The Ernst & Young 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 across banking and capital markets, asset management and insurance, as well as academics and policy-makers 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 comprised of 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 that are also practical in their focus. The Journal will be published three times a year.

gfsi.ey.com
Part 1: Banking and Capital Markets

7 Reworking the global financial architecture: is universal banking the best way forward?
Anthony Saunders, John M. Schiff Professor of Finance, Stern School of Business, New York University
Ingo Walter, Vice Dean of Faculty and Seymour Milstein Professor of Finance, Corporate Governance and Ethics, Stern School of Business, New York University

23 Regulating in the dark
Roberta Romano, Sterling Professor of Law and Director, Center for the Study of Corporate Law, Yale Law School

29 Impact of foreign banks
Stijn Claessens, Assistant Director in the Research Department and Head of Financial Studies Division, International Monetary Fund, Professor of International Finance Policy, University of Amsterdam, and Fellow, CEPR
Neeltje Van Horen, Senior Economist, Research Department, De Nederlandsche Bank (DNB), and Researcher, Economics Department, University of Amsterdam

43 Foreign-owned banks: (way) underestimated — and volatile — participants in the U.S. banking market
Daniel E. Nolle, Senior Financial Economist, Office of the Comptroller of the Currency

59 Valuing financial services firms
Aswath Damodaran, Kerschner Family Chair Professor in Finance Education, Professor of Finance, Stern School of Business, New York University

75 Getting risk governance right
Thomas F. Huertas, Partner, Ernst & Young LLP

Part 2: Insurance and Asset Management

87 Implied cost of equity capital in the U.S. insurance industry
Doron Nissim, Ernst & Young Professor of Accounting & Finance, Columbia Business School

105 The Dodd-Frank Act, Solvency II, and U.S. insurance regulation
Scott E. Harrington, Alan B. Miller Professor, Director, Wharton/Penn Risk and Insurance Program, The Wharton School, University of Pennsylvania

117 What drives the development of the insurance sector?
An empirical analysis based on a panel of developed and developing countries
Erik Feyen, Senior Financial Specialist, Financial and Private Sector Department, World Bank
Rodney Lester, Senior Consultant, World Bank
Roberto Rocha, Senior Advisor, Financial and Private Sector Department, World Bank

141 The customers have spoken, are the insurance companies listening?
A survey of the global insurance market
Paul Clark, Asia Pacific Insurance Leader, Ernst & Young LLP
Christine Delany, Director, Global Insurance Advisory, Ernst & Young LLP
Shona Burns, Senior Manager, Insurance Advisory, Ernst & Young LLP

159 Do firm characteristics influence mutual fund performance?
An empirical study for European mutual funds
Frank de Jong, Professor of Financial Markets and Risk Management, Tilburg University
Loes Wingens, Kempen Capital Management

169 Fund transfer pricing for deposits and loans, foundation and advanced
Jean Dermine, Professor of Banking and Finance, Chair Finance Area, INSEAD

179 The cultural revolution in risk management
Frank De Jonghe, Partner, Ernst & Young LLP, and University of Ghent
Mark Edeisten, Director, Ernst & Young LLP
Isabel Xavier, Senior Consultant, Ernst & Young LLP
Part 1: Banking and Capital Markets

Reworking the global financial architecture: is universal banking the best way forward?

Regulating in the dark

Impact of foreign banks

Foreign-owned banks: (way) underestimated – and volatile – participants in the U.S. banking market

Valuing financial services firms

Getting risk governance right
Reworking the global financial architecture: is universal banking the best way forward?

Anthony Saunders
John M. Schiff Professor of Finance, Stern School of Business, New York University

Ingo Walter
Vice Dean of Faculty, and Seymour Milstein Professor of Finance, Corporate Governance and Ethics, Stern School of Business, New York University

Abstract
Consolidation has been a fact of life in the wholesale financial services sector, resulting in fundamental change in the financial architecture and public exposure to systemic risk. The underlying drivers include advances in transactions and information technologies, regulatory changes, geographic shifts in growth opportunities, and the rapid evolution of client requirements, which in combination have obliged financial firms to rethink their roles as intermediaries. Moreover, financial sector reconfiguration has accelerated as a result of the global market turbulence that began in 2007, with governments either forcing or encouraging combinations of stronger and weaker financial firms in an effort to stem the crisis and improve systemic robustness.

In the process, financial firms that are “systemic” in nature and had a major role in creating the crisis have come out of it with even larger market shares and greater systemic importance. Given the episodic socialization of risk in the form of widespread use of public guarantees to firms judged too big or too interconnected to be allowed to fail, the role of systemically important financial institutions (SIFIs) is central to the financial architecture and the public interest going forward.

This survey paper considers the sources of systemic gains, losses, and risks associated with SIFIs in historical context, in the theoretical and empirical literature, and in public policy discussions – i.e., what is gained and what is lost as a result of the available policy options to deal with the dominant role of SIFIs in the financial architecture?

Introduction
A quarter century ago we formed a small group of financial economists to re-examine the relative merits of specialized versus universal banking with respect to efficiency and innovation as well as safety and soundness in financial intermediation (Walter (1985)). At the time in the United States, financial architecture had been characterized by specialized financial intermediation dating from the Glass-Steagall provisions of the Banking Act of 1933 – a structure that also characterized Japan after the American occupation following World War II and the British tradition of designated clearing banks alongside independent merchant banks, brokers, dealers, and insurers. Meanwhile, virtually all other countries followed some form of universal banking. This divergence suggested a “natural experiment” based on available evidence and stylized facts. Which system seemed to have delivered the best results based on a set of social welfare benchmarks, notably some composite of efficiency, growth, and stability?

Our conclusion at the time was that financial architecture ought to be essentially market-driven, based on a level regulatory playing field, and that market forces and the economics of industrial organization in financial intermediation should be allowed to determine optimum firm structures (Saunders and Walter (1994)). We argued that market-based drivers would harvest both scale and scope economies for the benefit of the public interest, and that market discipline and appropriate corporate governance would adequately promote safety and soundness in the financial system.

The arguments we put forward in the 1980s eventually carried the day. Gradual liberalization undertaken by the key regulatory agencies in the U.S. allowed progressively greater functional and geographic latitude for financial intermediaries, culminating in the 1999 Gramm-Leach-Bliley Act. Parallel liberalization occurred in the U.K. with the “Big Bang” in 1986 and a regulatory shift from the Bank of England to the Financial Services Authority, and in due course in Japan as well. With the disappearance of many specialized financial intermediaries into the major universal banks and financial conglomerates of the time, it seemed that the market had provided an answer – the forces of scale and scope would ultimately produce a better (more efficient and more stable) financial system, one dominated by a monoculture of national and global financial behemoths.

But would the apparently inexorable progression to exploit scale and scope necessarily be in the public interest? Since the dominant institutions were likely to be too big, too complex, and too interconnected to fail, would the evolving architecture, based on size and scope, generate negative externalities not priced in the market, so that what appeared to be a market-driven institutional result would turn out to be based on badly distorted signals? Do externalities rooted in scale and scope inevitably produce “privatization of returns and socialization of risks,” so that the institutional evolution we were observing would actually be wealth-redistributing rather than wealth-creating? Like the dog that caught the bus, what you wish for may not be exactly what you get.

Conversely, are those who today argue for breaking-up or constraining “systemic” financial conglomerates in order to constrain the system’s exposure to negative externalities, seriously underestimating the offsetting benefits of scale and scope in financial intermediation?

In this survey paper, we review the available evidence regarding financial architecture in the perspective of the systemic turbulence that surfaced in 2007 and grew to encompass virtually all major financial markets and institutions, inflicted massive financial losses on market participants, and went on to contaminate the real economy and public finances – and ultimately tested the limits of public policy. Section 2 of the paper traces the pathology of regulation, deregulation, and the evolution of the financial architecture from watershed events in the 1930s to the systemic failures of the recent past. Section 3 surveys the industrial economics of financial intermediation and explains why market forces could produce system-critical financial institutions as a dominant institutional form, identifying what would be lost and gained if they were to be broken-up. Section 4 considers the public-interest benchmarks against which institutional design in financial intermediation should be calibrated, and how scope-related regulation would be viewed in this context. Section 5 concludes.

