). In the past, demand for long-duration bonds increased while the relative attractiveness of stocks declined when U.S. corporate and European regulation changed.

\(^{34}\) An exception would be Elton et al. (1995), which focuses on using APT approaches to explain bond fund performance. Bonds pose a problem because the instrument continually ages and matures, hence, it is easier to test for appropriate factors on bond funds that have a somewhat stable duration and focus.
A counter experiment presented itself in 2013, when long-term rates rose because of the Fed taper (lowering liabilities) and stocks rallied causing pension funds to close the funding gap, leading to increasing demand for bonds as investors moved from loss aversion to risk aversion, thereby lowering rates in early-mid 2014. This again provides an interesting avenue for research on the impact of changing liability compositions and reactions of investors to changes in funded status.

6.4 Investor implications
6.4.1 Impact from increased LDI adoption
For innovative investors who recognize the value of this paradigm, there are substantial profits to be made. Van den Brink (2012) discusses the European LDI experience and notes how his strategist, Patrick Groenendijk, saw the trend toward LDI in Europe taking root in 2003. His strategist suggested that the PME pension fund take a large position in longer-duration bonds to take advantage of declining future yields — van den Brink’s only regret being that he was meek in the size of the trade as it proved to be very profitable!

As van den Brink (2012) notes and RAPM validates, liability-driven, or equivalently benchmark-driven, investments may create short-horizon momentum in asset returns due to portfolio rebalancing activities. As explained in this paper, in liability or benchmark-driven investment approaches, investors will borrow and invest in liability hedging or benchmark assets, and then invest in risk-seeking assets with the balance of their portfolio so that the portfolio is optimized. If for some reason the value of the liability or benchmark increases (e.g., changes in regulation or some new information is introduced to the market), then investors will end up with insufficient borrowing relative to the liability value. This in turn should lead to an increase in borrowing and investment in the non-hedging portfolio.

35 Vayanos and Woolley (2012) explore the implications of agency on momentum strategies and our own approach is similar because both are relative paradigms.
36 Here, liability asset is assumed to have a perfect correlation with liability value. If not, then the implication will be slightly different.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

This will induce increased demand for assets and will drive asset prices up and vice versa. This process of changing values/changing demand and the potential for capturing value is demonstrated through three simple examples in Appendix 2. Interestingly enough, the implication of MPT is quite opposite in that when risky asset value increases, investors will sell some of the risky asset to maintain the same risky-riskless asset composition.

6.4.2 Exploiting “non-liability” assets
In addition, as Section 4 suggested, assets not included in the “liability” portfolio will earn a premium and typically, once these premia are discovered, investors tend to add these assets to the “liability” proxy. This was clearly the case with small and value stocks in the U.S. in the mid-1990s, leading to a dilution in the premia. A more recent trend has been to include low volatility stocks in portfolios because of the perceived anomaly that they provided a premium over high-volatility stocks [Baker et al. (2011)]. The clear message to the innovative investor is that trying to play at the margin by holding assets in the “liability” portfolio is not likely to lead to sustainable excess returns; rather, investing in assets excluded in the “liability” portfolio before they are recognized by the general market is the way to go, but this will engender relative/peer risk and needs to be managed carefully.

6.4.3 Need for adoption of innovative LDI processes and products
Moreover, there has been a tendency in the investment industry to follow certain fads, and as Warren Buffett noted back in 1975, “Wall Street is a succession of fashions.”37 Many of these were only investment focused: technology stocks in the late 1990s; private equity/venture capital and hedge funds/fund-of-funds post the market decline of 2000–02; 120-20 equity strategies in the mid-2000s; “risk parity” strategies post 2008, and more recently, “smart beta” strategies that seek to weight assets in an index typically with a valuation/momentum/volatility bias as opposed to market cap weighting. However, all of these innovations in investment products are largely investment focused and ignore the liabilities completely.

The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

Sadly, these strategies have had poor performance after clients had piled into the trade leaving funds worse off from a funded status perspective as liabilities did not experience similar performance (i.e., there was a poor correlation between these strategies and liabilities). This focus on just investments (as opposed to investments relative to liabilities), in some part caused by the fact that MPT focused on just investments and is still the basis of investment decision-making in many funds globally, has caused major problems, as noted by van den Brink (2012) in the context of the Netherlands. This focus on just assets in investment decision-making, as opposed to funded status, for the better part of the last two decades is probably part of the reason why many pension funds have declined from a position of full funding in the late 1990s to being severely underfunded by 2012, as shown in Figure 2 (data taken from the 2013 U.S. Public Fund Survey covering 126 pension funds with more than U.S.$2.6 trillion in assets).

Figure 2 plots the aggregate actuarial funding level among plans in the Survey since its inception in FY 2001. If RAPM had been the prevailing paradigm in the 1990s, there might have been greater focus/pressure on funds, possibly even driven by regulators, to adopt the types of strategies that Elton-Gruber (1992) proposed to hedge liability risk. Instead, by utilizing MPT/CAPM approaches to portfolio construction, funds took unintended relative risk – investing in equities when the liabilities resembled a bond exposure. This has left many funds in a tough position, especially since weak economic growth has put pressure on the sponsors (the state and local governments), preventing them from covering these losses with new contributions. Given the weak projections for economic growth globally, there is a critical need to improve investment approaches for these funds, as any further decline in funding will lead to the closure of these defined benefit schemes (as has been the case with many U.S. corporate pension funds), and in some extreme cases even reduced pensions in their retirement years (e.g., public fund retirees in the city of Detroit). Barrett et al. (2012) recommend a more dynamic approach to managing portfolios, rebalancing the asset allocation to reflect changing market conditions, but ensuring that this process produces a stream of returns that improves the asset-liability position of the fund. In short, the acknowledgement and acceptance of the RAPM paradigm is of utmost importance, not only to improve investment outcomes, but also to improve social outcomes.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

Figure 2: Decline in funded status of U.S. public pension plans

![Graph showing decline in funded status of U.S. public pension plans from 2001 to 2012. The graph includes data for actuarial value of assets and liabilities, with a decline in the aggregate funding ratio from 100.9% in 2001 to 73.5% in 2012. The source is the Public Fund Survey, November 2013.]
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

6.4.4 Popularity of life cycle/target-date funds
Increasingly, investment products called life cycle funds or target date funds, where the allocation to equities and bonds change based on the age of the investor, are becoming popular default options and are gathering substantial assets. The World Bank is increasingly advocating such products for country-level social security systems [see, Price and Rudolph (2013)]. In the U.S., a target date fund (TDF) is considered Qualified Default Investment Alternative (QDIA), giving the employer safe-harbor protections from the Department of Labor, and the U.S. Congress recently passed regulation to make these funds the default option for the Federal Thrift Savings Plan.

RAPM and Merton (2014) would suggest that rather than changing allocations based on the age of the participant, which is easy to monitor, the allocation should change based on the achievement of some funded-status objective. The experience in 2008 with these TDFs was not positive as many such funds carried too much equity risk and a funded status-based rebalancing might lead to more secure social outcomes, especially since these funds are being made into default options on the assumption that this is the best alternative for unsophisticated investors. The ABN AMRO Pension Fund which rebalanced based on funded status weathered the crisis in 2008 much better than 99% of the funds globally [Muralidhar (2011)].

Therefore, it is our hope that the RAPM paradigm gradually helps replace the existing asset-only/absolute focus on portfolios and risk, and introduces a more “liability-centric” approach to portfolio construction. As more investors adopt an explicit liability proxy process, the RAPM framework becomes more interesting and relevant for asset pricing, asset allocation shifts as well as new methods for calculating risk-adjusted performance. As this transition takes place, there will be unique opportunities for the development of new investment products/processes (e.g., to dynamically improve the liability hedging process as the funded status of pension funds improves),38 and also opportunities to capture unique trend/momentum of various assets that could lead to one-time gains to portfolios.

38 Muralidhar (2011a) and (2011b) examine view neutral and view-based portfolio rebalancing. The former reference highlights how the ABN AMRO pension fund, by adopting view neutral ALM rebalancing, preserved its funded status in 2008 above the 105% threshold when most other Dutch funds suffered.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

7. Summary
In a liability-centric world of asset pricing, utility ought to be examined not just relative to wealth, but also relative to stochastic liabilities.

This paper postulates a simple RAPM anchored in a liability-centric world, to show how, if the seemingly contradictory theories and findings of various academics are reframed from this perspective, they might not all be so contradictory.

Reconciling these seemingly competing theories is important as it allows practitioners to adopt effective processes and not get confused as to which paradigm to adopt. It appears, by opening the relative tent wide enough, there may be a big enough platform to accommodate the traditional CAPM, Black’s CAPM, Prospect Theory, Fama-French and other factors, Agency Theory, Benchmarking Theory, and even the Home Bias or Background Risk phenomenon, thus providing fertile ground for some talented academic/s to develop a Unified Theory of Asset Pricing Finance [Muralidhar et al. (2014 a)].

It would seem paradoxical to invoke the name of the Buddha in a discussion of investment finance, but a simple couplet attributed to him in reconciling the views of the five blind men who examined an elephant and bickered over what it might be seems appropriate. Buddha is believed to have composed this verse to resolve the differences:

“O how they cling and wrangle, some who claim
For preacher and monk the honored name!
For, quarreling, each to his view they cling.
Such folk see only one side of a thing.”

The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

Appendix 1 – funded ratio versus surplus
As a relative measure inclusive of liabilities, two competing measures exist – surplus (or equivalently excess return) and funded ratio (or equivalently relative return index). Sharpe and Tint (1990), Leibowitz and Henrikson (1988), Rudolf and Ziemba (2004), and Cornell and Roll (2005) consider surplus as a relative measure, whereas Reisman and Lauterbach (2002) and many pension regulators consider the funded ratio (or relative return index) as the relative measure. RAPM discussed in Section 1 derives from assuming the funded ratio as a relative measure, but in this appendix, we show how the two different measures impact investors’ asset allocation decisions. Muralidhar et al. (2014b) develops RAPM based on both surplus and funded status as well from an investment and consumption approach.

Denote $A_0$ and $L_0$ as the asset and liability values at time 0, respectively, and $r(A)$ and $r(L)$ as the respective asset and liability returns. In the case of funded ratio optimization, the future funded ratio can be expressed as:

$$[A_0 \times (1 + r(A))]/[L_0 \times (1 + r(L))] = (A_0/L_0) \times [(1 + r(A))/(1 + r(L))]$$

Therefore, at time 0, the funded ratio plays the same role of an initial wealth in the asset-only optimization case. For log utility, for example, initial wealth does not matter in asset decisions and hence in RAPM using log utility, initial funded status will not matter in portfolio decisions. In the case of surplus optimization, future surplus can be expressed as:

$$A_0 \times (1 + r(A)) - L_0 \times (1 + r(L)) = A_0 \times [(1 + r(A)) - (L_0/A_0) \times (1 + r(L))]$$

Therefore, the inverse of initial funded ratio, $L_0/A_0$, acts as a scaling factor on liability portfolio return. For log utility in RAPM using surplus, for example, the initial funded ratio will matter in asset decisions.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

The two competing measures will result in different holdings of the liability portfolio. If, for example, \( A_0 \) and \( L_0 \) are U.S.$0.5 billion and U.S.$1 billion, respectively, the liability portfolio holdings will be U.S.$0.5 billion and U.S.$1 billion for the funded status and surplus approach, respectively. In other words, the liability holding amount for the funded ratio case will be the same as the asset value, whereas it will be the liability value in the surplus case. The RAPM in this paper does not have any funded ratio component because the relative measure is defined as the funded ratio. If, instead, it is defined as surplus, then the initial funded ratio will appear in the RAPM equation. This is explored in future research.

Appendix 2 - RAPM rebalancing based on funded status (versus surplus)

The following example may make the implications of the various approaches easier to follow. Suppose that we have three time points, \( t=0, 1, 2 \) and two periods, period 1 and period 2. At \( t=0 \), suppose that investor has assets valued at U.S.$100, and a liability valued at U.S.$80. Let us assume that there is a liability hedging asset and it has a perfect correlation with actual liability. For convenience, the borrowing rate is assumed to be 0%.

Case 1 (funded ratio utility): rebalancing because of a change in asset values

In our paradigm, the investor will borrow U.S.$100 (the initial asset value), and invest it in the liability hedge. Then, the investor will invest the asset pool of U.S.$100, without further consideration to the liability, in a “non-hedging portfolio” that is composed of riskless and risky assets according to the regular MPT approach. Suppose that during period 1 some positive news hits the market and the non-hedging portfolio value increases to U.S.$120 and the liability value increases to U.S.$110. Then the investor’s new endowment value at \( t=1 \) will be U.S.$130 (=U.S.$120+U.S.$110–U.S.$100), and new liability value will be U.S.$110. Therefore, the investor will borrow an additional U.S.$30 and invest U.S.$20 in the liability hedge and U.S.$10 in the non-hedging portfolio, which will result in price increase in risky assets in the non-hedging portfolio as well as in the liability hedge due to increased demand. Investors, who like van den Brink’s strategist, Patrick Groenendijk, can foresee such trends will be able to profit from such market movements as news hit the market.40

---

40 Barrett el al. (2011) discusses a form of dynamic beta management that rotates a portfolio’s asset allocation in response to data changes, but did not consider the implications of changes in investor behavior/rebalancing in response to these data changes.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

Case 2: comparison to the case where investors focus on surplus
The investor will borrow U.S.$80 (the liability value), and invest it in the liability hedge. Thereafter, as in Case 1, the investor will invest the asset pool of U.S.$100, in the non-hedging portfolio that is composed of riskless and risky assets. Suppose that during period 1, the same positive news hits the market and the non-hedging portfolio value increases to U.S.$120, as in Case 1, and the liability value increases to U.S.$90. Then the investor’s new endowment value at \( t=1 \) will be U.S.$130 (=U.S.$120+U.S.$90-U.S.$80), and new liability value will be U.S.$90. Therefore, the investor will borrow additional U.S.$10 and invest it in liability hedge, which will result in a price increase in liability hedge due to increased demand. This case will be more applicable to situations of benchmark-driven investment, but note the difference relative to Case 1.

Case 3: change in regulation
We examine the case where changes in regulation can impact how liabilities are valued and this is a potential live issue that could impact all U.S. public pension funds (from a reset in the discount rate) as noted earlier. Under such a situation, the instrument that was used to previously hedge the true liability may not change in value commensurate with the change in the true liability. In our example above, suppose that the change in regulation results in the value of the liability increasing to U.S.$110. Under the RAPM (funded status) approach, no further rebalancing is necessary because there is no change in size of the asset pool (=non-hedging portfolio value - borrowing + liability asset value). However, if an investor assumed that all funds were approaching their investments from a surplus perspective, as assumed in previous research, they would expect funds to be rebalancing (because the hedging amount is set according to liability value, which increased to U.S.$110 from U.S.$80). They would expect additional borrowing of U.S.$30 to be made and be invested in the liability hedge, which they expect will result in the price increase in liability assets. Therefore, investors who apply the wrong paradigm might potentially make the wrong expectation about future market movements.

---

41 Alternatively, this could happen because of a new actuarial method being applied to the liabilities.
The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

References


Duarte, F., 2013, “Inflation risk and the cross section of stock returns,” Federal Reserve Bank of New York Staff Reports No. 621, New York, NY


The relative asset pricing model: implications for asset allocation, rebalancing and asset pricing

Muralidhar, A., 2011a, A SMART approach to portfolio management: an innovative paradigm for managing risk, Great Falls, VA: Royal Fern Publishing
Muralidhar, A., 2011b, LDI = Less than desirable investing?” aiCIO, LDI special issue, Fall, 87, New York
Partington, G., 2013, “Death where is thy sting? A response to Dempsey's dispatching of the CAPM,” ABACUS 49, Supplement
Abstract
As presented in leading corporate finance textbooks, the predominant method for making capital budgeting decisions is discounted cash flow analysis. The primary benefit of this approach is that it allows for different discount rates to be used for different projects. In this paper, I argue that this alleged benefit is, in fact, a detriment for two reasons. First, the betas that determine the differing discount rates are not only measured with significant error, but tend to drift substantially over the life of typical capital budgeting projects. Second, introducing differential discount rates is likely to conflict with other “general equilibrium” management objectives related to developing a successful corporate culture that promotes collaboration and innovation. This explains why many companies, particularly companies that have a large number of growth options, choose alternatives to discounted cash flow, such as the internal rate of return.

I would like to thank Ivo Welch for ongoing discussions regarding the capital budgeting decision.
1. Introduction
As presented in leading corporate finance textbooks, the capital budgeting decision is a mathematical exercise involving two steps: (1) a procedure for estimating expected future cash flows and (2) an assessment of the required return associated with those cash flows provided by the application of an asset pricing model (most commonly the CAPM, but also the Fama–French three-factor model). The capital budgeting decision rule is to take projects for which the net present value (NPV) of the expected cash flows discounted at the required return is greater than zero.

A surprising empirical result is that sophisticated managers, often educated at elite business schools, fail to follow the standard procedure. For example, based on an extensive survey, Gompers et al. (2014) report that few private equity investors use discounted cash flow or net present value techniques to evaluate investments. They also observe that few private equity investors use the capital asset pricing model, or any asset pricing model, to determine the cost of capital. Instead, those investors typically target an internal rate of return in the range of 20% to 25%, a return well above what would have been produced by an asset pricing model.

The purpose of this short paper is to offer an explanation for this observed behavior that also leads to a cross-sectional prediction regarding when the textbook discounted cash flow (DCF) approach will be employed and when it will not. The key to the analysis is what I refer to as the “general equilibrium” factors involved in capital budgeting. The factors come into play because capital budgeting decisions are not exogenous financial exercises, but are a critical component of the overall operation of the firm.

---

2 See for example, Berk and DeMarzo (2014) or Brealey et al. (2013).
Finding and exploiting profitable opportunities requires that the firm be managed in a manner that provides proper incentives and contributes to cooperation among employees. More specifically, successful capital budgeting typically will involve issues including:

- Motivating, rewarding and retaining employees
- Promoting collaboration and discouraging infighting and empire building within the company
- Avoiding the appearance of management favoritism
- Nurturing creative thinking
- Monitoring excessive optimism with regard to projects in which employees are involved and other behavioral biases

As a result, the choice between different capital budgeting decision criteria, such as DCF versus the internal rate of return (IRR), or even the multiple of invested capital (MOIC), will depend not only on the financial calculations, but also upon the manner in which capital budgeting interacts with the other general equilibrium factors.

2. DCF versus IRR in a general equilibrium context
The fundamental distinction between the DCF and IRR decision rules is that DCF takes account of the possibility that different capital projects may have different required rates of return. This distinction arises because the stochastic process of returns for different projects potentially have differential covariances (betas) with either the market or the priced factors. If the firm chooses to ignore these distinctions, then the IRR and DCF decision rules become essentially equivalent with regard to the accept/reject decision of capital budgeting. The only remaining benefit of the DCF approach is that the net present provides an estimate of the value wealth created by the project.

Whereas the ability to take account of differential required returns for different projects may appear to be a benefit of the DCF approach, in general equilibrium it may in fact be a detriment. In doing this research, I was fortunate to have a detailed interaction with the CFO of a major technology company.
Capital budgeting: a “general equilibrium” analysis

His view was that using differential required returns for different projects was something to be specifically avoided. The reason was that choosing among different projects proposed by separate teams raised all the general equilibrium issues delineated above. There were risks of internal strife, claims of favoritism, and so forth. In such a situation, adding a complex and not particularly transparent calculation of differential risk premia to the mix, a procedure that would favor some projects over others, was particularly ill advised. A far better solution, according to the CFO, was to set a relatively high IRR that applies to all projects and provides a cushion that takes account of the tendency of teams to be overly optimistic regarding their own projects.

It is worth noting that the foregoing concerns would be moot if the discount rate could be measured with precision. But that is not the case, particularly over longer horizons that match the life of most capital budgeting problems. The betas that are the source of differential discount rates are not only measured with significant error, but drift markedly over time horizons on the order of most capital budgeting projects. Following a detailed study designed to assess the effectiveness of using financial models to assess differential discount rates in a capital budgeting context, Levi and Welch (2014) conclude that “the capital asset pricing model and the Fama–French three-factor model have very little ability to discriminate across similar maturity far-away expected cash flows. This renders the models of little use for practical capital budgeting for cash flows beyond a few years. There is little evidence that the model predictions were even directionally correct, much less quantitatively correct, for corporate users who had to estimate exposures and risk premia.”

In light of the deficiencies that Levi and Welch identify, and setting aside consideration of the general equilibrium factors, the marginal benefit of the DCF approach is likely to be small, if it exists at all. Taking account of the general equilibrium factors, which all point in the direction of avoiding attempting to distinguish between discount rates based on different betas for competing projects, it is not surprising that many firms use alternative criteria.
3. Cross-sectional predictions

The importance of the general equilibrium factors is not independent of the type of projects being considered. For routine projects at companies with long histories, such as a railroad replacing a locomotive or a utility building a gas-fired power plant, the internal management issues associated with the capital budgeting decision are likely to be much more muted. In addition, for projects like those that are repeated over time estimating betas is more feasible, although the drift problem documented by Levi and Welch remains. Nonetheless, the DCF approach should be a relatively more attractive alternative for evaluating such projects.

At the other end of the spectrum are companies like Facebook whose future projects are growth options that in many cases have not even been identified. For such companies, developing a “creative culture” that encourages collaboration and promotes innovation is critical. As the CFO I interacted with noted, the last thing senior management wants is that the process of capital budgeting interferes with the company’s effort to develop the proper corporate culture. In such an environment, the general equilibrium factors predominate when choosing among investment decision rules and the details of financial calculations are second order. Diverting attention to largely unresolvable and speculative issues related to estimating betas and required returns is at best an unrewarding distraction and at worst an added source of potential divisiveness.

In light of the foregoing, the analysis predicts that as one moves from projects that have the characteristics of routine replacement at one end of the spectrum to projects that are pure growth options at the other end, the frequency with which the DCF method is used will decline. This offers one possible explanation for the different survey results reported by Gompers et al. (2014) compared to those reported by Graham and Harvey (2001), who found more frequent adoption of the DCF method. Whereas Gompers et al. (2014) surveyed private equity managers, Graham and Harvey (2001) surveyed corporate CFOs.
Because private equity investors are generally attempting to transform firms and create value, their capital budgeting decisions will skew more toward growth options and less toward normal replacements than those decisions faced by a typical corporate CFO. The analysis predicts, therefore, that private equity managers will pay more attention to the general equilibrium factors and be less likely to use a DCF approach.

To make the analysis more concrete, consider the case of Amazon, which has a history of investing in a variety of growth options. Table 1 lists some of the major capital budgeting projects undertaken by Amazon in the last 18 years.

**Table 1: Selected Amazon capital budgeting projects**

<table>
<thead>
<tr>
<th>Date</th>
<th>Project Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep 1997</td>
<td>1-Click Ordering allows users to store preferred payment and shipping information so</td>
</tr>
<tr>
<td></td>
<td>that purchases can be completed with one click.</td>
</tr>
<tr>
<td>Sep 1999</td>
<td>zShops allows third parties to sell merchandise through Amazon for the first time.</td>
</tr>
<tr>
<td>Oct 2001</td>
<td>Look Inside the Book enables customers to discover books by searching and previewing</td>
</tr>
<tr>
<td></td>
<td>the text inside.</td>
</tr>
<tr>
<td>Jul 2002</td>
<td>AWS allows developers to outsource their online and application infrastructure needs at</td>
</tr>
<tr>
<td></td>
<td>commodity prices.</td>
</tr>
<tr>
<td>Feb 2005</td>
<td>Amazon introduces a unique, membership-based, express shipping program.</td>
</tr>
<tr>
<td>Sep 2007</td>
<td>Amazon's first foray into direct digital entertainment sales.</td>
</tr>
<tr>
<td>Nov 2007</td>
<td>Amazon introduces the first eInk specialized digital reader.</td>
</tr>
<tr>
<td>Sep 2011</td>
<td>Amazon introduces full featured mini tablet computers designed to interface specifically</td>
</tr>
<tr>
<td></td>
<td>with their stores.</td>
</tr>
<tr>
<td>Sep 2013</td>
<td>Amazon develops a new help service by which pushing a button immediately connects the</td>
</tr>
<tr>
<td></td>
<td>user to live video chat support.</td>
</tr>
<tr>
<td>Apr 2014</td>
<td>Amazon introduces a streaming media player tied to its digital video store.</td>
</tr>
<tr>
<td>Jun 2014</td>
<td>Amazon introduces a new cell phone with innovative features geared to electronic</td>
</tr>
<tr>
<td></td>
<td>commerce.</td>
</tr>
</tbody>
</table>

The list is not meant to be complete. It includes only large, and what I would consider successful, projects that were publicly disclosed. To be fair, Amazon undertook numerous smaller projects, some of which were not successful.
The point here, however, is not to assess the success of Amazon's decision making, but to consider the problems that Amazon faced when evaluating capital budgeting projects. In the context of Table 1, for example, would it be reasonable for Amazon to apply a different discount rate to Amazon Price than Amazon Web Services (AWS) or the Kindle? Remember that at the time Amazon was considering these projects, they were largely new undertakings. This makes it even more difficult to estimate betas than even Levi and Welch suggest because there are no direct historical proxies for the stochastic properties of the projected cash flows. If differential betas were to be adopted, management would have to explain to say the Kindle team why the beta for their project was greater than that applied to AWS. Given the immense uncertainty associated with the beta calculations, the wisdom of avoiding making such determinations is largely self-evident.

4. Caveats and extensions
The prediction that IRR is preferable to DCF in capital budgeting situations in which the general equilibrium factors are predominant should not be taken to imply that the IRR approach solves the political problems associated with capital budgeting decisions, only that it does not exacerbate them in the way the DCF approach could. To be fair, most of the internal disputes regarding choice among competing projects involves the level of the cash flow forecasts, not their betas. Because all the capital budgeting procedures use the same cash flow forecasts, there is no reason to choose between them on this basis. The point of the previous analysis was to demonstrate that in the context of debates regarding the level of cash flow forecasts, there is little to be gained, and potentially much to be lost, by allowing side disputes about differential risk premiums to complicate the conversation.

To conclude, there is one caveat worth mentioning. The foregoing has focused on differential discount rates associated with differential betas. However, discount rates also vary with term as well as with systematic risk. Unlike betas, the term premium can be measured with greater precision and with fewer disputes directly from the yields on government securities. For that reason, it may make sense for managers to divide projects into different buckets according to their duration. The longer duration projects would then be assigned slightly higher hurdle rates to take account of the term premium. Such a procedure, if undertaken transparently, should not run afoul of the general equilibrium factors.
Capital budgeting: a “general equilibrium” analysis

References
New private equity models: how should the interests of investors and managers be aligned?

Joseph A. McCahery  
Professor of International Economic Law, Tilburg University

Erik P. M. Vermeulen  
Professor of Business & Financial Law, Tilburg University

Abstract  
The recent global turbulence in the credit markets had a severe impact on all aspects of the private equity industry. In response, lawmakers introduced legislation that subjects fund managers to a registration requirement and provisions targeted at improving fund monitoring and accountability. Yet, little is known about the post-crisis scrutiny of private equity funds by investors. In this article, we examine the post financial crisis trends in the private equity industry. The evidence indicates that investors’ demands for the inclusion of more investor-favorable compensation terms have begun to take hold in European funds. Fund manager responsiveness to the demand for better terms seems more general, extending to increased investor control over fund investment decisions. The new pattern also reveals the inclusion of more straightforward co-investment rights. Finally, our findings suggest that, besides the contractual “improvements,” investors want to see more skin in the game from the managers/general partners.
New private equity models: how should the interests of investors and managers be aligned?

1. Introduction
Why has the alternative asset sector been subject to the less regulatory scrutiny than the broader financial sector? (McCahery and Vermeulen (2008)). The results of prior research on contractual mechanisms that protect investors and compensation arrangements (also known as “two and twenty”) may provide an answer (Fleischer (2008); Litvak (2009); Metrick and Yasuda (2010)). However, scholars are increasingly redirecting their attention to answering basic questions regarding the optimal level of regulatory intervention in the alternative asset market. The global turbulence in the credit markets, triggered by the turmoil in the subprime mortgage market in the U.S. in 2007–08, brought an end to the private equity bonanza as well as the laissez-faire era in the alternative asset sector. In response, policymakers and governments introduced legislation that subjects fund managers to a registration requirement and provisions targeted at improving fund monitoring and accountability.