Legacies of the past
Counterfactuals have become a favorite technique of modern historians. Suppose Napoleon had won a resounding victory at Waterloo in 1815. Suppose the British and the Germans had formed an alliance after Archduke Ferdinand’s assassination in
Sarajevo in 1914. What would the world have looked like in the ensuing years and decades? The imagination is allowed free rein – immune to definitive rejection, since the argumentation is based on some key assumptions.

Here we can develop a counterfactual on the functional separation of “banking” and “commerce” in 1933, marking the end of decades of universal banking in the U.S. and a key part of serious regulatory reform in the wake of debilitating financial and economic turbulence that took hold in 1929. What might the U.S. and global financial system have looked like in the absence of forced break-up of the legacy American universal banks? There are some available comparisons, after all, since the countries of continental Europe specifically did not follow the American approach as they struggled through the Depression, and continued with the universal banking model to the present day.

The so-called Glass-Steagall Act actually consisted of four provisions of the Banking Act of 1933, arguably the most substantial piece of U.S. financial legislation to emerge from the Great Depression – legislation that also encompassed deposit insurance and other reforms designed to restore and maintain financial stability. One of the key provisions of the Act mandated a virtually complete separation of investment banking (commerce) from deposit-taking activities (banking). It thus eliminated involvement by firms with a commercial banking charter in the securities business – specifically, underwriting and dealing in corporate debt and equity securities and municipal revenue bonds (obligations not guaranteed by the full faith and credit of municipal governments). That business, which gained dramatically in volume during the 1920s, was dominated by an amalgam of universal banks such as JP Morgan and National City Corporation and broker-dealer banks such as Goldman Sachs and Lehman Brothers. The former were listed companies engaged in a full array of commercial and investment banking activities, and the latter were private partnerships engaged mainly in securities underwriting and trading, and investing their partners’ capital.

Contemporary critics of the U.S. universal banking model feared that bank involvement in securities underwriting had directly and indirectly led them to ramp-up (warehouse) their holdings of long-term financial instruments – funded in part by customer deposits – thereby exposing themselves, their clients, and the system to potentially dangerous market, credit, and liquidity risk. When this risk materialized with a vengeance after 1929, it was thought to have contaminated the entire U.S. financial system by triggering the collapse of banks nationwide, which in turn had disastrous consequences for the real economy. About 40% of all U.S. banks failed during this period, undermining their role as financial intermediaries and cutting off the air supply to the real economy. Little wonder, in the heat of the moment, that investment banking might have been considered “too risky” for commercial banks.

The fact is that the big universal banks did increase their holdings of equities and long-term debt securities during the 1920s, probably because the demand for short-term corporate loans declined substantially. But there is little evidence that the quality of bank securities holdings was responsible for the cascading bank failures of 1930-1933. Rather, the Federal Reserve System’s unwillingness to provide liquidity to a banking system beset by depositor runs was largely responsible for the banking collapse of the time. Under the circumstances, most of the banks that failed would have collapsed even if they had held no long-term bonds at all. Evidence that commercial banks’ securities activities somehow directly caused the Great Depression has remained elusive, although the indirect causality remains arguable.

Moreover, modern studies of financial practices preceding passage of the Banking Act of 1933 suggest that the large universal banks did not in fact exploit conflicts of interest that arose from their role as commercial lenders and securities underwriters – for example using their privileged insight as lenders to prompt clients to issue securities and use the proceeds to repay commercial loans, thereby shifting losses from themselves to investors. Securities underwritten by (potentially conflicted) universal banks did not underperform those underwritten by (non-conflicted) investment banks [See, for example Gande et al. (1997), Gande et al. (1999), Puri (1994), and Puri (1996)]. Most of the questionable transactions that appeared to have occurred between banks and their securities affiliates in the 1920s would have been illegal under the financial laws and regulations that followed – the Securities act of 1933, the Securities and Exchange Act of 1934, and Investment Company Act of 1940.

The goal of financial reformers in the 1930s was to prevent a recurrence of the financial hurricane that had just passed,
Reworking the global financial architecture: is universal banking the best way forward?

had become heavily market-dominated while financial systems (liquidity requirements). By the 1980s, the U.S. financial system burdens they had to bear (notably capital requirements and market share, aided by the substantially lighter regulatory innovations were in part a product of investment banks’ Commercial paper markets, high-yield securities, asset accelerating this process. In turn, the investment banks had a great deal to do with the formation of broker-dealers, integrated investment banks and investment banking divisions of universal banks, and financial conglomerates. In 1948, Japan was forced under the U.S. occupation to adopt a version of Glass-Steagall in the form of Article 65 of the Japan Securities and Exchange Law, which contained strict separation of commercial and investment banking.

Without access to the markets for deposits and commercial loans but protected from competition by commercial banks, U.S. investment banks’ share of financial intermediation grew rapidly as financial flows progressively shifted from the balance sheets of commercial banks and other credit institutions to the financial markets. In turn, the investment banks had a great deal to do with accelerating this process.

Commercial paper markets, high-yield securities, asset securitizations, money market mutual funds, and similar innovations were in part a product of investment banks’ successful incursions into the credit institutions’ intermediation market share, aided by the substantially lighter regulatory burdens they had to bear (notably capital requirements and liquidity requirements). By the 1980s, the U.S. financial system had become heavily market-dominated while financial systems in continental Europe remained universal bank-dominated, with the major banks understandably resistant to cannibalization of profitable businesses at home via capital market instruments. Without the disruptive pressure exerted by independent broker-dealers, European equity and corporate bond markets were slow to develop, and banks took significant shareholdings and control positions in nonfinancial companies, creating close and stable client relationships but arguably eroding the role of market discipline in capital allocation and corporate governance.

Continental Europe, in contrast, engaged in no such functional separation and largely continued with the universal banking tradition, complemented by a few independent securities and advisory firms, mainly in Italy (notably Mediobanca), France (the banques d'affaires), and the Netherlands. The U.K. went its own way with a commercial banking structure centered on a short list (determined by the Bank of England) of listed clearing banks and a long tradition in the securities sector of “single capacity” jobbers (dealers), brokers, and merchant banks, which was changed in Britain’s 1986 “Big Bang” reforms to allow the formation of broker-dealers, integrated investment banks and investment banking divisions of universal banks, and financial conglomerates. Without the disruptive pressure exerted by independent broker-dealers, European equity and corporate bond markets were rapidly evolving offshore and onshore capital markets worldwide. Surviving American broker-dealers, whose competitiveness was enhanced by the disappearance of fixed brokerage commissions in the SEC’s “Mayday” financial reforms in 1974, began a sustained offensive in financial markets abroad.

Penetrating the fortresses of universal banking in one country after another, they mounted a sustained 20-year attack to wean European and later Asian corporates and governments from their reliance on intimate relationships with domestic universal

2 An exception was the 1984 failure of Continental Illinois Bank and Trust Company, which required resolution by the Federal Reserve and the Federal Deposit Insurance Corporation in a U.S.$4.5 billion rescue that wiped-out its shareholders, saw management and directors replaced, and led to eventual acquisition by Bank of America - but which successfully avoided any contamination of the banking system as a whole. Arguably the role of the Federal Reserve in the resolution of Long Term Capital Management in 1998 was another exception, with Fed-sponsored intervention and liquidation of the firm by a group of its prime-broker commercial banks intended to contain possible systemic impact on financial markets. Fed intervention came despite the existence of a commercial offer to restructure the firm in a private transaction led by Warren Buffet.
banks, offering lower funding costs and innovative financings. Meanwhile, they cultivated the global buy-side of the market—insurance companies, pension funds, and other institutional investors—with new investment alternatives and ideas for improving portfolio efficiency.