According to analysts, the economic downturn has had a severe impact on all aspects of the private equity industry. Yet, if we compare the total amount of cash raised by private equity funds on a global scale in 2005 and 2013, the industry has not significantly changed. According to available data (Preqin (2014)), the aggregate annual fundraising amount was U.S.$362 billion in 2005 and U.S.$454 billion in 2013. What has dramatically changed is that the economic downturn has affected the level of private equity fundraising and investment activity, particularly in Europe. Market forces also continue to alter the dynamics of the private equity market. It is estimated that there were 601 funds in the market in 2005 (Preqin (2013c)). By 2013, there were more than 2000 funds engaged in fundraising activities, and it is evident that the economic downturn unleashed a wave of fierce competition among private equity funds. The increase in the average time to the final closing of a fund, which was approximately 11.3 months in 2006 and 18.3 months in 2013, is clear evidence of the competition in the private equity industry. The second important change is the increase in scrutiny of private equity funds from regulators and policymakers, but also from investors (especially in the area of the compensation arrangements) (Mulcahy et al. (2012)).

---

1 See also State Street (2013). Eighty-two percent of the 391 alternative investment managers questioned indicated that fundraising would be the most challenging activity over the next five years.
From this discussion, we ask the question: what can be done to increase investors' interest in private equity while at the same time increasing managerial responsibility? If we focus on Europe, the reliance on a regulatory overhaul for many has been the core response to the effects of the financial crisis on the private equity industry. Having assumed that interventions should aim at increasing transparency toward investors and stakeholders, the approach by regulators has been to introduce a set of harmonized rules and regulations imposing stringent registration and reporting requirements for alternative investment fund advisers/managers. These regulations seek to reduce the systemic risk emerging from the operation of private equity funds and promote the stability and efficiency of the financial markets. The Alternative Investment Funds Managers Directive (AIFMD) offers a good example of these regulatory initiatives.

So, what can we expect from the post-financial crisis' legal and regulatory interventions? There is no easy answer. The response can be separated into two different categories. At one extreme, many have argued that the AIFMD addresses the inconsistencies and gaps that currently exist in the fragmented European regulatory framework, thereby reducing risks and stimulating growth for E.U. fund managers. At the other extreme, others expect that the AIFMD will increase costs and create regulatory uncertainty [Burrows (2013)]. Some argue that the regulatory burden is overwhelming. For example, the one-off costs are estimated to range from U.S.$300,000 to more than U.S.$1 million [Wang (2013)]. Indeed, the concerns over high costs is likely to deter non-E.U. fund managers from offering portfolios to European investors and induce E.U. fund managers to shift offshore [Marriage (2013)]. The result might be that the AIFMD will significantly crowd out private equity investments that are (particularly in a declining stock market) necessary to create economic stability and jobs. In fact, there is some regulatory support for the latter view. The architects of the AIFMD have acknowledged that strict application of the stringent and costly rules and regulations would be detrimental to the formation of smaller private equity funds (that often operate as venture capital providers). As a consequence, the AIFMD includes certain exemptions that are applicable to venture capital funds.2

---

2 See Articles 3(2), 16(1), 21(3) second subparagraph, and 26(2)A of the AIFMD.
New private equity models: how should the interests of investors and managers be aligned?

Against this background, there is a general belief that the AIFMD will negatively impact the private equity industry in Europe. However, we take a different view. Those expecting a nightmare (or a miracle) from the AIFMD should not hold their breath. In the end, it is highly questionable whether the proposed regulatory initiatives will form the basis of a coherent and effective regulatory regime that sufficiently protects the interests of most sophisticated investors. For this reason, we predict that the AIFMD label will just become a necessary “boilerplate” formality to raise funds in the future. In fact, we already observe that most of the private equity funds take measures to deal with the regulatory requirements of the AIFMD by either enhancing their back office capacity or outsourcing their compliance units to specialized consultants [Duffel (2013)].

The question that remains is: will the AIFMD make fundraising more efficient? The experience with fundraising during the financial crisis suggests that fundraising levels per private equity fund will continue to lag those of the pre-financial crisis levels in Europe. Recent empirical evidence shows [Preqin (2013b)] a number of trends in private equity finance and investments in the post-financial crisis era. One development is immediately noteworthy: the “survival of the fittest” trend. To be sure, there are signs of a fundraising recovery in 2013 [Winfrey (2013)]. However, only high-quality funds with impeccable track records seem able to attract new investors. The results are clear. Since fewer managers were able to secure sufficient capital to close their funds in 2013 (873 funds) compared to 2012 (1,035 funds), the percentage of successful closings of private equity funds in the post-financial crisis era increased, unsurprisingly, to 66% in 2013 from 54% in 2012.

In this paper, we focus not only on the value of the AIFMD’s new secondary reporting and transparency requirements, but also on the other factors that may influence investors to commit funds to private equity investment. This approach is supported by the evidence that compensation arrangements in the private equity agreements between investors and fund managers may require updating and retooling in order to better deal with the high degree of information asymmetries. Although the literature has followed the pre-crisis trends, our evidence shows that investors are demanding the inclusion of more investor-favorable compensation terms and conditions in the private equity fund agreements.
New private equity models: how should the interests of investors and managers be aligned?

Thus, the benefits ascribed to the new contractual terms provide investors with more favorable management fee and profit distribution arrangements, but also give them more control over the fund's investment decisions [Donato (2011)], as well as more straightforward co-investment rights [Favas (2013)].

In addition to examining the role of investor-favorable terms and the impact of such provisions on the benefits for limited partners, this paper also studies the extent of general partners having more skin in the game as a natural market reaction to the pre-crisis transactions, where they often had little skin in the game. The economic reasoning behind top performing general partners making significant capital contributions to their own fund is that managers' interests can be better aligned with investors.

This paper is divided into four parts. Part 2 briefly describes the regulatory measures introduced by the European AIFMD. We also address the exemptions to the AIFMD. Part 3 begins with a discussion on the traditional contractual framework that governs fund formation and operation, management fees and expenses, profit sharing and distributions, and corporate governance. It then discusses how the practice of private equity contracting is changing in the post-financial crisis era, making the dominant compensation arrangements obsolete [Private Equity International (2013a)]. To assess the landscape at the macro level, we compile a list of 520 limited partnership agreements of buyout funds with vintage years of 2005 to 2014. We make sure to break the agreements down into fund sizes of 0 to U.S.$1 billion and more than U.S.$1 billion. Our list contained 199 European and 321 U.S. agreements. Part 4 concludes.
2. The regulatory infrastructure of the private equity industry in Europe

2.1 The pros and cons of private equity regulation in Europe

In economics jargon, the private equity market is replete with information asymmetries. There is inevitably a high degree of information asymmetry between the fund managers, who play a relatively active role in the development (or restructuring) and growth of portfolio companies, and the passive investors, who are not able to closely monitor the prospects of each individual company. To be sure, national “private placement” rules and regulations often offer some protection to investors. Here, private placement is understood as the marketing and sale of “investment interests” in private equity funds to a limited number of professional investors, such as institutional investors and wealthy individuals. However, the downside of the application of these rules is that attracting investors significantly increases the compliance costs and fundraising complications. This is particularly prevalent in Europe where the regulatory systems of the member states are still fragmented and only harmonized to a certain extent.

In a controversial move, European regulators introduced a directive that (if implemented properly) will enable fund managers to obtain a European passport. A possible solution to the regulatory barriers of setting up a European-wide fund is to allow fund managers to ask for a European registration in the home member state, which would then automatically be mutually recognized in other member states. The application of a harmonized and uniform regulation that would govern the marketing and sale of “investment interests” in private equity funds should make it easier for and provide incentives to investors to participate in foreign funds. The passport system would help defragment the private equity market, allegedly resulting in more cross-border oriented private equity funds. In the next section, we will turn to the AIFMD and explain the arguments in favor and against this type of regulation in Europe.

2.2 The Alternative Investment Fund Managers Directive (AIFMD)

The AIFMD provides a marketing passport for managers of Alternative Investment Funds (AIFs) that fall outside the scope of the Undertakings for Collective Investment in Transferable Securities (UCITS) Directive, such as hedge funds, private equity funds and real estate funds. The rationale behind the AIFMD is to develop a uniform set of rules and regulations for AIFs that protects investors and other market participants.
AIF managers that comply with the rules of the Directive and have obtained the “passport” will be allowed to manage or market funds to professional investors throughout the European Union. Since AIF managers’ decisions affect investors in different member states, the AIFMD aims to introduce a comprehensive and secure regulatory framework that ensures proper monitoring and prudential oversight of alternative investments that pose systemic risk. Strict rules on transparency and disclosure, valuation, risk and liquidity management, the use of leverage, remuneration, conflicts of interest and the acquisition of companies are expected to enhance public accountability and the protection of investors (see Table 1). In order to further reduce the problems arising from information asymmetries, the AIFMD requires the AIF’s assets to be safe-kept by an independent depository, which is subject to high liability standards.

**Table 1: The AIFMD in a nutshell**

<table>
<thead>
<tr>
<th>Categories of rules</th>
<th>AIFMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authorization and/or registration procedures</td>
<td>All AIFMs managing AIFs must apply for authorization with the authorities of their home member state (MS) (Art. 6 and 7 AIFMD). Disclosure of information concerning the AIFM, its members/shareholders, the managers of the AIFM, the program of activity and structure of the AIFM, remuneration policies and delegation/sub-delegation of functions (Article 7(2) AIFMD), the AIFs, their investment strategies, leverage policies, risk profiles, countries of establishment, instruments of incorporation, appointment of depositaries and the additional information of Article 23(1) AIFMD (Article 7(3) AIFMD).</td>
</tr>
<tr>
<td>Initial capital and own funds</td>
<td>For internally managed AIFs: at least €300,000 – (Article 9(1) AIFMD). For AIFs with an external manager: at least €125,000 – (Article 9(2) AIFMD). If the value of the portfolios managed by the AIFM exceeds €250 million: additional amount equals to 0.02% of the value of the difference between €250 million and the total value of the portfolios of AIFs managed by such AIFM (Article 9(3) AIFMD).</td>
</tr>
<tr>
<td>Operating conditions</td>
<td>(i) Fiduciary duties of AIFMs toward AIFs and their investors; (ii) restrictive remuneration policies; (iii) duty to identify, prevent and disclose conflicts of interest; (iv) duty to establish effective risk management systems; (v) strict rules on valuation and appointment of an internal/external valuer; (vi) strict rules on delegation/sub-delegation of functions; (vii) appointment of depositary for each AIF whose function's delegation is restricted (Articles 12 to 21 AIFMD).</td>
</tr>
<tr>
<td>Transparency requirements</td>
<td>(i) Audited annual report with audited financial statements for each managed AIF, to be disclosed to each of AIF’s investors and to authorities of home MS of the AIFM (and of the AIF if applicable); (ii) pre-investment disclosure towards prospective investors of all material information items concerning the managed AIF; and (iii) regular reporting duties to authorities of AIFM’s home MS on the markets and instruments in which it deals and the principal exposures and most important concentrations of each managed AIF (Articles 22 to 24 AIFMD).</td>
</tr>
<tr>
<td>Rules on fund managers managing specific types of AIF</td>
<td>Exemption of AIFMs managing AIFs that acquire control of non-listed companies that are SMEs (Article 26(2)(a) AIFMD). AIFMs managing VC funds are exempted.</td>
</tr>
</tbody>
</table>
We conjecture that the strict application of the stringent (and costly) AIFMD rules is likely to have a decreasing effect on the supply of private equity, thereby seriously hampering the workings of the private equity cycle. Not surprisingly, therefore, the AIFMD contains certain exemptions that are applicable to smaller funds. Article 3(2) states that, besides certain registration and notification duties, the AIFMD does not apply to (a) AIF managers that either directly or indirectly (through a company with which the AIF manager is linked by common management or control, or by a substantive direct or indirect holding) manage portfolios of AIFs whose assets under management, including any assets acquired through use of leverage, do not exceed a threshold of €100 million; or (b) AIF managers that either directly or indirectly (through a company with which the AIFM is linked by common management or control, or by a substantive direct or indirect holding) manage portfolios of AIFs whose assets under management do not exceed a threshold of €500 million provided that the AIFs are unleveraged and do not provide for redemption rights exercisable during a period of five years following the date of initial investment in these AIFs.

It should be noted that if investors are to benefit from the harmonizing effects of the proposed regulatory measures, we would expect an increase in investor confidence and interest in the private equity industry throughout the E.U. More specific details on the state of the trends in the private equity industry emerge from an examination of fundraising. Data provider Preqin (2014) observed that the aggregate amount of capital raised by funds closed was U.S.$105 billion in 2013, which is more than the U.S.$94 billion in 2012 and U.S.$72 billion in 2011. It is striking, however, to observe that the number of private equity funds that had a final closing declined from 414 in 2007 to 174 in 2013.

3 In October 2013, the European Securities and Markets Authority (ESMA) published its final guidelines on the disclosure requirements for AIFMs.

4 The AIFMD provisions slightly deviate from the registration measures introduced by the Dodd-Frank Act in the U.S. The U.S. counterpart of the AIFMD significantly extended the registration requirements under the Investment Advisers Act of 1940 to include advisers of private funds, such as hedge funds and private equity funds. The rationale behind the Dodd-Frank Act is, similarly to the AIFMD, to reduce financial market failures or systemic risk. Venture capital funds are exempted, because they do not threaten the stability and continuity of the financial system.
It should be noted that high quality funds appear to receive continuous funding for their investment activities. This is confirmed by data derived from Dow Jones LP Source (which is based on multiple closings). The result is that the median fund size has significantly increased in Europe (Table 2).

This brings us to the question of whether it is reasonable to expect that financial governance reforms, such as the AIFMD, can stimulate private sector investments in the private equity industry. Does the Directive deal with the key investment issues? Attempts to answer this question have generated a great deal of discussion in the literature. On one hand, optimists will argue that the AIFMD label will not only help reduce uncertainty and information asymmetry in the industry, but also provide an international stamp of quality. If they are correct, institutional and other investors will be more inclined to invest again in private equity (and not only high quality funds). We claim, however, that the improved fundraising view is too optimistic. Our view is supported by a recent empirical study conducted by data provider Preqin [Preqin (2013a)]. The study shows that only 19% of the 450 responding private equity investors (42% of the respondents were European investors and 40% were located in the U.S.) expected a positive impact from regulatory initiatives, such as the AIFMD. 41% stated that an increasingly regulatory approach would not benefit the industry and 40% were still unsure.

### Table 2: Private equity fundraising in Europe and the U.S. (based on multiple closings)

<table>
<thead>
<tr>
<th></th>
<th>Funds raised</th>
<th>Number of funds</th>
<th>Median fund size</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.S.</td>
<td>$133.2b</td>
<td>$160.4b</td>
<td>453</td>
</tr>
<tr>
<td>Europe</td>
<td>$59.2b</td>
<td>$58.1b</td>
<td>182</td>
</tr>
</tbody>
</table>

Source: Dow Jones LP Source
In addition to showing that European investors are not overly optimistic about the benefits of regulation aimed at contributing governance improvements to the operation and efficiency of funds, the data indicates that in order to deal more effectively with the post-financial crisis challenges of uncertainty, information asymmetry and opportunism, investors in private equity funds are becoming tougher in negotiating the terms and conditions of the funds [Preqin (2013); Burrows (2013)]. Predictably, the more active approach of investors is likely to bring about a cultural change in the private equity industry. We are already seeing signs of demands for greater clarity in private equity fund terms. In our assessment, investors increasingly prefer to invest in private equity funds that are willing to better accommodate their specific concerns, particularly related to co-investment rights and compensation arrangements and disclosures. Potentially, the emergence and standardization of new disruptive contractual practices in the private equity industry will be quickly adopted to ensure better firm performance and productivity. The next section will provide an overall assessment of the changing fundraising landscape in more detail.

3. The contractual infrastructure of the private equity industry
In Europe and elsewhere, the limited partnership structure (or an equivalent flexible business structures) is the dominant legal vehicle used in private equity structuring. Its popularity is due to its contractual nature, which allows the internal and external participants to reduce opportunism and agency costs. Indeed, the limited partnership structure permits fund managers (general partners) to achieve extensive control over the operation of their funds subject to few intrusive legal obligations. Other features, such as tax benefits, the flexibility surrounding its structure and terms and its fixed life, contribute to its continuing viability as the business model of choice for collective investment vehicles. The limited partnership has other important advantages as well. First, it is familiar to most investors and intermediaries, which contributes to its enduring popularity. Second, there is a risk that other business structures, operating internationally, could be treated as a non-transparent foreign entity and taxed as a corporate body.
In order to obtain fees and shield the individual managers from liabilities of the fund, two entities are usually created: a limited partnership and a management company, which is generally organized as a corporation. Moreover, the management company is either a separate entity from the general partner or affiliated with one of the general partners, or is a subsidiary of a bank or insurance company and, accordingly, will exercise effective control over the limited partnership. With a management company, the day-to-day management is separated from the fund, which may help in resolving some tax issues while limiting doing business and other concerns.

We have seen that the flexibility of the limited partnership plays a critical role in aligning the interests of venture capitalists and investors. This brings us to the contractual provisions that are typically employed by the private equity investors in Europe (as well as in the U.S.).

3.1 Private equity contracting and compensation arrangements
The relationship between the limited partners and general partners is, as we have seen, usually characterized as a principal-agent relationship. In order to make this work, legal practice tends to include boilerplate clauses in the limited partnership agreement that are designed to reduce agency costs by aligning the incentives of the general partners with those of the investors. The boilerplate arrangements in private equity limited partnerships can roughly be split into three separate categories:

(1) fund formation and operation provisions, such as limits on the fund-raising period, the lifespan of the fund and the required managers’ contribution, (2) management fees and carried interest, and (3) the governance structure to ensure that the fund is organized and managed in the most effective manner.

For example, a fund’s duration is usually ten years with a five years investment period, making it possible for investors to estimate with reasonable accuracy when the private equity firm can make fresh investments and, most importantly, when they ultimately will be able to recover their investments, including profits. In order to align the interests of investors and the managers, the latter are also required to make a capital commitment.
Typically, the managers invest between 1% and 3% of the fund’s total commitments. Another key contractual technique is the compensation arrangement between the fund managers and the investors. Compensation usually derives from two main sources (the so-called “two and twenty” arrangement). First, fund managers are typically entitled to receive 20% of the profits generated by each of the funds, the carried interest. A second source of compensation for the fund managers is the annual management fee, usually 2% of a fund’s committed capital, depending on the fund size (see Figure 1).

**Figure 1: Average management fee arrangements during the investment period**

Source: Preqin Fund Terms Advisor
Overall, the contractual mechanisms for determining the general partners’ compensation include both a profit sharing arrangement that balances investors’ concerns regarding pay-for-performance, and a distribution scheme to investors to limit overall fund risk-taking. Moreover, there are a variety of factors that affect fund manager compensation. First, the private equity fund’s focus and characteristics can result in quite different outcomes for fund managers. For example, funds focusing on venture capital investment are more likely to require professional staff and expertise, which leads to lower yields and higher performance fees to align interests. Conversely, larger funds, such as buyout funds, are more likely to have lower fixed compensation because they require less staff than funds focusing on venture capital. Second, the type of institutional investor and their risk appetite may influence the fixed and variable fee structure of fund managers.

Arguably, the tried-and-tested compensation arrangements in the limited partnership agreement lower transaction costs and offer contractual transparency necessary to induce investors to make their funds available for investments in their portfolio companies. However, limited partners are increasingly convinced that they have mistakenly believed that the boilerplate “two and twenty” fee structure ensures a proper alignment of interest and incentives.

It should, therefore, come as no surprise that provisions in the limited partnership agreement increasingly include measures to protect investors from managerial misbehavior. What is interesting is that European funds demand more protective clauses than their peers who operate in the more mature U.S. private equity market. A plausible explanation for this is that the attention to the contractual and organizational structure of the fund arguably increases when the private equity market, which is prone to violent cyclical movements, lacks the implicit mechanisms necessary to protect against opportunistic behavior and misappropriation that are found in mature markets, such as during the recent financial crisis.

European funds can generally be distinguished from their U.S. counterparts in terms of the stricter rules regarding the distribution of profits to managers and the preferred return to the investors.
Yet, despite these differences, the increasingly global nature of the private equity industry, as well as the recent financial crisis, may lead to substantial convergence of contractual arrangements between Europe and the U.S. in a number of important ways, particularly in the area of management fees and profit distribution (see also Figure 1).

A key contractual technique, for example, is the “scale down” compensation arrangements between the fund managers and investors. This is quite common (and more investor-favorable) in the U.S., but is now also increasingly applied by European private equity funds (see Figure 2).

**Figure 2: Mechanisms for management fee reductions - post investment period**

Source: Preqin Fund Terms Advisor - Buyout agreements vintage 2005-2014
What is a scale down arrangement? Historically, a significant majority of funds set management fees as a constant percentage of committed capital. This is where the scale down arrangement comes into play. There has been a decrease in the management fees in recent years due to a number of economic factors. In particular, some funds are more likely to have a fixed fee of 2% of the fund’s assets, which is paid annually for five years and then decreases by 25 basis points for the next five years. Other fund managers will allow reductions of the fixed fee based on a change from committed capital in years one to five to net invested capital in year six to ten. Given these changes, a substantial proportion of buyout firms’ median compensation has been reduced to 12% of the committed capital.

Another example of convergence in contract terms is the profit distribution provision for the preferred return (or hurdle rate) arrangement. Due to this arrangement profits can only be distributed to fund managers after a certain profit threshold – a minimum annual internal rate of return – has been met. The preferred return rate ranges from 3% to 25% in European limited partnership (buyout) agreements with a median of 8%. As already mentioned, preferred return provisions are increasingly included in the limited partnership agreements in the U.S. Note that the median preferred return rate in the U.S. is also 8%, ranging from 7% to 20%. The convergence is reflected in the introduction of such a provision by the Institutional Limited Partners Association (ILPA) (www.ilpa.org).

These profit distribution arrangements are attempts to maximize fund managers performance, which means that profits can only be distributed to the managers after a certain threshold has been reached. Clearly, these arrangements, which as mentioned earlier, usually require private equity managers to first provide a preferred return to the investors before being able to distribute the “carry,” significantly reduce the chance that managers receive more than their fair share of the profits. In order to keep the managers focused and incentivized, most priority returns arrangements have a catch-up provision, which permits a reallocation of the profits to the general partner after the priority return has been distributed to the investors. In short, the catch-up provision entitles fund managers to receive most of the profits until the contractually agreed profit-split between the investors and the managers has been reached.
Certainly, if at a later stage it transpires that the general partners have received more than their fair share of the profits, investors will be entitled to call upon the clawback provisions under which the managers have to pay back the excess carried interest distributed earlier.

The question is whether we can expect to see more dramatic revisions to the fee and profit distribution provisions, rejecting the long-standing contractual practices. We will address the critical issues involved with this question in the next section.

3.2 Post-financial crisis trends in private equity contracting

It follows from the previous section that private equity investors are increasingly successful in finding ways to improve management fee and profit distribution arrangements [Jannarone (2012)]. This is confirmed by empirical evidence, which finds that 59% of the investors who were interviewed in June 2013 indicated that investor-favorable fee and profit arrangements were on top of their priority lists [Preqin (2013c)]. One year earlier, this percentage was still 68%, which is an indication that private equity investors are increasingly successful in negotiating more favorable terms. Moreover, our work shows that an increasing number of investors still seek a higher degree of control and bargaining power over the terms of the limited partnership agreement or attempt to avoid the payment of management fees altogether [McCahery and Vermeulen (2014)].

Consider Blackstone’s Tactical Opportunities, which is a portfolio of separately managed accounts [Witkowsky (2012a)]. Separate accounts, for example, are different from the organization of traditional funds in that an investor’s capital contribution will only be invested in accordance with its specific investment strategies and interests. From the perspective of the limited partners the benefits are twofold. First, separate account arrangements are flexible in the sense that they are usually tailored to the investors’ risk appetite and diversification needs. Second, and perhaps more importantly, it is obvious that the direct and close relationship between a single limited partner and a private equity firm enables investors to bargain for better terms and conditions, including “disruptive” and investor-favorable management fees and carried interest provisions [Kelleher (2013); Canada (2013)]. It is, therefore, to be expected that institutional investors will be more inclined to invest in separate accounts in the future.
For instance, 19% of institutional investors surveyed have set up a separate account arrangement in 2013, a significant increase compared to the 7% of the investors who invested through a separate account arrangement in 2012 [Preqin (2013a)].

It could, of course, be argued that these separate account arrangements should be viewed as a style or strategy that will gradually disappear in the aftermath of the financial crisis [Moix (2004)]. But since segregated accounts appear well designed to reduce investors’ risks, it is unlikely that investors will change their preferences so long as the structure meets their needs. The evidence is consistent with this view: 64% of investors that currently invest in separate accounts believe that these arrangements will be a permanent feature in the relationship with fund managers [Preqin (2013a)]. Particularly, small and medium investors, such as many family offices and wealthy individuals, are increasingly being attracted to separate account arrangements [Myles (2013)]. Unsurprisingly, therefore, these arrangements promptly gain popularity (see Figure 3). According to a recent survey of about 100 private equity fund managers based in the U.S., Europe and Asia, 7% have already set up separately managed accounts, with another 10% planning to do so in the next five years [State Street (2013)]. This also explains why lawyers and advisors are attentive to separate accounts. In order to avoid conflicts of interests between the separately managed funds and the traditional (co-mingled) funds, they help in setting out clear investment and disclosure policies for the separate accounts [Dai and Canada (2013)].

---

5 Please note that separate accounts arrangements are already common in the area of hedge funds and real estate funds. State Street (2013) shows that 35% of hedge funds managers have set up separately managed accounts, with another 22% indicating that they intend to follow suit. For real estate funds these percentages are 26% and 16%, respectively.
These findings point to investors’ search for greater control over both the investment decisions of general partners and the negotiations of the fund terms, which has also led to an increase in direct investments in private equity opportunities. We would like to stress that the uncertainties and information asymmetries often deter institutional investors from investing directly in high growth companies.

**Figure 3: The increasing popularity of separate accounts**

Source: Preqin Funds in the Market
We observe, however, that in an attempt to make investments in the best performing companies more lucrative (without the need to negotiate thorny compensation arrangements), there are increasing private equity deals where institutional investors piggyback on the due diligence and selection efforts of their fund managers by demanding the ability to pursue a co-investment strategy. Recent studies show that co-investments have gained in popularity. A study by Dineen (2012) indicated that co-investment rights provisions are already a must-have for institutional investors. Empirical evidence supports this view: (1) 32 out of 100 investors invest as co-investors with their fund manager and (2) 97% of these investors expect that they maintain (44%) or even increase (53%) their co-investment strategies [Preqin (2013a)]. A study by Dineen (2012) indicated that co-investment rights provisions are already a must-have for institutional investors. Empirical evidence supports this view: (1) 32 out of 100 investors invest as co-investors with their fund manager and (2) 97% of these investors expect that they maintain (44%) or even increase (53%) their co-investment strategies [Preqin (2013a)].

It is also likely that investors’ growing demands for greater control over their portfolios will lead to a greater use of alternative fund structures, such as pledge funds [Blake and Judd (2012)]. Pledge funds offer investors the opportunity to make investment decisions on a deal-by-deal basis. In order to get access to investment opportunities, investors must pay an annual fee and carry is calculated based on the amount invested. Although admitted investors can review potential portfolio companies, they are usually not obliged to participate in the deal. If the managers receive sufficient commitments from the “member investors,” they can prepare and negotiate the deal documents on behalf of the fund. In most cases a separate limited partnership is set up to make the investment. Again, the advantages are clear. Besides the greater control over portfolio acquisitions, the pledge fund alternative also provides investors with the possibility to avoid high management fees, as they pay lower fees on uninvested capital, and carried interest. The downside is that the newly designed pledge funds structures usually come with higher transaction costs [Jesch (2010)], and may suffer from lower returns on investment due to lack of diversification and cherry picking problems by general partners.
New private equity models: how should the interests of investors and managers be aligned?