The U.S. investment banks’ global offensive was so successful that virtually all the major universal banks in Europe launched vigorous efforts to develop investment banking divisions of their own, but without their U.S. rivals’ advantages of a captive domestic market and having been battle-tested in global competition. By the early 1990s, American investment banks basically dominated their industry worldwide, with a global market share approaching 75% in terms of transactions volume. As a consequence, investment banking developed into one of the top U.S. export industries. Had universal banking remained in place in the U.S. after 1933, the lack of competitive pressure across very different strategic cohorts might well have had a very different outcome in terms of global competitiveness. So, with the passage of time, Glass-Steagall may well have had important positive unforeseen consequences.

In sum, it seems plausible that separated banking substantially delivered on its promise of efficiency, stability, and competitiveness for well over half a century, against the counterfactual that the same or even better results would have been achieved if U.S. universal banks and financial conglomerates had been allowed to persist.

Who knows? But revealed preference suggests otherwise, as the major U.S. wholesale commercial banks—notably Morgan Guaranty Trust Company, Bankers Trust, Chase Manhattan, and Citicorp—began to agitate vigorously by the early 1980s for reinstatement of universal banking powers to redress what they had come to regard as a debilitating competitive disadvantage. The grass was obviously much greener on the other side of the Glass-Steagall fence. While they could and did compete vigorously with investment banks in government bond, foreign exchange, and other traded markets as well as corporate advisories, they were hamstrung in the high-growth “Glass-Steagall illegal” sectors of the market that were critical for competing in high-growth financial intermediation channels.

Indeed, by the late 1980s commercial banks had gained the limited right to sell investment and insurance products to retail customers, as well as the right to operate separately capitalized, size-constrained wholesale securities subsidiaries (so-called Section 20 subsidiaries) under various safeguards, to prevent the commercial banking arm from contamination in the event of investment banking losses. Liberalization came in the form of administrative rulings on the part of the regulators, not legislative change. Perhaps a dozen of the major wholesale commercial banks took early advantage to build significant securities subsidiaries, especially in the bond business, to complement their powerful wholesale commercial banking and government bond activities and their emerging presence in corporate advisory work.

One important area in which the commercial banks made little headway was equities, a highly profitable growth market that was far removed from their traditional expertise in debt finance, and in which they had little sales and trading expertise and few natural relationships with companies undertaking IPOs. Moreover, lack of a market presence in equities seriously hampered their ability to build a competitive fee-based corporate finance business. This gap in their product line-up lent even more urgency to removal of the remaining Glass-Steagall restrictions though legislative action.

So, in the 1980s the commercial banks mounted a vigorous attack on Glass-Steagall, reflected in a drumbeat of political initiatives to get the rules changed. These included “high-road” arguments that the structure of financial intermediaries should be driven by competitive and strategic consideration, not anachronistic legislation. They also included “low-road” initiatives such as political lobbying and audacious tactical initiatives. Bankers Trust’s technically illegal, from a Glass-Steagall perspective, underwriting of commercial paper in 1985 led to litigation by the investment banks protecting their turf and forced the courts to decide the merits (Bankers Trust won). Citicorp’s technically illegal merger, again from a Glass-Steagall perspective, with Travelers to form Citigroup in 1998 was an audacious bet that Glass-Steagall would soon be swept into the dustbin of history by creating facts on the ground. Such bold pre-emptive strikes by commercial banks were soon validated by passage of the Gramm-Leach-Bliley Financial Services Modernization Act of 1999 (GLB).
Reworking the global financial architecture: is universal banking the best way forward?

Within two years of its enactment, every major commercial bank that took full advantage of its new freedom of access to investment banking was involved in the most serious spate of corporate financial scandals of modern times ~ including the collapse of Enron and WorldCom ~ resulting in large losses for the banks themselves and their investor clients, major fines and legal settlements, and a general erosion of confidence in financial markets. Using their enormous balance sheets, the new financial conglomerates had become fee-chasing Goliaths that were vulnerable to rogue clients playing them off against each other and against the independent investment banks.

All of the new investment banking units of bank holding companies became embroiled in major regulatory violations and exploitation of conflicts of interest, including corrupted equity research, facilitating late trading and market timing by hedge funds against the interests of ordinary shareholders of in-house mutual funds, and acting simultaneously as principal and intermediary in corporate actions. The much-criticized Sarbanes-Oxley Act followed as one effort to try to redress abuses and restore market integrity.

At the same time, the power of the newcomers in competing for investment banking market share was impressive. Less than a decade after the 1999 deregulation, only five major independent investment banks remained ~ Bear Stearns, Goldman Sachs, Lehman Brothers, Merrill Lynch, and Morgan Stanley. All the rest succumbed to the newly unchained financial conglomerates’ use of their massive balance sheets to capture investment banking business, ranging from underwriting and market-making to merger advice and proprietary trading. The dominant U.S. financial conglomerates were joined by three foreign universal banks in the global wholesale league tables ~ Credit Suisse, Deutsche Bank, and UBS AG.

Together, the top-ten global wholesale firms led over 80% of investment banking transactions by value in 2006, the year before the start of the global financial crisis. As a consequence, all ten global wholesale banks were at the epicenter of the crisis. Besides encountering “pipeline” exposure to market, credit, and liquidity risk in pursuit of what proved to be toxic flow-business, many also took on “warehouse” exposure on their massive balance sheets or in off-balance sheet conduits set up to avoid regulatory capital requirements (Acharya and Richardson (2009)).

In the event, of the independent investment banks one (Bear Stearns) failed and was absorbed by JP Morgan Chase with the aid of large-scale taxpayer assumption of risk. A second (Merrill Lynch) was absorbed by Bank of America with the encouragement of the Treasury and Federal Reserve on highly favorable terms for Merrill shareholders and employees. A third (Lehman Brothers) was allowed to fail, triggering substantial turmoil in global financial markets. The other two (Goldman Sachs and Morgan Stanley) became bank holding companies in order to obtain access to Federal Reserve refinancing. So ended the era of globally-competitive, independent U.S. investment banks.

---

3 Among the remaining constraints, the 1999 deregulation did not remove the restrictions on banks under the Bank Holding Company Act of 1956 (BHC), which prevented financial institutions from owning non-financial corporations. It conversely prohibited corporations outside of the banking sector from entering deposit taking and commercial lending. This prompted many nonfinancial corporations, such as General Electric and BMW, to set up industrial loan corporations (ILCs), mainly chartered in Utah, which enabled them to take FDIC insured deposits and make commercial loans. It also allowed broker-dealers and investment banking units of financial conglomerates to set up FDIC-insured ILCs to offer their clients in the form of brokerage sweep accounts. Also left in place was a cap of 10% on total U.S. deposits booked by any single bank holding company, although the largest financial conglomerates soon lobbied for the cap to be lifted.
Among the foreign-based wholesale universals, UBS and Royal Bank of Scotland were rescued by their respective governments. Others like Credit Suisse, Barclays, and Deutsche Bank managed to survive on their own, but some would doubtless have failed in 2008 had they not by then become “systemic” institutions, benefiting from critical central bank funding and government debt guarantees that transferred to taxpayers key risks they had assumed in executing their wholesale banking strategies.

Emerging from the global financial crisis and the ensuing public-sector debt problems in the Eurozone, the dominance of large, complex, interconnected bank-based financial conglomerates active across the spectrum of banking and “shadow” financial intermediation – renamed “systemically important financial institutions” (SIFIs) – was greater than ever before. Events demonstrated that in such a financial architecture there is no substitute for government risk absorption in extremis. They also demonstrated that efforts to monitor and constrain systemic risk centered on such firms – and to resolve them when they fail – remain unproven. Except for the U.K., proposals to reconfigure the financial architecture toward smaller, more specialized, easier to regulate and resolve, and ultimately less systemic financial institutions have been uniformly turned back.