So far we have focused on the empirical research that shows that investors are taking an increasingly active approach in their negotiations about compensation terms of the limited partnership agreement. But there is more. Besides demanding ‘improvements’ to the limited partnership agreement, we observe that investors want to see more skin in the game from the managers/general partners. Recall that the industry norm is 1% to 3% of the committed capital. By requiring that the general partners make significant capital contributions to their own funds, the investors can reasonably expect that the managers’ interests are better aligned with theirs. It appears that ‘higher-than-average’ capital commitments (ranging from 5% to 10%) are gradually becoming the norm in the private equity industry [Witkowsky (2012b)]. This view is not only confirmed by our data (see Table 3), but also supported by a poll conducted among general partners during the Private Equity International’s Forum in January 2011 (see Table 4) [Mitchenall (2011)].

Table 3: Skin in the game in buyout agreements (vintage 2005–14)

<table>
<thead>
<tr>
<th>Region</th>
<th>Agreements in dataset</th>
<th>Average GP commitment to fund (%)</th>
<th>Median GP commitment to fund (%)</th>
<th>Minimum GP commitment to fund (%)</th>
<th>Maximum GP commitment to fund (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.U.</td>
<td>115</td>
<td>4.14</td>
<td>2</td>
<td>0.4</td>
<td>48</td>
</tr>
<tr>
<td>U.S.</td>
<td>187</td>
<td>4.76</td>
<td>3</td>
<td>0.1</td>
<td>40</td>
</tr>
</tbody>
</table>

Source: Preqin Fund Terms Advisor

Table 4: Skin in the game in the private equity industry

<table>
<thead>
<tr>
<th>Contribution as % of committed capital</th>
<th>Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1%</td>
<td>21%</td>
</tr>
<tr>
<td>1%-2%</td>
<td>28%</td>
</tr>
<tr>
<td>2%-5%</td>
<td>35%</td>
</tr>
<tr>
<td>&gt;5%</td>
<td>16%</td>
</tr>
</tbody>
</table>

Source: Private Equity International
New private equity models: how should the interests of investors and managers be aligned?

4. Conclusion
This paper analyzed the regulatory alternatives for the private equity industry. We discussed important regulatory checks that enhance investor protection and support alignment of interests between GPs and LPs. In this context, we examined some of the recent proposed changes introduced by the Alternative Investment Fund Managers Directive in Europe.

Having explored the benefits and costs of a regulatory regime, we then turned to review the typical contractual measures in place, designed to align investor and fund manager's interests. This paper argued that the fundraising landscape has evolved and changed significantly in the post-financial crisis era, leading to “new” limited partnership compensation arrangements. The evolution of “new” limited partnership terms can largely be attributed to the major challenges in private equity fundraising and returns.

In response to increased LP bargaining power, this paper has identified two strategies that are currently deployed by private equity fund managers. The first strategy relates to the negotiation of improvements to the “two and twenty” compensation arrangements in the traditional limited partnership agreement. The adoption of scale down provisions and preferred returns provide substantial protections to potential investors. The second strategy involves the creation of “innovative” private equity structures that constrain fund managers’ discretion to make investment decisions. Greater customization through separate accounts arrangements and deal-by-deal investment opportunities will significantly increase investors’ bargaining power and the possibility to negotiate more effective compensation arrangements. Together, these results are consistent with the view that the “end” of the traditional “two and twenty” fee structure will arguably have a more positive impact on the private equity industry than the need to comply with the cumbersome AIFMD rules.
New private equity models: how should the interests of investors and managers be aligned?

References
Canada, H., 2013, “Fee cuts aside, PE firms have much to gain from separate accounts,” Wall Street Journal (Private Equity Beat), 17 September
Myles, D., 2013, “How fundraising models will bifurcate PE market,” IFLR, 11 September
Mulcahy, D., B. Weeks, and H. S. Bradley, 2012, “We have met the enemy... and he is us: lessons from twenty years of the Kauffman Foundation's Investments in venture capital funds and the triumph of hope over experience,” Ewing Marion Kauffman Foundation, May
New private equity models: how should the interests of investors and managers be aligned?

Preqin, 2012b, “Preqin special report: European private equity,” March
Preqin, 2013a, “Preqin investor outlook: alternative assets” H2 2013
Preqin, 2013b, “2013 Preqin private equity fund terms advisor,” September
Preqin, 2013c, “Private equity spotlight,” October
Private Equity International, 2013a, “Flocking to quality,” Friday letter, 19 April
Private Equity International, 2013b, “Blind pool blindness,” Friday Letter, 8 March
State Street, 2013, “The next alternative: thriving in a new fund environment,”
Wang, W., 2013, “AIFMD compliance remains exclusive,” Global Custodian, 22 July
Winfrey, G., “Fundraising on course for post-crisis record in 2013,” Private Equity International, 2 October
Witkowsky, C., 2012b, “Fundraising GPs put more ‘skin in the game’,“ Private Equity International, 14 November
Auditing estimates: what will the future bring?

Paul de Beus
Partner, EY, The Netherlands

Maarten Koning
Partner, EY, The Netherlands

Abstract
Under Solvency II, national competent authorities are requesting audit procedures to be performed on all supervisory reporting. Reports will contain more information based on projections of future cash flows. The level of assurance from the Solvency II external audit is expected to be similar to the opinion provided by auditors on financial statements, i.e., reasonable assurance. Current guidelines represent only minimum requirements. So, when is good, “good enough” for the external auditor? In this article, we suggest to analyze the (audit) findings on model outputs against a reasonable range. And in specific situations, this range can exceed the traditional “tolerable error” (TE). We translate the qualitative assessment of a reasonable range and make it quantifiable and objective. We will do so by defining a tolerable range based on sensitivities of inherent variability of underlying key risk drivers. This is needed because audit materiality and TE measures are mainly focusing on “errors” (or misstatements), and are often based on less substantiated and sophisticated approaches. However, when using model outputs in which key model design and parameter choices have to be made, and cash flows are projected long into the future, it is quite often not a question of simply being “right or wrong.”
1. Introduction
In June 2013, the European Insurance and Occupational Pensions Authority (EIOPA) issued its second “unofficial” pre-consultation paper concerning “Proposals on Guidelines on External Audit.” The proposed guidelines require the statutory auditor to perform an audit on quantitative and qualitative elements of the Solvency Financial Condition Report (SFCR) of insurance companies. In particular, the Solvency II balance sheet, own funds (including the eligibility and tiering of fund items), and minimum – and solvency capital requirements (MCR and SCR).

Not awaiting EIOPA final guidelines, the PRA in the U.K. released their Solvency II audit instructions to the market on 16 October 2014. Likewise, the National Bank of Belgium (NBB) is requesting audit procedures to be performed by the insurance company’s statutory auditors on all supervisory reporting within the scope of the Solvency II interim measures.

Reporting for internal and external purposes will contain more information about projections of future cash flows. The level of assurance from the Solvency II external audit is expected to be similar to the opinion provided by auditors on financial statements, i.e., reasonable assurance. The (proposed) guidelines represent only minimum requirements. So, when is good, “good enough” for the external auditor?

2. Solvency II audit
In the audit of a Solvency II economic balance sheet and risk-based capital requirements, the auditor evaluates the outcome of complex models and complicated processes. The model outputs are often the result of various model choices, (risk) data vendor selection and significant assumptions (such as scenarios of future events or stochastic model specifications), among others. Auditors will need to involve actuarial experts in the audit.¹

¹ International Standards on Auditing (ISA) 620, ‘Using the work of an auditor’s expert.’
A model outcome, or in audit terminology “the accounting estimate,” is highly sensitive to assumptions used and, therefore, subject to high estimation uncertainty. As a consequence, the range of possible reasonable outcomes can be quite wide. This range could be much wider than the reporting materiality threshold applied for auditing the Solvency II balance sheet, taken questions should be answered:

- Is it acceptable to apply a range of reasonable outcomes that is broader than the traditional audit materiality?
- Can we develop a method to define reasonable ranges in case of high estimation uncertainty?

In this article, we aim to instigate a wider discussion on this topic in the insurance sector, and to advocate for a closer working relationship between auditors and actuaries. We will develop an audit approach for quantifiable and justifiable derivation of reasonable ranges for model estimations in balance sheet items, where a simple “right” or “wrong” cannot be determined with classical audit methods. But first we start by providing some background on the concept of audit materiality.
Auditing estimates: what will the future bring?

**Box 1: Audit guidelines**

The “Proposals on Guidelines on External Audit” issued by EIOPA require that auditors will place reliance on the model validation process, since it is part of the supervisory approval. Such reliance is subject to the auditor's assessment of the effectiveness of the internal controls related to the internal model validation process. The Solvency II model validation standards, however, do not give materiality guidance for the evaluation of the model validation findings and results, nor does it focus on the Solvency II balance sheet and application of the “standard formula.” The guidance provided is more from a technical and governance model validation perspective; for example, requirements around processes, controls and calibration. Validation reports often conclude with sensitivity analyses or an extrapolation of a sample test outcome. But when are these results within an acceptable range?

The scope in the PRA (U.K.) guidelines\(^2\) includes all “internal model” firms plus the larger firms that apply the so-called “standard formula” under Solvency II. They require a “reasonable assurance” opinion on the balance sheet, technical provisions (excluding risk margin) and own funds as at 31 December 2014. This opinion needs to be from an external audit firm – not necessarily from the external auditor. To prepare for this audit, firms are first requested to provide a “review and recommend” report on their preparedness to implement the Solvency II regulatory framework.

NBB is also requesting audit procedures to be performed by the insurance company’s statutory auditors on all supervisory reporting within the scope of the interim measures. Aware that entities are still preparing for Solvency II, the NBB does not request a standard audit opinion on this quantitative and qualitative information, but a detailed report of findings. This should provide the regulator with a comprehensive overview of the level of readiness for Solvency II in the entire industry, as well as differences in practices and interpretation of the rules.

3. Background on audit materiality

With regard to the question of whether it is acceptable to apply a range of reasonable outcomes, which is greater than the traditional audit materiality, consideration should also be given to what is written in the current auditing standards (ISAs: International Standards on Auditing). We discuss two standards: ISA 320 and ISA 540.

ISA 320,\(^3\) which provides a general guidance, stipulates that in auditing financial statements, the overall objective of the auditor is to obtain reasonable assurance about whether the financial statements as a whole are free from material misstatement. Misstatements are considered to be material if they, individually or in aggregate, could reasonably be expected to influence the economic decisions of users taken on the basis of the financial statements. In addition, ISA 320 indicates that to determine the audit materiality, a percentage is often applied to a chosen benchmark. Examples include a percentage of “profit before tax” and a percentage of “equity.”

When applying this concept to a Solvency II balance sheet, we expect to see the audit materiality set at a percentage of “own funds.” Based on this threshold, a TE is defined. The TE is the maximum misstatement amount(s); set by the auditor at less than audit materiality for the financial statements as a whole. This is done to mitigate the probability that the aggregate of uncorrected and undetected misstatements exceeds audit materiality.

ISA 540\(^4\) more specifically discusses the issue described above. It recognizes that developing “a range” to evaluate management’s estimates may be an appropriate audit response. It requires that range to encompass all “reasonable outcomes” rather than all possible outcomes (the latter would be too wide to be effective for purposes of the audit). Also, ordinarily, a range that has been set to be equal to or less than TE is adequate for the purpose of evaluating the reasonableness of management’s estimate.

---

\(^3\) ISA 320, “Materiality in planning and performing an audit”.

\(^4\) ISA 540, “Auditing accounting estimates, including fair value accounting estimates, and related disclosures”.
However, the standard continues to say that, particularly in certain industries, it may not be possible to narrow down the range to below the TE amount, and that this does not necessarily preclude recognition of the accounting estimate. It may indicate, however, that the estimation uncertainty associated with the accounting estimate is such that it gives rise to a significant risk. In such cases, the auditor’s further substantive procedures are focused on evaluating how management has assessed the effect of estimation uncertainty and the adequacy of the related disclosures.

In this regard, the standard describes that management may evaluate alternative assumptions or outcomes of the accounting estimates through a number of methods, such as undertaking a sensitivity analysis. Please note that, in addition to business and risk management purposes, certain accounting frameworks require disclosure of this information, such as IFRS4 and IFRS7.5

Based on the above guidance, it can be concluded that it would be necessary to analyze the (audit) findings on model outputs against a reasonable range. And in specific situations, like for the financial services industry, this range can exceed the traditional TE. The challenge we face now is: how to translate the qualitative assessment of a reasonable range, and make it quantifiable and objective? We believe that we can take a step toward such an assessment by defining what we refer to as a “tolerable range” based on sensitivities of inherent variability of underlying risk drivers. We believe we need to add this concept because audit materiality and TE measures are mainly focusing on “errors” (or misstatements). Thresholds for acceptable ranges are often based on less substantiated and sophisticated approaches (e.g., a certain percentage of the booked reserve). However, when using model outputs in which key model design and parameter choices have to be made, and cash flows are projected long into the future, it is quite often not a question of simply being “right or wrong.”

4. Assessing models
When assessing (the output from) a valuation or risk model, the auditor should, among others, evaluate the model's theoretical soundness, mathematical integrity and appropriateness of the model parameters (ISA 540). These are common elements as part of a model validation process, and are also referred to as follows:

- **Theoretical soundness**: representativeness or fit for purpose
- **Mathematical integrity**: arithmetical correctness
- **Appropriateness of the model parameters**: objective measurability of parameters

We will initially focus on these elements in order to be able to define a tolerable range in a quantifiable and objective measure. As will be described later in this article, just doing this is not sufficient to form the auditor’s opinion — similar to model validation being much more comprehensive than merely looking at these three elements.

In assessing the representativeness of input and methodology the verification focuses on whether the model and input used are “fit for purpose” (e.g., interest curve used, basis spread correction) and whether the methodology is widely used and is a market (leading) practice. The arithmetical correctness can be tested, for example, by parallel modeling using the same input as that used for the assessment value or even parallel modeling using underlying contractual input and market information obtained independently. The objective measurability of input parameters can be determined by reconciliation of static input data with underlying contracts and/or systems and reconciliation of market input data with objective systems (data providers, such as Bloomberg and Reuters). By assessing model outputs using these criteria, assessing one of these criteria often implicitly contains a (partial) assessment of the other elements.

To determine the arithmetical correctness and representativeness of a valuation or risk model, the analysis could be based on the difference between the model output under a parallel replicated model, alternative method(s) and the assessment value.
We appreciate that parallel modeling is not always feasible in full, but model points and/or benchmark portfolios and/or parameters can provide sufficient insight for audit purposes. A conclusion can be drawn based on objective measures, such as:

- The absolute difference
- The relative difference
- The absolute difference expressed as a percentage of the sensitivity of the assessment value to changes in key assumptions/parameters

The objective assessment measure selected depends on the goal of the audit procedures performed. For example, in validation procedures for financial derivatives valuation, we generally use the third method described above. This provides a sufficient level of comfort on the arithmetical correctness and, to a certain extent and as a bonus, also the representativeness and even objective measurability of the derivative valuation method.

5. **DV01: dollar value of 1 basis point**

The objective measure based on the absolute difference expressed as a percentage of the sensitivity of the assessment value to changes in key assumptions/parameters is generally referred to as the DV01 method. A DV01 is the sensitivity of the value of a financial instrument, like a derivative, to a 1 basis point (bp) change (parallel shift) in the underlying interest curve. Mathematically, DV01 estimates the “first derivative” to the interest rate of the valuation function.

Looking at a DV01 for interest rate derivatives makes sense. Many derivatives are based on a fictive notional, meaning that the notional is not transferred but used as reference to base the coupon payments on, and are structured on large (implicit or explicit) long and short positions. For example, an interest rate swap (IRS) has a value of nil at issue date because the (large) notional/value of the floating and fixed leg are equal. After the issue date, the fair values of the floating and fixed leg change due to remaining maturity and changes in underlying market interest rates.
The fair value of the IRS is the net value of both legs. This net value is, therefore, very sensitive to the underlying interest rate curve. Hence, differences with other models and/or parameters can easily occur, but are not necessarily an error, and the auditor wants to know when “good is good enough” for audit considerations.

Differences can be explained by varying definitions and/or interpretations and/or choices of day-count conventions, inter- and extrapolation, choice of data vendors to extract interest rate curves from etc. When auditing the value of interest rate derivatives, 3 times DV01 could be an acceptable range, the tolerable range, for the setting of a materiality threshold. This “3” is the inherent variability and chosen for European IRS based on the following rationales:

- Movements of 3 bps to 5 bps during a business day are normal for euro IRS rates.
- Differences of 2 bps to 4 bps at a point in time is seen for euro IRS rate between different data vendors.

This inherent variability is the objective range in which the risk driver lies “as a fact of life” and can be seen as the intuitively reasonable range that is not neither right nor wrong by definition. Hence, a tolerable range based on valuations from 3 bps change to interest rates could be reasonable to audit the assessment value. This means that when the difference between the assessment value and the output of the parallel model is below 3 times the DV01 of the parallel model, this difference would not be considered as being material.

With this in mind, we are able to set a materiality threshold objectively for the purpose of mathematical correctness. Note that this measure will be proportional to the underlying assessment value. Indeed, in the New World where many reports are based on model output, we believe that materiality is an important consideration in model risk management, and a proportionate risk-based approach is practically sensible.

---

6 In this example, we assume euro denominated derivatives; for other currencies and/or other complex derivatives with, for example, Steepener features, the DV01 and/or TR might not be appropriate.
This means that when the auditor determines the materiality for the object on which assurance is expressed, the tolerable range should be based on different materiality considerations per model, per risk driver, per usage and per disclosure; like the case for TE to reduce the probability to an appropriately low level that the aggregate exceeds materiality.

To illustrate the concept of DV01, let us consider an IRS example by assessing the value of a €100 million notional receiving 2.5% fixed/pay floating swap. The assessment value equals €1,492,594.

We determined a value using a parallel model of €1,255,579, resulting in a difference of €237,015.

The auditor has a tolerable error threshold of €150,000. Remember, this is the “traditional” threshold in terms of error for reporting purposes. Is the difference material?

In this example, we probably would not conclude so. The DV01 is determined as €81,405, and 3 times DV01 at €244,215 (TR), which is higher than the difference between the assessment value and the value determined using a parallel model.

By looking at the inherent variability of the risk driver, we also need to look at the objective measurability of the underlying risk driver. And the parallel model used to assess the audit estimate is not necessarily the exact same model as that used for the assessment value. One could have used a stochastic model to value a European equity option whereas this parallel model could be built on, for example, a binomial tree or closed form Black-Scholes formulas when these are also market practice for the valuation of the option. So, when determining the arithmetic correctness and the representativeness (fit for purpose), the model outcomes are only “as good as the input” - the parameters (or risk drivers) needed in the valuation function or in the risk function. Inherently, there is uncertainty around the parameters chosen. In setting a parameter, there is often expert judgment involved. As long as the parameter is within a reasonable range defined by the expert, there might not be a misstatement issue.
As part of the audit procedures, it should be determined whether the expert uses all relevant data and should be able to substantiate their choice. This is a governance discussion that continues to be crucial in audits. To what extent are policies set, controls on assumption setting available and effective, model choices documented and approved etc. We will elaborate on this further in the section describing “the audit of the estimation process.”

Can we apply this concept to assess acceptable ranges for the valuation and risk measurement of insurers businesses as well?

6. DVOC method
As previously mentioned, a DV01 model where interest rate is the key parameter is defined as the difference in value from the model as a result of a 1 bp change in the yield curve. A DV01 model where volatility is the key parameter could be defined as the difference in value from the model by changing the volatility by 1% (i.e., estimating the “Vega”). Next, we can determine the inherent variability of (implied) volatilities to set the multiplier for that “DV01.” The same could be generalized for some noneconomic parameters (risk drivers) like lapses, mortality, etc.; hence, looking at the (estimated) set of (mathematically spoken) “first partial derivatives” of the valuation (or risk metric) function.” We have defined these sensitivities based on (inherent) variability as Delta Value of Change (DVOC).

Insight in the DVOC has clear advantages. One advantage is that it allows us to determine the key parameters driving the majority of the model output. Experience has shown that more than 95% of the results, in terms of an economic value like “solvency own funds” or a risk metric like “solvency capital requirement,” can be explained by only a limited number of input parameters – these are the function’s key risk drivers.

---

7 In mathematics, a partial derivative of a function (in this article the function is the model used to estimate the economic value or risk metric) of several variables is its derivative with respect to one of those variables, with the others held constant. Partial differentiation is the act of choosing one of the lines to every point on this surface and finding its slope. The slope of this line is a measure of the sensitivity of the functions outcome around the input parameter.

8 This has already been acknowledged in the evolution of embedded value principles for life insurers over the years where sensitivities have to be disclosed.
Auditing estimates: what will the future bring?

Another advantage is to continue the analysis by using these key risk drivers to determine and analyze their (inherent) variability. This can and should be used to determine the acceptable range of resulting outcomes—the tolerable range for audit purposes by setting objectively a multiplier of DVOC as that range (like the “3” for DVO1).

We believe there are other advantages to using DVOC—one being for “analyses of changes” (or “movement analyses”). We will describe this later in the article.

Like for swaps, the value of an insurance liability is often also at net value: the difference between the discounted value of future premiums and the discounted value of future benefits. Likewise for risk metrics. Therefore, we see the benefit of generalizing the concept of DVO1 to DVOC.

We will use two case studies in the remainder of this article to elaborate on the concept of DVOC and to show some remaining challenges. We will cover both assets and liabilities, as well as the solvency capital requirement (based on Solvency II standard formulae). In the first case study we will look at a stylized life insurance company and its capital position. Based on Solvency II principles, we will determine a market value balance sheet and its solvency capital requirement. Because we know upfront that the figures are dependent on choices (preferences?) in parameter setting and inherent variability of various risk drivers, we will apply the DVOC concept to this particular audit example.

This means that we want to gain insights into the key parameters (those that matter the most), the objective range of parameter possibilities (for example, based on the inherent variability of these parameters) and hence a measure of model materiality, the tolerable range.

The second case study covers a non-life insurer for which the estimation process seems to have a much wider range of sensitivities; hence, additional audit procedures need to be performed. For this case study we limit the example to the economic values of the insurers liabilities.

---

9 Like the solvency capital requirement as a surplus-at-risk measure, the surplus is the net difference between all market valued assets and liabilities.
7. Case study 1: life insurance
We consider the insurance company Solve & See Life, that has an insurance book consisting of the following two products:

- Annuities: insurance contracts consisting of annual payments to the policyholder for as long as the policyholder is alive; the annuities are financed by a single premium payment by the policyholder at the start of the insurance contract.
- Term insurance: life insurance contracts that pay out in case of death (whole life coverage); the whole life insurance contracts at Solve & See Life are financed by a single premium payment by the policyholder at the start of the contract as well.

Table 1 presents the base assumptions used for valuation of the life insurance liabilities by Solve & See. Please note that in this example, for simplification purposes, the risk margin has been set to zero and annual expenses are not included in the best estimate calculation.

For discounting the future payments, we used the yield curve as published by the Dutch National Bank. This yield curve is based on European swap rates for maturities up till 20 years. For maturities above 20 years, the yield curve is extrapolated by converging the one-year ultimate forward rate to 4.2% over a 40-year period.

Table 1: Valuation assumptions of the life insurance liabilities of Solve & See

<table>
<thead>
<tr>
<th></th>
<th>Annuity</th>
<th>Term insurance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality forecast</td>
<td>AS 2012-62</td>
<td>AS 2012-62</td>
</tr>
<tr>
<td>Mortality experience factor</td>
<td>ES-L2 10</td>
<td>100%</td>
</tr>
<tr>
<td>Life expectancy at 67 male/female</td>
<td>21.4 / 22.5 years</td>
<td>18.6 / 21.1 years</td>
</tr>
<tr>
<td>Yield curve</td>
<td>EUR Swap curve with UFR 31 Dec 2013</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 presents the base assumptions used for valuation of the life insurance liabilities by Solve & See. Please note that in this example, for simplification purposes, the risk margin has been set to zero and annual expenses are not included in the best estimate calculation.

For discounting the future payments, we used the yield curve as published by the Dutch National Bank. This yield curve is based on European swap rates for maturities up till 20 years. For maturities above 20 years, the yield curve is extrapolated by converging the one-year ultimate forward rate to 4.2% over a 40-year period.

---

10 Mortality experience based on observations in Dutch insurance market for annuities during 2007–09. ES-L2 (Ervaringssterfte Lijfrente 2) Mortality experience Annuities 2. The mortality assumptions (and acronyms used in the table) will be discussed further.

11 DNB curve including UFR
Method yield curve: http://www.toezicht.dnb.nl/5/19/10/50-205324.jsp
Yield curve: http://www.statistics.dnb.nl/usr/statistics/excel/t1.3nm.xls
7.1 Mortality assumptions
For the chosen products, it seems clear that mortality rates and interest rates are the key risk drivers in the valuation of these products. So we start by identifying the inherent variability of mortality range in particular (for interest rate we refer to the DV01 section), as the first step to determining the tolerable range. In recent years, there have been initiatives to securitize longevity and mortality risk through, for example, the construction of longevity bonds, mortality swaps or “qx forward/future” contracts.\(^\text{12}\)

However, currently the market volume of these mortality-related investment products is not yet sufficient to derive (local) market prices for mortality rates. Objective measurability of used mortality assumptions by reconciliation to quoted market data is therefore not possible, contrary to, for example, swap rates.

Within an audit where the focus is on mortality assumptions, an actuary would typically use the base table mortality rates and future mortality improvements available. The base table is derived from general population data adjusted to reflect the portfolio’s current mortality experience, while future mortality improvements are obtained from industry or countrywide forecasts.

After analyzing “equally valid” mortality tables and approaches for mortality improvement factors, we conclude that the inherent variability is equivalent to a change of 5% mortality rate improvement and equivalent to the difference between two mortality tables under inspection.

So, let us have a look at some numbers. In our example, we assume, after testing, that the following model points are representative of the total life insurance portfolio of Solve & See:

- Male and female age 67, with annual annuity payments of 105
- Male and female age 67, with a term insurance coverage of 1,000

Auditing estimates: what will the future bring?

The assets on the balance sheet are listed equities, government bonds (coupon 3% and remaining maturity 10 years), and an existing 15-year 3% fixed/Euribor floating swap with a notional of 1,650 (all free from default, spread and concentration risk). Table 2 presents a (simplified) Solvency II balance sheet of the firm. The Solvency Capital Requirement based on Solvency II standard formula amounts to 562, resulting in a solvency ratio of 150%.

Figure 1 and 2 show the impact of the sensitivity of the different risk drivers on the balance sheet, as well as the Solvency II capital requirements.

**Figure 1:**
**Impact of change of risk driver on balance sheet compared to best estimate scenario**
Auditing estimates: what will the future bring?

Figure 2:
Impact of change of risk driver on capital requirements compared to best estimate scenario

Table 2: Market value balance sheet of Solve & See, year-end 2013

<table>
<thead>
<tr>
<th>Assets</th>
<th>Liabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity</td>
<td>Own funds</td>
</tr>
<tr>
<td>Bond</td>
<td>Life insurance liabilities</td>
</tr>
<tr>
<td>Swap</td>
<td>Annuity</td>
</tr>
<tr>
<td>Term insurance</td>
<td>3,517</td>
</tr>
<tr>
<td>Term insurance</td>
<td>1,166</td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
</tr>
</tbody>
</table>

| Equity      | 1,000                             |
| Bond        | 4,431                             |
| Swap        | 94                                |
| Term insurance | 1,166                             |
| Total       | 5,525                             |
| Total       | 5,525                             |
Auditing estimates: what will the future bring?