The SIFI issue is amplified in the context of today’s global financial markets, where different institutional structures compete with one another, as well as international financial firms competing in various local markets around the world. Passage of the 2010 Dodd-Frank Act in the U.S., the Basel III capital and liquidity rules, regulatory initiatives to “ring-fence” retail banking in the U.K., and regulatory initiatives in the E.U., Switzerland, and elsewhere have moved the debate along. It invariably comes back to the linkage between size and organizational structure of financial intermediaries – either as a source of systemic risk or as a key issue in crisis-resolution after systemic damage materializes. Again, implicit in all of these efforts is the assumption that the large financial conglomerate model will dominate the financial architecture going forward. [See Acharya et al. (2009 and 2011) for a detailed discussion of the issues in retrospect and prospect].

**Value creation and value destruction in SIFIs**

If SIFIs today dominate global financial intermediation and are indeed the source of serious systemic risk, then the lack of interest in systemic risk-reduction and regulatory simplification by separating some of their key functions in the wake of the last crisis may not survive the next one. Given the severity and duration of the fallout from the 2007-09 financial turbulence, the perceived inequity of its impact, and the high-pressure lobbying by banks to undermine some of the key regulatory reforms that have been enacted, it seems unlikely that the political economy will tolerate another “made in finance” debacle anytime soon – especially if resolution of failing firms turns out to be impossible without once again calling on taxpayer support. Next time around, SIFIs could be targeted as a malignant institutional form that tends to “manufacture” systemic risk. Yet breakup of SIFIs in extremis is unlikely to be costless, imposing potential efficiency offsets to any improvement in safety and soundness that is achieved. Why? Because SIFIs are, at least in part, the product of market forces whose benefits would have to be sacrificed in any institutional restructuring that breaks them up.

Like good architecture, “form follows function” in optimum financial system design. Industrial economics suggests that the structural form of firms in particular lines of economic activity should follow the dictates of institutional comparative advantage. If there are significant economies of scale that can be exploited, it will be reflected in firm size. If there are significant economies of scope that can be exploited – either in costs or revenues – it will be reflected in the range of activities in which successful firms are engaged. If productive linkages can be built across geographies or client segments or business lines, this too will be reflected in the breadth and geographic scope of the most valuable and productive intermediaries.

In a pure market-driven context, optimum institutional structure is driven by the production functions of financial intermediaries on the one hand, and preference functions of end-users on the other. Distortions in markets for financial intermediation in the form of taxes, explicit or implicit subsidies, and regulatory constraints will alter structural optimization and create efficiency losses in the financial system. Financial institutions become larger or smaller, broader or narrower than what is optimal for the intermediation tasks to be performed. From a public interest perspective, these distortions are justifiable only if the associated social gains – including improved safety and soundness and reduced moral hazard – exceed the associated efficiency losses. This balance is not easy to achieve, yet it is at the core of the financial architecture debate.
The underlying drivers of the structure of financial intermediaries thus center on two questions. Is bigger better? Is broader better? Put simply, scale and scope. Better for whom? For shareholders? For clients? For regulators? For society? The empirical evidence so far allows no definitive conclusions on some of the key issues, with the datasets needed to find defensible evidence reflecting many moving parts and making ceteris paribus assumptions difficult.

Scale
Whether significant economies or diseconomies of scale exist in financial services has been at the heart of strategic and regulatory discussions about optimum firm size in the financial services industry for decades. Are larger firms associated with increased scale economies and hence profitability and shareholder value? Can increased average size of firms create a more efficient financial sector? Credible empirical work must isolate the impact of pure size-effects imbedded in the production function of the financial intermediary, and isolate them from other gains or losses related to firm size.

Cost estimation has consistently found that economies of scale are achieved with increases in size among small commercial banks (e.g., below U.S.$100 million in asset size). A few studies have shown that scale economies may also exist in banks falling into the U.S.$100 million to U.S.$5 billion range [Cornett et al. (2003)]. There is limited evidence so far of scale economies in the case of banks larger than U.S.$5 billion in asset footings. There is some scattered evidence of scale-related cost gains for banks up to U.S.$25 billion in asset size [Houston and Ryngaert (1994)]. But there is no credible evidence so far of significant firm-wide scale economies among very large banks. Some studies have found the relationship between size and average cost to be U-shaped. This suggests that small banks can benefit significantly from economies of scale, but that large banks seem to suffer from diseconomies of scale, resulting in higher average costs as they increase further in size. Overall, the consensus seems to be that scale economies and diseconomies generally do not result in more than about 6% difference in unit costs interquartile range among U.S. commercial banks.

A basic problem is that most of the available empirical studies focus entirely on firm-wide scale economies, even though the important scale issues are encountered at the level of individual businesses. Economies of scale appear to drive operating economies and competitive performance in areas such as global custody, transactions processing, and institutional asset management, for example. Economies of scale may be far less important in other areas such as private banking and M&A advisory services. Unfortunately, information on the business-line cost functions needed to identify and calibrate economies of scale are generally proprietary and rarely available to researchers. Nevertheless, findings of relatively small-scale economies in firm-wide cost functions, in the presence of significant observed scale economies in specific activity lines, suggest that important unobserved diseconomies of scale exist in other parts of the firm-wide cost functions.

Moreover, the presence of significant economies of scale at the business-line level does not necessarily provide a compelling reason for market dominance by large financial conglomerates, since these benefits could also be harvested by firms that specialize in scale-driven activities. Examples include financial specialists in scale-sensitive activities like custody, clearance and settlement, and transactions processing. Specialized scale-driven firms such as Bank of New York Mellon, State Street, Euroclear, and BlackRock are among the most successful in the industry.

Operating efficiency
Adjusting for scale, there is ample evidence that significant performance differences exist among financial intermediaries — as reflected for example in cost-to-income ratios among banks, insurance companies, and investment firms of comparable size [for a survey, see Walter (2006)]. Operating efficiency differentials in production functions center on efficiency and effectiveness in the use of labor and capital, sourcing and application of available technology, and acquisition of inputs and organizational design, alongside the design of compensation and incentive systems. This suggests that the way financial intermediaries are run may be more important than raw size or the selection of businesses that they pursue [Berger and Humphrey (1992)].

This is good news for smaller financial intermediaries, suggesting significant managerial degrees of freedom in the drive for efficiency — although if very large institutions are systematically better managed than smaller ones, there may indeed be a link between firm size and operating efficiency. It is also possible that
very large organizations may be more capable of the massive and “lumpy” capital outlays required to install and maintain cutting-edge information-technology and transactions-processing infrastructures. If extremely high recurring technology spend-levels result in greater operating efficiency, then large financial services firms will tend to benefit in cost-competition with smaller ones.

Scope
Adjusting for scale, the range of activities engaged in by financial intermediaries can have both cost and revenue benefits.

Cost reductions may be achieved by selling a broader rather than narrower range of products – i.e., the joint production of two or more products or services may be accomplished more cheaply than producing them separately. Cost economies of scope can be harvested through the sharing of IT platforms, branch networks and other overheads, information and monitoring costs, and the like. Like economies of scale, cost-related scope economies should be directly observable in production functions of financial services suppliers and in aggregate performance measures. Most empirical studies have failed to find significant cost-economies of scope in the banking, insurance, or securities industries [De Long (2001a)]. They do suggest that some cost-diseconomies of scope are encountered when firms in the financial services sector add new product ranges to their portfolios – as the product range widens, unit-costs seem to go up, although not dramatically so. Like scale economies and diseconomies, significant cost-scope economies or diseconomies would be reflected in valuations of financial intermediaries, and therefore affect optimum institutional structure.

Revenue economies of scope in financial intermediation arise when the all-in cost to the buyer of multiple financial services from a single supplier is less than the cost of purchasing them from separate suppliers. This includes search, monitoring, and contracting costs. Additionally, financial intermediaries that are diversified into several types of activities or several geographic areas tend to have more contact points with clients, supporting cross-selling. On the other hand, revenue-diseconomies of scope could arise from management complexities and conflicts of interest associated with greater breadth. Because of the proprietary nature of in-company data required to confirm or refute the existence of revenue economies of scale at the individual product level, few empirical studies of revenue-scope economies in financial services are available so far [De Long (2001b), Houston and Ryngaert (2001)].