Table 3: Solvency capital requirement Solve & See, year-end 2013

<table>
<thead>
<tr>
<th>SCR</th>
<th>562</th>
</tr>
</thead>
<tbody>
<tr>
<td>Market risk</td>
<td></td>
</tr>
<tr>
<td>Equity risk</td>
<td>465</td>
</tr>
<tr>
<td>Interest rate risk</td>
<td>53</td>
</tr>
<tr>
<td>Diversification</td>
<td>-24</td>
</tr>
<tr>
<td>Total market risk</td>
<td>493</td>
</tr>
<tr>
<td>Life risk</td>
<td></td>
</tr>
<tr>
<td>Longevity</td>
<td>178</td>
</tr>
<tr>
<td>Mortality</td>
<td>38</td>
</tr>
<tr>
<td>Catastrophe</td>
<td>1</td>
</tr>
<tr>
<td>Diversification</td>
<td>-45</td>
</tr>
<tr>
<td>Total life risk</td>
<td>172</td>
</tr>
<tr>
<td>Market risk</td>
<td>493</td>
</tr>
<tr>
<td>Life risk</td>
<td>172</td>
</tr>
<tr>
<td>Diversification</td>
<td>-104</td>
</tr>
<tr>
<td>SCR</td>
<td>562</td>
</tr>
<tr>
<td>Solvency ratio</td>
<td>150%</td>
</tr>
</tbody>
</table>

The reference point is the risk dashboard in table 3, for which the base assumptions (BE = best estimate based on the mortality table from the Dutch Actuarial Society - AS 2012-62) are used as described at the start of this case study.
Box 2: Inherent variability of mortality rates

Base table mortality rates
The base table represents the current mortality rates for the insurance portfolio. Starting point for this base table is general population data for the country in which the insurer operates, as these general population mortality rates are based on the largest sample possible for that region. The population in the insurance portfolio is a subset of the general population and can have different mortality characteristics than the average person. Consumers who buy immediate annuities are often healthier than the general population (and therefore have lower mortality rates). The difference in mortality rate is expressed in a mortality experience factor, based on experience in the portfolio. An experience factor of, for example, 80% means that the mortality rates of the policyholder is 20% below general population.

Mortality experience factors are estimated based on an experience analysis of the observed deaths in a portfolio relative to expected deaths according to the general population table. The uncertainty in these experience factors depends on a number of factors including the size of the portfolio, the observation period chosen and the extent of detail of the experience rates (one factor, or gender, and/or age dependent).

As an indication of the sensitivity of the market value of life insurance liabilities, we have analyzed a parallel relative shock of experience rates by 1% (i.e., where a mortality factor of 80% on general population mortality rates is applied, this is shocked to 80.8%) and the relative impact on the market value. The difference, or DVOC, is shown in Figure B1. The result can then be used for setting a tolerable range depending on portfolio specific uncertainty of experience rates - the inherent variability.

For life annuities, the difference, or DVOC, of the parallel shock on the market value of the liabilities increases with age; while for term insurance, the difference decreases by age.

Auditing estimates: what will the future bring?
Auditing estimates: what will the future bring?

Figure B1: DVOC for the market value of liabilities as a function of age in case of 1% relative shock (decrease) on mortality experience factors
Future mortality improvements
So, what is the inherent variability for mortality experience factors? There is a wide variety of methods to forecast future mortality improvements.\(^\text{13}\) These methods can be divided into three general approaches:

1. Expectation of mortality development
2. Extrapolation of historical mortality development
3. Explanatory approach by using causes of death

Given the wide variety of methods and approaches available, there are multiple (equally acceptable) mortality forecasts for the same population that all have a valid rationale but generate different results. Analysis shows that these methods lead to differences in mortality improvement assumption over 1%. Differences in the base assumed mortality rates (different mortality tables that are available to the insurer) should also be investigated. The knowledge of both analyses can then be used to determine the inherent variability of the underlying risk driver (i.e., mortality assumptions), and hence the factor to multiply the DVOC with to determine the tolerable range for the value of the liability.

In several countries, the local society of actuaries have set up mortality forecast models; some with variable parameters and variable output (for example, CMI in the U.K.)\(^\text{14}\) and others with fixed parameters and output (for example, the Dutch Actuarial Society).\(^\text{15}\) In addition to these forecasts, other projections are available from government institutes like the Bureau of Statistics (CBS).


\(^{15}\) Antonio et al., 2012, Prognosetafel 2012-2062, Actuarieel Genootschap.
As previously mentioned, inherent variability could, therefore, be seen as variability among the different mortality rate forecasts and subsequently the tolerable range as the relative spread of the liability value results using these different (generally accepted) population mortality forecasts. For example, we have analyzed the following mortality forecasts available in the Netherlands:

- CBS forecast 2012–60:¹⁶ projected mortality rates taking into account smoking related deceases and convergence to Western European mortality trend.

Both forecasts are updated every two years. As the Actuarial Society forecast is directly derived from observed mortality, we have also created forecast 2013–62 by extending the used observation period for the short- and long-term average to 2012.

Figure B2 shows the variability of the present value of the annuity and term insurance product for the different equally accepted mortality projections. The update of the Actuarial Society model with an extra observation year led to a slight increase in future mortality rates compared to the 2012–62 forecast. The impact has, therefore, been to decrease the present value of the annuity and increase the term insurance product. Further analysis showed that using an alternative forecast, the CBS 2012–60 forecast, led to a further decrease/increase in present value for the annuity/term insurance.

Auditing estimates: what will the future bring?

**Figure B2:** Cumulative DVOC for annuities (left) and term insurance (below), for impact of Actuarial Society model by new observations (yellow) and further change as a result of using a different accepted model/forecast (CBS forecast).
The following risk drivers are concluded to be key and hence, taken into account:

- **Mortality forecast:** the assumption for future mortality improvements is adjusted to other equally accepted forecasts, in this case CBS 2012–60; this should reflect the inherent variability on mortality forecast.

- **Mortality experience rates:** to reflect current mortality rates observed in the portfolio, general population rates are adjusted with experience factors. The uncertainty of these experience factors depends on a number of factors, like the number of observations used in the experience study. We analyze the impact of a relative 5% increase of all mortality rates and a 5% decrease of all mortality rates; this reflects the inherent variability on mortality experience rate assumptions.

- **Interest rate:** the yield curve is increased and decreased with three basis points; this should reflect the inherent variability of the use of the interest rates (instead of 1bp sensitivity shock and multiplying with 3 afterwards to show differences in this stylized example easier).

### 7.2 Analyzing the results

Graphically it is easy to see that SCR is relatively stable whereas the own funds are more sensitive. Hence the Solvency II ratio is sensitive. The impact of the mortality forecast on the market value of the liabilities amounts to -2% on the annuities and +2% on the term insurance, which is in line with analysis performed (see Box 2). As the annuities are the biggest part of the liabilities, the total liabilities decrease and the own funds increase by 5%. In the calculation of the required solvency capital however, the decrease in interest rate risk is offset by an increase in longevity risk, resulting in a relatively stable SCR (+0.5%). Still, the solvency ratio increases by 6% point.

The impact of a permanent increase or decrease in mortality rates, indicating possible inherent uncertainty in the current level of mortality rates, is similar to a change in mortality forecast, in this case of a 5% shock of mortality rates. This demonstrates the importance of statistical analysis on the experience factors used by an insurance company. The impact on liabilities is about 1% for both the annuity and term insurance portfolio (and in line with the analysis shown in Figure B1 in Box 2 for a 1% shock - 0.2%).
The 3 bp shock shows the interest rate sensitivity of both assets and liabilities. Note that the swap is, as expected, most sensitive to the interest rate shock (6% for the swap but only 0.3% for the insurance liability). Furthermore, we see that both mortality forecasts, using BE or CBS rates, are similarly sensitive to the interest rate shocks. This is not too surprising as the sensitivity is just the delta (the difference between assessment value and shocked value).

The main risk drivers in this example, per the balance sheet item, seems clearly to be mortality and interest rates. The mortality forecast available produces a solvency ratio range of 150%-156%, while the level of mortality leads in this case to a range of 144% -156%. The solvency ratio itself seems quite stable based on the DVOC for interest rate (or in this case DV01) - the swap seems to do its work quite well. These ranges could, therefore, be used to set the total tolerable range for the life insurance portfolio; the range could potentially be set to 144%-156% or even larger when combined with the DVOC for mortality forecast. However, like setting the tolerable error as part of total materiality, the tolerable range could and should also be translated using underlying key components of the total balance sheet to reduce to an appropriately low level the probability that the aggregate of uncorrected and undetected misstatements exceeds materiality.

Concluding that DVOC for SCR for interest rate is close to zero, should also trigger the question of whether the swap position is correct in the first place. The model tells us that if no swap was entered into, or the existing swap was agreed against less favorable conditions (e.g., at a time the fixed interest rate was, say 2.5% instead the current 3%) the SCR ratio would be much worse. In the base scenario, the SCR would be 130% when interest was agreed at 2.5% and even only 113% if no swap was entered into at all.
That is why it is important to also look carefully at the constituents of the balance sheet and the quality of data used (exposure and risk data assessed for completeness, accuracy and reliability). This might even have a larger impact on the ratio than the validity of the model itself. And this is not an inherent variability.

8. Case study 2: general insurance

In this case study we will evaluate the model materiality of claims reserves in property and casualty insurance. General insurance liabilities are known for their wide variety of underlying insurance risk types, ranging from medical expenses to environmental liability and natural catastrophe losses. The actuarial techniques used to estimate the claims reserves are also known for their extensive use of expert judgment. We will show why model assessment here has to focus on theoretical soundness and objective measurement of parameters, given the large material impact of model selection.

Let us analyze the company Solve & See P&C, with a book of business consisting of:

- Motor/truck liability commercial lines
- General third party liability (GTPL)
- Motor liability private lines
- Product liability
- Workers compensation

A general characteristic of liability claims is possible delays in the claims reporting, as well as long duration of claims handling.

Consequently, the claims reserves per event year can have a run-off lasting more than 20 years. The actuarial department of Solve & See uses a structured way to select the valuation model in three steps:

1. Optimized paid chain ladder: a run-off analysis on the cumulative payment data; for the two most recent event years, if they have developed less than 75% of the estimated ultimate claim size, the method switches to a loss ratio calculation derived from the older event years, applied to the earned premium of the current event year.
2. **Optimized incurred chain ladder**: the same analysis on the cumulative incurred claims data (paid plus outstanding case reserves), with the same switching regime.

3. **Final average**: the best estimate reserve is chosen as the nominal average of both results.

The approach here is just one example of widely used approaches. In practice, the model selection depends heavily on the reliability of the available data. Note that payment data are “exact,” but outstanding case reserves are estimates made by claims handlers that apply their expert knowledge combined with standardized methods. It is very common that these reserves are overestimated. Typically, actuaries take into account additional information from the claims department to interpret the data. For parameter setting, more expert judgment is applied, for instance, in smoothing the outliers, trend fitting and other selections. In addition, more detailed data can be used to estimate the claim numbers and the claim amounts separately, or more statistically validated regression techniques can be applied.

As stated previously, there is no common opinion on the best practice methodology choice. To illustrate the large volatility of model results, we have tested the theoretical soundness of this valuation approach on a large market dataset from the U.S.\(^1\) This is a dataset of the National Association of Insurance Commissioners, made available to the Casualty Actuarial Society for public use. It consists of over a thousand sets of 10-year run-off statistics (both payments and incurred data) for six different lines-of-business\(^1\) from 376 U.S.-based property and casualty insurance companies. This dataset can be used for back testing valuation models over a 10-year run-off period.

Figure 3 provides a comparison of the best estimate reserve method with the reserve realization after 10 years.\(^1\) It shows the nominal deviation between the actuarial estimate and the “true” value, compared to the premiums received the previous year.

---


\(^1\) The number of datasets used in our calculation for each business line is given in brackets: Motor/truck liability commercial lines (95), GTPL (160), motor liability private lines (94), product liability (33), workers compensation (69).

\(^1\) Note that after 10 years the runoff was not finalized, so the realization still bears an element of estimation.
Auditing estimates: what will the future bring?

Without this scaling, the results are far more erratic. We see here, for example, that in General third party liability (160 datasets) the median estimate is 10% higher than the scaled true value, and the 30th to the 70th percentile would deviate in a range of -11% to +56% from the scaled true value.

**Figure 3: Relative spread of reserve estimates per branch**

This boxplot provides the 10th, 30th, 50th, 70th and 90th percentile of the deviation of the branches relative to last years earned premium.
These results provide insight into the large volatility of the results of a methodology that seems at first sight to be a sound approach. We derived similar results for other structured model selection methods. This proves that the model selection has a very large impact on the results, and justifies the common practice of the use of detailed expert judgment in model selection and parameter selection.

**Table B1: Implicit considerations and residual risks when using DV01 for assessing interest rate derivatives**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Implicit</th>
<th>Example</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arithmetic correctness (only potentially found when difference &gt; &quot;X&quot; DV01)</td>
<td>Objective measurability</td>
<td>Assessing valuation of IRS by parallel modeling using own market data (result &lt; 3DV01)</td>
<td>Errors due to wrong market input are not detected</td>
</tr>
<tr>
<td></td>
<td>Objective measurability</td>
<td>Assessing valuation of IRS by using client data from systems (result &lt; 3DV01)</td>
<td>Errors due to wrong input (wrongly fed into system) are not detected; for example, errors in static data or incorrect assumptions</td>
</tr>
</tbody>
</table>
|                                  | Representativeness                     | Assessing valuation of IRS by parallel modeling using own market data (result < 3DV01) | Wrongfully concluding representativeness of valuation methodology used by clients; for example, from difference DV01 not visible that client did not use OIS*
| Objective measurability (define the difference in bps or unit measure that is acceptable) | Representativeness                      | Input data used can be reconciled with objective sources (e.g., credit spreads and static data) | Underlying market data is not representative to be used for valuation or static data not fit for purpose (wrong entries/links) |
|                                  | Arithmetical correctness               | Input data used can be reconciled with objective sources (interest curve) | Calculation errors in (enriched) data not detected                    |
| Representativeness               | Arithmetical correctness               | Use of discounted cash flows methodology based on OIS, construction of curve | Agreeing upon an interpolation method that is wrongly implemented     |
|                                  | Arithmetical correctness               | Use of certain source (DNB curve, Barclays index) is market practice     | Errors in information that is used widely and generally accepted for the purpose |
|                                  | Objective measurability                | Use of basis spread correction in the valuation (according to client's valuation manual) | Basis spread wrongly taken from data provider/wrongly implemented     |

* Overnight Indexed Swap. ** Dutch Central Bank.
8.1 Audit approach
Given the justified use of detailed expert judgment, how can we derive audit materiality ranges for claims reserves? The ranges derived from the analysis could be applied as ranges for reasonable estimates around an independent model choice of the auditing actuary. A remark here is that the auditing actuary who performs an independent recalculation will have the same deviation distribution. The materiality interval will, therefore, become very large, making this approach less useful in practice to set a tolerable range (at this stage) although it still provides a lot of insight in the key items driving the value.

In the audit practice, the soundness of model selection and of assumption and parameter setting for claims reserves is usually audited by testing the results against experience; benchmarking the method against market practice and testing the estimates for sensitivity to parameter choices. Since expert judgment needs thorough independent review, checks on the documentation and internal review assessments will help to evaluate the reasonability and justification of the model selection and parameter choices.

Having gained insight into the key sensitivities and the key parameters driving these, the auditor can use this information to assess the objective measurability of these parameters. And users of the information often want to understand these key risk drivers and their impact on the value and/or risk drivers. By using a sensitivity or DVOC approach the company does have this type of information available. It should also be clear that simply setting a tolerable range alone is not sufficient for drawing conclusions. Actuaries and auditors should look at the underlying process of why certain choices have been made. This is because some audit risk will remain when one would only look at DVOC’s and model materiality (see Box 3). To mitigate the remaining audit risk, additional control procedures need to be performed. And by looking at those elements that really do matter, by using DVOC information, the overall audit on modeled reported results can be made more efficient, effective and much more transparent. For this part, let us continue considering the audit approaches for estimates.
Auditing estimates: what will the future bring?

Box 3: Remaining audit risks when using DVOC

There will remain some audit risk when one only looks at the DVOC’s and model materiality. Models can never predict the future or describe the real world exactly, so model risk will remain. We should also recognize that audit risks remain in our suggested DVOC approach for assessment purposes. As the experience in the asset management industry (by applying DV01’s to IRS, for example) is more mature than applying DVOC’s for insurance firms, we have listed in Table B1 some implicit considerations and risks that remain for applying DV01. These could also be extended to DVOC.

Hence, in the case of DVOC, the use of wrong static data input in the model might not be detected when the difference between the assessment value and parallel model value is within the DVOC range. Reason could, for example, be incorrect use of day count convention, wrong coupon rate, wrong interest rate curve, wrong lapse rate used, wrong experience assumption factor used, wrong interpolation/extrapolation methods have been used, etc. However, the impact is limited on the assessment value. Note that ‘wrong’ here could mean ‘different from market (leading) practice or different from the own company’s policy’. Even if these errors could be detected during the qualitative assessment within the audit procedures, traditional audit reporting materiality considerations will be applied to conclude whether corrections have to be made or not, even when differences stay within the tolerable range applicable for the modelled reported figures.
9. The audit of the estimation process
Economic values and risk measures are the result of an estimation process that can be summarized as in Figure 4. In general, each element in the estimation process should be supported by sufficient governance and internal controls addressing the risk of what could go wrong in the (sub)process, substantiated by sufficient documentation supporting the design and operating effectiveness of the controls. In this regard, the following key procedures and/or controls are relevant for the auditor to consider in its audit of the estimation process.

Figure 4: The estimation process
9.1 Data and data processes
The data that are used as input for the valuation and risk models need to derived from the (routine and nonroutine) data processes, such as premiums, claims and policy master file maintenance. The insurance company should demonstrate accuracy and completeness of such data in a systematic manner. A useful data quality framework that can be used by insurance companies to demonstrate their ongoing quality of data is based on a continuous process monitoring approach, which is illustrated in Figure 5.

**Figure 5: Generic data quality framework**

<table>
<thead>
<tr>
<th>Phase 1</th>
<th>Diagnose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rapid assessment including data profiling, visualization and impact analysis in the context of targeted end-to-end process lifecycle(s)</td>
<td></td>
</tr>
<tr>
<td>Note: Diagnosis is based on assessment of foundational elements including both technical and functional data quality.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 2</th>
<th>Develop and stabilize</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean data</td>
<td></td>
</tr>
<tr>
<td>Harmonize processes</td>
<td></td>
</tr>
<tr>
<td>Establish governance</td>
<td></td>
</tr>
<tr>
<td>Enable monitoring</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Phase 3</th>
<th>Implement and sustain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional data governance framework</td>
<td></td>
</tr>
<tr>
<td>Exception notification and settlement</td>
<td></td>
</tr>
<tr>
<td>Embedded monitoring program with performance and risk indicators</td>
<td></td>
</tr>
<tr>
<td>Fact-based decisionmaking through the organization</td>
<td></td>
</tr>
</tbody>
</table>

People and culture
Process and data
Technology and enablers
In addition, where model points are extracted from databases, procedures should be in place to assess whether they are representative of the total portfolio.

9.2 Parameters
The selection of parameters by the insurance company involves procedures for calibration to market observables (e.g., benchmarking), selection of sources of data, rationale for assumption changes against experience and documentation of management best estimates for the parameters where no external data is available, and final review and approval of management.

9.3 Model
With respect to the models used in the estimation process, the following procedures need to be considered:

- **The model validation process:** in its “Proposals on Guidelines on External Audit,” EIPOA stipulates that insurance companies should ensure that the contract of engagement states that the Solvency II external audit will place reliance, where possible, on relevant internal controls and internal validation processes, including internal audit, as explicitly required by Solvency II for the purpose of the design and implementation of their audit procedures, subject to conditions set out in the international auditing standards or in nationally accepted auditing standards. This means that to be able to ensure compliance, the auditor should perform the following procedures (refer to ISA 610, “Using the work of internal auditors” and ISA 620, “Using the work of an auditor’s expert,” including the use of managements’ expert):
  
  1. Evaluate the competence, capabilities and objectivity of that expert
  2. Obtain an understanding of the work of that expert
  3. Evaluate the appropriateness of that expert’s work

- Monitoring of model risk by continuously analyzing model results against actual experience.

- **The change management procedures:** are checks on plausibility of the model results and analysis of experience against model outcome performed to provide a continuous impulse to model enhancements? Are adjustments and enhancements of the models as a result of changes in the environment or new products, distribution methods, etc. monitored?
9.4 Output: analyses and selection
The output (e.g., analysis of change, plausibility checks, analytics on results, exception reporting, etc.) should be monitored and analyzed by management. The quality and documentation of this analysis are important factors for the auditability of the information.

9.5 Monitoring experience and management action
The documentation of management’s response to the model output in relation to monitoring the progress of the strategic planning and of the risk budgets is important to ensure that the measure is used by management to manage sensitivity to the relevant risks. The rationale and signoff of any management action should be documented.

10. Analysis of change
As mentioned above, in the audit of the estimation process, an analysis of change is an important tool in getting better insight into the “why” question and also in becoming comfortable with the reliability of the actual estimates. Let us consider the following example for a specific product and align to our DVOC approach.

On 1 January 2014, the portfolio “value” included 50,000 death benefit contracts with the following characteristics:

- Maturity: 31 December 2023
- Annual premium, payable 1 January per contract: €1,250
- Acquisition costs per contract, paid 1 January 2014: €1,000
- Death benefit (on occurrence of death), payable 31 December, per contract: €100,000
- Required capital for risk margin purposes: €12,500,000 (the risk margin is calculated by determining the cost of providing an amount of eligible own funds equal to the solvency capital requirement necessary to support the insurance obligations over their lifetime).
Furthermore, assume cost of capital rate of 6% a (flat) risk-free rate (curve) of 3% and best estimate mortality rate of 1% for all future years.

At t=0, the basic own funds (BOF * €1,000) for this portfolio equals €61,219.

After one year, the BOF equals €58,498. We want to analyze the change between year t=0 and t=1.

We can undertake an “analysis of change” by using the information about the portfolio at t=0:

1. In the market consistent risk neutral world, the BOF will generate a return, in the best estimate assumption, equal to the risk-free return
2. Similarly, the BOF will increase with the release of the risk margin (+risk-free interest)

Variance in period (0,1)
3. The BOF will change due to difference between best estimate mortality and realized mortality.
4. The BOF will change due to delta in actual investment return (under- or outperformance relative to risk-free return).

Assumption changes after t=1
5. The BOF will change due to a change in mortality assumption.
6. The BOF will also change due to a change in interest rate curve.

Items 1 and 2 are fixed, and items 3 to 6 can be estimated by using DVOCs calculated on t=0 and actual parameter changes known on t=1. No additional “full balance sheet and/or P&L” calculations are needed.
10.1 Results (* €1,000)
1. BOF will make a return in the best estimate assumption equal to the risk free return; hence a euro return of €1,837
2. BOF will increase with the release of the risk margin for the period between t=0 and t=1 (+risk free interest): €773
3. DVOC death benefits: €100,00 per individual
4. DVOC investment return: €1.25 per bp increase in return (over risk free)
5. DVOC mortality assumption: –€3,747.52 per basis point increase in mortality rate
6. DVOC interest rate: –€0.60 per basis point increase in interest rate

After one year, the changes (variances and assumption changes) are also known. In the case of this example:

- 20 more individuals died in the past year (period between t=0 and t=1)
- Investment return turned out to be 3% above risk free
- Mortality assumption increased by 1bp (hence: 1.01% per year instead of 1.00%)
- Interest rate (curve) increased by 50bp (hence: 3.5% flat rate as of year 1 onwards)

Table 4 presents the analysis of change using the above information.

<table>
<thead>
<tr>
<th>t=0 calculated BOF</th>
<th>€ 61,219</th>
<th>% delta BOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 additional deaths</td>
<td>(€ 2,000)</td>
<td>72%</td>
</tr>
<tr>
<td>0.50% change in interest rate</td>
<td>€ 30</td>
<td>1%</td>
</tr>
<tr>
<td>3% additional investment return</td>
<td>€ 375</td>
<td>13%</td>
</tr>
<tr>
<td>1 bp increase in mortality rates</td>
<td>€ 3,748</td>
<td>134%</td>
</tr>
<tr>
<td>Risk-free change BOF</td>
<td>€ 1,837</td>
<td>66%</td>
</tr>
<tr>
<td>Release risk margin Y1</td>
<td>€ 773</td>
<td>28%</td>
</tr>
<tr>
<td>t=1 estimated BOF</td>
<td>€ 58,426</td>
<td>Estimated using the DVOCs</td>
</tr>
<tr>
<td>Calculated BOF</td>
<td>€ 58,498</td>
<td></td>
</tr>
<tr>
<td>Unexplained</td>
<td>€ 72</td>
<td>3%</td>
</tr>
</tbody>
</table>

At t=0, the BOF for this portfolio equals €61,219. At t=1, BOF equals €58,498

---

20 Delta BOF equals €2.721 (€61,219–€58,498).
The unexplained amount seems, in this simplified example, relatively small compared to the key risk drivers. Unexplained contains not only the estimation error for using DVOCs, it also contains a harder-to-split nonlinear impact due to changes in parameters that occur simultaneously during a year (the DVOC concepts could be further enhanced by looking at the (estimated) first partial derivative to two risk drivers). Furthermore, it contains the interest rate convexity (which could be estimated at t=0 as well by using convexity - not done here yet). Having insight in the DVOCs should also help risk management to focus on those key risk drivers that really do matter. As long as the “unexplained” is reasonably small, this approach saves a lot of calculation time.

11. Summary
Solvency II audit will certainly happen. With it, comes a number of questions, such as how does one assess “right or wrong” when reported results are based on modeled/projected future cash flows? And, what does model materiality mean in this New World? Different methods and assumptions could be equally valid, and as such there is a need to evaluate the (expert) judgment applied, and define an acceptable or tolerable range to be able to assess whether projections can be considered acceptable. In contrast, an error (when found) in the model itself, the source data, the model parameters (key risk drivers) or underlying process is a “hard” error (factual misstatement) and should be evaluated by the “traditional” materiality considerations. Using the concept of DVOC in audit processes helps making the tolerable range objective, based on those key risk drivers that really do matter.

DVOC describes the sensitivity of the (key) parameters to which an assessment value is (most) sensitive. Mathematically speaking, DVOCs are the first partial derivative of the valuation or risk function. A tolerable range as model materiality measure for audit considerations can be defined by multiplying the DVOC with a factor that is determined on the inherent variability of that specific risk driver.

We believe that the concept of using DVOC is a concrete and objective solution, in line with Solvency II standards. Technical Provision 6.3421 sets out that the assessment of an error may be carried out by expert judgment or by more sophisticated approaches, for example:

---

Sensitivity analysis in the framework of the applied model: varying the parameters and/or the data to observe the range at which a best estimate might be located.

Compare with results of other methods: applying different methods gives insight into potential model errors. These methods would not necessarily need to be more complex.

Descriptive statistics: in certain cases, the applied model allows for the derivation of descriptive statistics on the estimation error contained in the estimation (of course, this would not include the uncertainty arising from a misspecification of the model itself). Such information may help to quantitatively describe the sources of uncertainty.

Back-testing: comparing the results of the estimation against experience may help to identify systemic deviations that are due to deficiencies in modeling.

Quantitative assessment scenario as benchmark.

When assessing economic values or risk metrics, one should use various components to align with the overall model risk appetite. Assessment for a stand-alone model should be made with consideration to materiality per legal entity versus group consolidation. One should distinguish between avoidable risk (e.g., model risk stemming from not modeling known risks such as credit spread implied volatility risk) and unavoidable risk (e.g., insufficient data available in order to get a clear picture of how fat the tail of a probability distribution function is). In conclusion, tolerable range boundaries should be defined based on inherent variability, risk appetite for model risk and approval from senior management.

We see an increased need by many stakeholders to look beyond realized (ex-post) cash flows and to become comfortable with forward-looking (ex-ante) figures such as economic values and risk metrics. These figures are nothing more than an estimate within a range of possible outcomes. To assess reasonableness, auditors and actuaries need to work side by side to translate the qualitative considerations and make it quantifiable and objective as part of the overall process of accepting the figures used to manage the business.

With this article we hope to having taken a step in this direction.
The EY Global Financial Services Institute brings together world-renowned thought leaders and practitioners from top-tier academic institutions, global financial services firms, public policy organizations and regulators to develop solutions to the most pertinent issues facing the financial services industry.

The Journal of Financial Perspectives aims to become the medium of choice for senior financial services executives from banking and capital markets, wealth and asset management and insurance, as well as academics and policymakers who wish to keep abreast of the latest ideas from some of the world’s foremost thought leaders in financial services. To achieve this objective, a board comprising leading academic scholars and respected financial executives has been established to solicit articles that not only make genuine contributions to the most important topics, but are also practical in their focus. The Journal will be published three times a year.

gfsi.ey.com