In terms of revealed behavior, persistent management focus on cross-selling suggests that revenue economies of scope in financial intermediation may indeed exist at both the wholesale and retail levels. But they are likely to be very specific to the types of services cross-sold and the types of clients served [Mitchell and Onvural (1996)]. As always, the devil is in the detail – especially in the design of incentives and organizational structures to ensure that cross-selling actually occurs.

Universal banks and financial conglomerates like to argue that any imposition of size and scope constraints would inhibit their ability to provide financial intermediation to non-financial corporate and institutional investor clients that are themselves large and diverse. On the other hand, it seems likely that the broader the activity-range of financial firms in the presence of imperfect information, the greater the probability that the firm will encounter potential conflicts of interest, the higher will be the potential agency costs facing clients, and the more difficult and costly will be the internal and external safeguards necessary to prevent conflict exploitation. The competitive consequences associated with conflict exploitation may counteract the realization of revenue economies of scope among financial services firms.

Market power
As in other industries, financial intermediaries will seek to dominate markets in order to extract economic rents – sometimes referred to as economies of “size” as opposed to classic economies of “scale.” Exploitable market power allows intermediaries to extract rents from consumers or users of financial services and redistribute them to shareholders and employees, cross-subsidize other areas of activity, invest in wasteful projects, and reduce pressures for cost-containment. It may also affect safety and soundness, since excess margins allow financial intermediaries to be better capitalized and more resistant to shocks – an argument that may apply to the comparatively good performance of the big Canadian and Australian banks in the 2007-09 financial crisis, for example. In effect, end-users of financial services may pay a monopoly-surcharge, part of which takes the form of an insurance premium that supports improved firm robustness.
Still, despite very substantial consolidation in recent years in the most concentrated segment of the financial services industry, wholesale banking and capital markets, there is little evidence of market power based on standard metrics like the Herfindahl-Hirschman index. This is reflected in the meager risk-adjusted total returns to investors who own shares in the principal players in the industry and the low valuation metrics exhibited by many of these firms. Post-crisis consolidation in retail financial intermediation, on the other hand, may have reached unacceptable concentration in some national markets. E.U. competition authorities have required substantial post-consolidation divestitures, although rising banking concentration levels in the U.S. are not yet considered problematic.

Proprietary information
One argument supporting large, diverse financial intermediaries is that internal information flows are substantially better and involve lower costs than external information flows available to more narrowly-focused firms. Consequently, a firm that is present in a broad range of financial markets, functions, and geographies can find client-driven trading, financing, and investment opportunities that smaller and narrower firms cannot. Moreover, some areas of financial intermediation, like wholesale banking and asset management, have become the repository of dedicated expertise—embodied human capital in specialist businesses that are conducted by specialists to meet specialist client requirements.

It seems unclear whether size or breadth has much to do with the proprietary-information attribute of financial intermediation. Indeed, there has been plentiful anecdotal evidence of high-performance talent deserting SIFIs to join hedge funds and investment banking boutiques which depend on external information and derive competitive advantage from the ability to interpret and act on it.

Diversification and financial stability
Greater diversification of earnings attributable to multiple products, client segments, and geographies may be associated with more stable, safer, and ultimately more valuable financial institutions. The lower the correlations across the cash flows from a firm’s various financial intermediation activities, the greater the benefits of diversification. This should produce higher credit quality and higher debt ratings (lower bankruptcy risk), and therefore, lower cost of capital than faced by narrower, more focused firms. Likewise, greater earnings stability should bolster share prices.

Empirical evidence does indeed suggest relatively low correlations among cash flows from various businesses conducted by financial conglomerates, based in part on studies of earnings stability in hypothetical mergers of different types of financial intermediaries over time [see simulations in Saunders and Walter (1994)]. If true, this would be an important argument against breaking up financial conglomerates and universal banks.

Conglomerate discount
It is often argued that the shares of multi-product firms and business conglomerates tend to trade at prices lower than shares of more narrowly focused firms (all else being equal). There are two basic reasons why this “conglomerate discount” is alleged to exist.

First, it is argued that, on the whole, conglomerates tend to use capital inefficiently. This may be attributable to managerial discretion to engage in value-reducing projects, cross-subsidization of marginal or loss-making projects that drain resources from healthy businesses, misalignments in incentives between central and divisional managers, and the like. The bulk of value erosion in conglomerates is usually attributed to over-investment in marginally profitable activities and cross-subsidization [Berger and Ofek (1995)].

Second, a source of a possible conglomerate discount is that investors in shares of conglomerates find it difficult to “take a view” and add pure sectoral exposures to their asset portfolios. Investors may want to avoid such stocks in their effort to construct efficient asset-allocation profiles. This is especially true of performance-driven managers of institutional equity portfolios who are under pressure to outperform cohorts or equity indexes. Why would a fund manager want to invest in yet another (closed-end) fund in the form of a conglomerate—one that may be active in retail banking, wholesale commercial banking, middle-market lending private banking, corporate finance, trading, investment banking, asset management insurance, and perhaps other businesses as well?

Both the capital-misallocation effect and the portfolio-selection effect may weaken investor demand for shares of universal banks.

4 The HHI for the composite of wholesale banking activities has been less than 1,000 (sum of the squares of percentage market shares 0<HHI<10,000), and the industry remains subject to ruthless competition in most of the constituent businesses.
and financial conglomerates, lower their equity prices, and produce a higher cost of capital than if the conglomerate discount were absent. Recent empirical studies using large data panels have attempted to ascertain whether or not functional diversification is value-enhancing or value-destroying in the financial services sector. They conclude that there is a value-reduction effect of about 20% [Laeven and Levine (2007), Schmid and Walter (2009)].

The impact of business scope on the balance between greater stability (low earnings correlations among the businesses) and value destruction imbedded in highly diversified structures (associated with the conglomerate discount) raises some intriguing implications. Shareholders of universal banks and financial conglomerates may in effect be paying for diversification-driven systemic risk reduction through share-price erosion and relieving taxpayers of some of their exposure to systemic risk. This would be in line with an objective to pass some of the costs of potential bailouts back to firm shareholders, who might then pressure their boards to redirect strategy toward greater focus and less systemic firm characteristics, while at the same time lifting some of the conglomerate discount.

Monetary policy and surveillance
Two important issues often neglected in the universal banking debate are the effects on monetary policy and bank surveillance. Here it seems that both are somewhat open issues.

First, with respect to monetary policy, if there are just a few large banks in the financial system (such as in Canada and Australia), open-market operations can be quickly transmitted to the economy via the lending decision of just a few intermediaries. That is, the transmission time would be fairly short. In contrast, if there are large numbers of widely dispersed small and big banks in the system, the effect of central bank open-market operation might first affect large banks and then gradually feed down to smaller banks and small firms and consumers. This may be one reason why the lag between a monetary policy action and its effect on the real sector has been long and relatively unpredictable in the U.S. A second potential monetary policy advantage of a universal banking system is the possibility of the central bank or appropriate regulatory agency imposing moral suasion. It would seem far easier for regulators to apply pressure to undertake certain actions (e.g., increase corporate lending) in a system with 8 banks rather than 8,000.

With respect to surveillance, there are clear costs and benefits of a universal banking structure. On the one hand, regulators are faced with auditing a highly complex organization with systemic links both domestically and internationally. This may be especially so given likely differences in financial sophistication of a relatively underpaid bank examiner up against a relatively overpaid financial engineer. On the other hand, the cost and expense of auditing on an annual basis 8 rather than 8,000 banks may go some way to offset these costs, at least in an operational if not a systemic risk sense.

Benchmarking the role of SIFIs
The previous section of this paper suggests that the growing dominance of SIFIs in the global financial architecture reflects a complex balance of positive and negative economic drivers – scale, scope, market power, diversification, and conglomerate. The facts on the ground suggest that the positives must outweigh the negatives. Otherwise, SIFIs would play a much more limited role in financial intermediation than they actually do. It follows that efforts to break them up or to “ring-fence” certain activities, in the interest of reduced systemic risk and improved regulatory feasibility, will result in significant offsets in terms of erosion of financial system performance.

On the other hand, the fact that these institutions are systemic may itself explain their dominant role in the financial architecture. The argument is well-known. Without state assurances, uninsured depositors and other liability holders demand a risk premium. When a financial institution is not permitted to fail, the risk premium is no longer necessary. So, given the unacceptable consequences of institutional collapse, debt and equity investors in systemically sensitive financial services firms that surpass certain thresholds of size, complexity, and interconnectedness expect to be bailed-out by taxpayers. This results in materially lower cost of capital in comparison with non-systemic financial intermediaries, tilts the competitive playing field in their favor, and in turn allows them to capture progressively greater market share.5

5 This argument is similar to that associated with Fannie Mae and Freddie Mac in the U.S., two government sponsored enterprises (GSEs) that gradually encroached on private sector mortgage finance to achieve dominant positions in the market, based on an implied government debt guarantee and hence a AAA debt rating. In the event, the implied guarantee of the two toxic GSE-SIFIs became a reality during the 2007-09 financial crisis, with no structural solution yet at hand.
Moreover, financial intermediaries that expect to be covered by bailouts have an incentive to increase their risk in order to enjoy higher equity returns (moral hazard). In effect, they receive an un-priced subsidy from the taxpayer, which arguably distorts the underlying scale, scope, and other economic drivers of firm structure that drive their structure. Second-order distortions turn up in employee compensation levels far in excess of other sectors of the economy, out of proportion to value added to the real economy by the financial sector, and encouraging financial activity that is wealth-redistributing rather than wealth-creating. SIFIs’ competitive dominance, this argument suggests, is largely a product of their own systemic importance, with the taxpayer held hostage in the process.6

In the U.S., bailout policy toward SIFIs became explicit as early as 1984, when the Comptroller of the Currency testified to Congress that 11 banks were so important that they would not be permitted to fail. In other countries the same policy existed, and covered even more of the local financial systems than in the U.S. There were numerous examples of government support of banks in France, Switzerland, Norway, Sweden, Finland and Japan during the 1990s. The financial turbulence of 2007-09 further hardened the SIFI bailout presumption in virtually all of the affected countries, and, along with crisis-driven consolidation, further reinforced their competitive dominance.

The key question is how SIFIs contribute to, or detract from, the performance of the financial system, viewed from the perspective of the public interest. We suggest there are four benchmarks against which the financial system should be calibrated:

- **Static efficiency** – the metrics include the weighted mean spread between what ultimate savers (predominantly households) receive and what ultimate users of capital (households, nonfinancial businesses, and governments) have to pay. That spread is some composite of operating costs, regulatory costs, and intermediation losses. Fairness and transparency are additional aspects of static efficiency.

- **Dynamic efficiency** – here the metrics are less transparent, and include product and process innovation and technology change in financial intermediation and the role of the financial system in promoting economic growth by continually allocating and denying capital to competing uses in the production function that drives the real economy.

- **Stability** – the financial system itself should be sufficiently robust to withstand shocks that will inevitably emanate from the real sector from time to time. Furthermore, the financial sector itself should not generate shocks that spill-over into the real sector, generating serious negative externalities.

- **Competitiveness and robustness** – in a macro sense, the financial sector is an industry like any other, and generates income, employment, and international trade in services. Countries compete vigorously to maintain financial systems that add value in this regard.

These four benchmarks may well be in conflict with one another, and involve trade-offs that are often hard to identify and measure. Nevertheless, options for financial reform should be consistently calibrated against these benchmarks.

The more complex the industry, the greater the challenge to sensible regulation. This is nowhere as striking as in massive, complex, global financial services conglomerates that may be too hard to manage, too hard to oversee and govern, and almost certainly too hard to monitor and regulate. Consequently, the organizational structure of financial firms invariably enters into the regulatory debate which, in the real world, confronts heavily entrenched and politically connected players, and runs up against the personal financial interests of some of the brightest minds and biggest egos in business.7

There are two ways to address the central issue of internalizing systemic risk. One relies on market discipline and the other relies on functional separation. We believe the first is clearly preferable to the second, but depends more heavily on overcoming political economy, regulatory arbitrage, and corporate governance

---

6 Determining precisely when a financial institution becomes systemic raises some difficult issues. Citicorp was already the largest bank holding company in the U.S. before it merged with Travelers in 1998. So the bailout benefits of becoming still larger or more complex may be marginal if a firm already enjoys SIFI status.

7 Practitioners [e.g., Ackermann (2009)] argue that forcing structural change on systemic financial firms would be bad for global allocation of capital and risk, and new mechanisms would have to be found to preserve as much efficiency and innovation as possible in cross-border financial flows.
frictions that threaten its ability to produce a balance across the four benchmarks that is both durable and in the public interest.8

Under the market-discipline approach, if explicit or implicit subsidies of financial intermediaries, such as too-big-to-fail support, can be eliminated or properly priced, so that systemic externalities are successfully internalized, then the managements and boards of SIFIs will presumably rethink their strategies and their firm structures and react accordingly, probably becoming less systemic and certainly easier to monitor and regulate. Benefits (net of offsets) go to society by reducing its exposure to systemic risk. Benefits also go to shareholders by reducing or eliminating any conglomerate discount and allowing them to choose in which functions of financial intermediation to invest their capital – as opposed to holding shares in multi-function financial conglomerates.

Specifically, under Basel III, designated SIFIs will be subject to a capital surcharge for being systemic, and a further surcharge reflecting the potential systemic risk imbedded in their business models. Under this rule, Basel III capital ratios for Citigroup, HSBC, JP Morgan Chase, and Deutsche bank will be 9.5%, while BNP Paribas and Barclays will be 9%. Eight banks, including Bank of America, Goldman Sachs, Morgan Stanley, and Royal Bank of Scotland will be at 8.5%, and fourteen large banks that are more retail-oriented will be at 8%. These BIS Basel III capital levels do not preclude national regulators adding further surcharges. Switzerland announced a proposed “Swiss finish” with core capital ratios up to 16% for its two global banks. It would be hard to imagine bank managers and boards not rethinking their basic business models were such capital levels to be widely adopted. If they do not, investors may do it for them in the form of metrics that assign low values to current businesses and future prospects.

Indeed, there are grounds for optimism. For example, UBS in late 2012 announced a strategic redirection for its investment banking activities, to exit fixed income and commodities trading and focus on activities that support its valuable private banking franchise – with a loss of some 10,000 jobs. At the new capital levels, the risk-adjusted returns of these highly competitive activities make no business sense. Similar rethinking has been underway at other major wholesale banks such as Barclays, Citigroup, Royal Bank of Scotland, and others.

If the market-discipline solution prevails and systemic risk externalities are substantially internalized, then the issue of optimum organizational structure will resolve itself, and the associated incremental costs (offsets) will be disseminated among the end-users of financial intermediation as well as shareholders in an efficient general equilibrium solution. But if full application of the market-discipline solution to internalizing systemic risk is not possible – due to political blockage, regulatory arbitrage, or unresponsive corporate governance that leaves a residue of systemic risk – then resorting to mandated structural change involving activity constraints on SIFIs may be a second-best alternative to achieving better balance across the financial architecture benchmarks. Some of the separated financial firms will remain systemic, but they will be narrower and more specialized, and therefore more amenable to competent functional oversight by specialized regulatory agencies.

Two very different approaches to mopping-up residual systemic risk in the ongoing debate are the “Volcker Rule” under the Dodd-Frank Financial Reform and Consumer Protection Act of 2010 in the U.S. on the one hand, and the “ring-fence” proposal of the 2011 U.K. Independent Commission on Banking (Vickers Commission) recommendations as well as the E.U.’s 2012 Liikanen Report on the other.

Under the Volcker Rule, financial institutions with access to government lender of last resort (Federal Reserve) facilities and performing a vital public utility role in the payments system and as a transmission belt for monetary policy, should not be engaged in activities that generate systemic risk. Following vigorous debate, these prohibited activities now include proprietary trading in financial instruments and their derivatives (except U.S. government and municipal securities) and principal investing in hedge funds and private equity funds in excess of a cumulative total of 3% of core capital. Passage of the Volcker Rule quickly led to shut-downs and divestitures of proprietary trading desks, although the distinctions with respect to market-making and client-driven trading remained subject to debate. Nevertheless, the necessary risk-taking no longer allowed in bank-related financial firms equally quickly migrated to independent hedge funds and private equity firms without major disruptions in the market (Acharya et al. (2011)).

8 This includes periodic threats of geographic migration to more permissive regulatory environments that allegedly would accept a lender of last resort role.
Reworking the global financial architecture: is universal banking the best way forward?

The U.K. “ring fence” proposal basically limits taxpayer guarantees to retail banking and certain other commercial banking services from the bulk of wholesale banking conducted unguaranteed outside the “fence” but within the same institution – with separate governance structures and differential capital buffers. It remains unclear whether either of these models neutralizes whatever systemic risk remains after firms have adjusted to capital buffers and other efforts at systemic risk internalization [Vickers (2011)].

The Liikanen report took a different ring-fencing approach. It envisages a strict ring-fencing of proprietary and third-party trading activities both in terms of risk management and capital allocation, essentially through a holding company structure which would control separately capitalized subsidiaries with credible firewalls between them. The trading division would have to hold its own capital and stands or falls by its own activities and cannot contaminate the firm’s retail and commercial banking units. The objective is to protect taxpayers from future bailouts without protecting stockholders or subordinated debt-holders of the trading unit of the holding company from loss, thereby promoting market discipline. The Liikanen Report also advocates a better alignment of executive pay with long-term corporate and taxpayer interests by raising the “debt” component of bankers’ bonuses by linking them to long-term profitability.

Observers who have thrown in the towel on a pure market-discipline solution – the core supporters of Volcker, Vickers, and Liikanen – point to the continued presence of specialists in the financial system. Despite government supported and implicitly subsidized financial conglomerates, this suggests that a modern version of functional separation would not be ruinous when benchmarked against the four aforementioned criteria. It suggests that a new financial architecture would emerge, populated by relatively transparent financial firms that lend themselves to comparatively straightforward oversight by functional regulators in tandem with a systemic risk regulator empowered to deal with banks as well as non-bank financial firms. SIFIs tend to argue that separation limits synergies that are essential to their business models, while the other side argues that a host of nonbank financial firms will step up to conduct those activities that contribute static or dynamic financial efficiency gains without risking systemic consequences, as long as they are appropriately regulated along functional lines with systemic risk oversight.

Conclusions
Regulators face the daunting task of designing an “optimum” regulatory and supervisory structure that provides the desired degree of stability at lowest possible cost to efficiency, innovation, and competitiveness - and to do so in a highly politicized environment in a way that effectively aligns such policies among regulatory authorities functionally and internationally and avoids “fault lines” across regulatory regimes that can be gamed and arbitraged. There are no easy answers. There are only “better” and “worse” solutions as perceived by the public, to whom the regulators are ultimately accountable.9

On the surface, functional separation in the U.S. from 1933 to 1999 may have in fact done little harm and a lot of good. Growth was respectable, financial efficiency and innovation eventually attracted imitators around the world, American firms dominated global financial markets, and financial crises were largely avoided. But that scenario hinges on the plausibility of the counterfactual that these gains would not have been realized if pre-1933 universal banking had persisted in the U.S. The rapid emergence of financial conglomerates thereafter coincided with plenty of problems – a spate of corporate scandals in which financial conglomerates were the leading facilitators, financial innovation aimed at thwarting regulation and redistributing wealth, market bubbles in tech stocks and real estate, slower growth, a financial crisis of massive size and long-lasting effects, and unprecedented taxpayer assumption of risk and losses. Circumstantial evidence? Possibly so.

Certainly, the emergence of the financial system from the latest crisis has provided a window of opportunity to once again recalibrate the rules in the direction of the public interest. To be sure, the 2010-11 Financial Crisis Inquiry Commission (Angelides Commission) was a pale imitation of the Pecora Hearings leading up to the bold reforms of the Banking Act of 1933. The findings included a clear rejection of forced structural change in SIFIs. The U.K.’s Vickers Commission and E.U.’s Liikanen likewise shied away from full-blown

9 Perhaps the most respected and experienced observer of the US financial system, former Federal Reserve Chairman Paul A. Volcker, noted in a speech in April 2008 that “... today’s financial crisis is the culmination, as I count them, of at least five serious breakdowns of systemic significance in the past 25 years – on the average one every five years. Warning enough that something rather basic is amiss…. Simply stated, the bright new financial system - for all its talented participants, for all its rich rewards – has failed the test of the market place… [A] demonstrably fragile financial system that has produced unimaginable wealth for some, while repeatedly risking a cascading breakdown of the system as a whole, needs repair and reform.” Paul A. Volcker, remarks at a meeting of the Economic Club of New York, 28 April 2008, transcript at http://econclubny.org/files/Transcript_Volcker_April_2008.pdf.
SIFI reforms in favor of ring-fencing. Regulatory initiatives at the E.U. level uniformly accept SIFIs as a fact of life in the financial architecture. All accept the raison d’être and institutional dominance of large, complex, interconnected financial conglomerates, that Basel III and other efforts to internalize systemic risk will eventually succeed, and that the economic and social costs of forced structural change exceed the benefits. The good news is that all are converging on some form of separating the “casino” business from the core intermediation functions that form the financial air supply of the modern economy, either (a) by letting market forces push them into structurally simpler nonbank intermediaries that are either non-systemic or easier to oversee or regulate and resolve, or (b) by locating them in separately capitalized affiliates of financial holding companies, with credible firewalls capable of preventing contagion.

Arguably the most defensible approach to improving the financial architecture by addressing systemic risk, given the facts on the ground – and assuming it can be carried out in a disciplined, consistent, internationally coordinated and sustained manner with a firm eye to the public interest – will have to be heavily reliant on market forces in response to higher regulatory costs.

Forced to pay a significant price for generating systemic risk, SIFIs will have to draw their own strategic conclusions in the context of the microeconomics and industrial organization of global wholesale financial intermediation. But this assumes market discipline works effectively. Those who have become cynical about the political economy of regulation and “regulatory capture” have continued to advocate specific activity carve-outs as a second-best alternative to breaking the systemic risk grip of universal banks and financial conglomerates as a dominant form of organization in the financial architecture.

References

Aachrya, V., and M. Richardson (eds), 2009, Restoring financial stability, New York: John Wiley & Sons

Houston, J., and M. Rynkaert, 1994, “The overall gains from large bank mergers,” Journal of Banking and Finance, 18, 1155-1176
Kane, E. J., 2001, “Relevance and need for international regulatory standards,” Brookings-Wharton papers on financial services, 87-115
Walter, I., 2004, Mergers and acquisitions in banking and finance, Oxford University Press, New York
Regulating in the dark

Roberta Romano
Sterling Professor of Law and Director, Center for the Study of Corporate Law, Yale Law School

Abstract
Foundational financial legislation is typically adopted in the midst or aftermath of financial crises, when an informed understanding of the causes of the crisis is not yet available. Moreover, financial institutions operate in a dynamic environment of considerable uncertainty, such that legislation enacted even under the best of circumstances can have perverse unintended consequences, and regulatory requirements correct for an initial set of conditions can become inappropriate as economic and technological circumstances change. Furthermore, the stickiness of the status quo in the U.S. political system renders it difficult to revise legislation, even though there may be a consensus to do so.

This essay contends that the best means of responding to this dismal state of affairs is to include, as a matter of course, in crisis-driven financial legislation and its implementing regulation, two key procedural mechanisms: (1) a requirement of automatic subsequent review and reconsideration of the legislative and regulatory decisions at some future point in time; and (2) regulatory exemptive or waiver powers that encourage, where feasible, small scale experimentation, as well as flexibility in implementation. Both procedural devices will better inform and calibrate the regulatory apparatus, and could thereby mitigate, at least on the margin, the unintended errors which will invariably accompany financial legislation and rulemaking originating in a crisis.

Given the centrality of financial institutions and markets to economic growth and societal well-being, it is exceedingly important for legislators acting in a financial crisis with the best of intentions, to not make matters worse.

1 This article is an abridged version of a paper that is part of the book, Regulatory breakdown: the crisis of confidence in U.S. regulation (Cary Coglianese editor) published in 2012 by the University of Pennsylvania Press. The paper appears here with the Press’s permission. I would like to thank Alain Pietrancosta for suggesting what became the title of this paper; Cary Coglianese, Jill Fisch, Jonathan Macey, Andrew Verstein and participants at the Penn Program on Regulation conference, the Quinnipiac University School of Law Federalist Society Student Chapter Lunch Lecture, and the Yale Law School Graduate Student Seminar for helpful comments; and Aaron Ferner for suggesting the abridged form.
Introduction

How should one regulate in the midst of a financial crisis? This is a fundamental question for financial regulation, and it is not readily answerable, as the issues implicated are truly complex, if not intractable. Yet foundational financial legislation tends to be enacted in a crisis setting [Romano (2005)], and over the past decade, when confronted with this question, the U.S. Congress has answered it reflexively by enacting legislation massively increasing the scope and scale of the regulation of business firms, and especially financial institutions and instruments, in a manner seemingly oblivious to the cost and consequences of its actions. By tending to enact comprehensive financial legislation only in reaction to an immediate financial crisis, Congress acts most swiftly precisely when greater deliberateness is called for, given the paucity of information available to produce a high-quality decision.

In order to understand financial regulation undertaken in a crisis, we need to take account, as Frank Knight (1965) put it, of “human nature as we know it.” Human nature in this context is that legislators will find it impossible to not respond to a financial crisis by “doing something,” that is, by ratcheting up regulation, instead of waiting until a consensus understanding of what has occurred can be secured and a targeted solution then crafted, despite the considerable informational advantage from such an approach, which would, no doubt, improve the quality of decision-making. Compounding the problem, Congress tends not to move nimbly to rework financial legislation when it becomes widely acknowledged as flawed or seriously deficient. For instance, it took decades to repeal the Glass-Steagall Act’s separation of commercial and investment banking; eleven years to make relatively small revisions to accounting and bribery provisions of the Foreign Corrupt Practices Act; and eight years to amend the Sarbanes-Oxley Act to exempt only the smallest firms from the auditor attestation of internal controls’ effectiveness requirement, despite substantial consensus regarding the statutes’ problems.

In addition, financial firms operate in a dynamic environment in which there are many unknowns and unknowables and where state-of-the-art knowledge quickly obsolesces [Diebold et al. (2010)]. In such a context, even the most informed regulatory response – which Congress’s reaction in the recent crises was not – will be prone to error, and is likely to produce backward-looking regulation that takes aim at yesterday’s perceived problem, rather than tomorrow’s, for regulators necessarily operate under considerable uncertainty and at a lag behind private actors. Further, institutions and individuals adapt their behavior in response to regulation, and their reactions change over time, interacting with the regulatory environment in nonlinear, often unpredictable, ways, greatly complicating analysis. Accordingly, the unintended consequences of crisis-driven financial legislation are legion.

This essay contends that the best means of responding to the typical pattern of financial regulation – legislating in a crisis atmosphere under conditions of substantial uncertainty followed by status quo stickiness – is for Congress and regulators to include as a matter of course in financial legislation and regulation enacted in the midst or aftermath of a financial crisis, procedural mechanisms that require automatic subsequent review and reconsideration of those decisions, along with regulatory exemptive or waiver powers that create flexibility in implementation and encourage, where possible, small-scale, discrete experimentation to better inform and calibrate the regulatory apparatus. Such an approach, in my judgment, could mitigate, at least at the margin, errors which invariably accompany financial legislation and rulemaking originating in a crisis atmosphere. Given the fragility of financial institutions and markets, and their centrality to economic growth and societal well-being, this is an area in which it is exceedingly important for legislators acting in a crisis with the best of intentions, to not make matters worse.

Improving the quality of crisis-based financial regulation

There are two key components that should be included in financial regulation to mitigate the adverse effects of crisis-based financial legislation: (1) a sunset requirement that the legislation and implementing regulation be reviewed and reconsidered within a fixed period after enactment (e.g., five to six years) to stay on the books; and (2) a structure that is hospitable to regulatory experimentation wherever possible. By permitting legislators and regulators to incorporate new information into the decision-making process, and simultaneously increasing the likelihood that new information will be generated from the regulatory variety, generated by experimentation, the quality of decision-making has a better chance of being improved.
Sunsetting financial regulation

Sunsetting — providing that a statute expires on a specified date unless re-enacted — is a time-honored legislative tool.\(^2\) It has been used by Congress and state legislatures since the nation's founding, although its use as a lawmaking strategy has ebbed and flowed over time. For instance, in the late 1970s, sunset legislation rapidly coursed through the states, with 35 legislatures enacting sunset laws to review administrative agencies widely perceived to be ineffective and wasteful (Davis (1981); Price (1978)). At the same time, Congress considered, but did not enact, a broad sunset statute, yet it still followed the trend in sunsetting the newly created Commodity Futures Trading Commission (CFTC) in the Commodity Futures Trading Commission Act of 1974.

By 1990, enthusiasm for administrative agency sunsetting waned, given the time and cost of reviews, but over twenty states still have some form of active sunset review, and in recent years, as states' fiscal situations have deteriorated, they have once again adopted or reinvigorated the process (Kearney (1990); Price (1978)). Articles discussing the effectiveness of state sunset reviews in their heyday in the 1970s indicate that they were, on balance, successful, resulting in the termination of agencies (although no major entities were terminated) and improvements in agency operations, even in states that discontinued sunset reviews (Kearney (1990); Price (1978)).

Sunsetting is particularly well-suited for crisis-driven financial legislation. Of the rationales for adopting a sunsetting strategy, the key justification in the financial regulatory domain is that it mitigates the predicament of legislating with minimal information, and therefore running the risk of getting things seriously and, for all practical purposes, permanently wrong. Congress can, of course, in principle modify crisis-legislation that turns out to be misplaced. But the U.S. political system's organizing principles of separation of powers and checks and balances create numerous veto points throughout the legislative process (i.e., approval of both chambers, then presidential approval, or approval by a supermajority of both chambers) that make repealing a statute extremely arduous. Sunsetting loosens the institutional stickiness of the status quo, by putting a statute in play, with a need for affirmative legislative action at a specific date to remain in effect.

But more important, in the financial regulation context, sunsetting sets in motion a process by which post-enactment information can be incorporated into the regulatory regime. For instance, by the time of a statute's sunset review, several years after enactment, there should be a better understanding of the causes of the crisis that the legislation sought to address, along with knowledge of the enacted legislation's consequences — information indispensable for getting regulation right but unavailable when a crisis necessitates a response. In addition to permitting a more clear-eyed assessment, with the benefit of hindsight, of the crisis-enacted regulation, economic and technological conditions may have dramatically changed in the interim, with financial innovation occurring apace, and that information can also be taken advantage of in the legislative “second look,” for the most appropriate regulatory responses will undoubtedly have shifted as well.

To be effective, it is important that the sunsetting process be crafted in light of the states’ experiences with what works. To guide the collection and analysis of information in a sunset review, and hence the reassessment of whether legislation should be retained or revised, evaluative criteria for the sunset review, and not simply an expiration date, need to be specified in the statute responding to the crisis. Otherwise, a review will lack focus and may become a pro forma process, as legislators will often have more immediate concerns that they wish to pursue rather than undertake a serious reassessment, especially if, as is probable, constituent concerns in a crisis that motivated the statute in the first place have drifted to new matters (Davis (1981)). The evaluative criteria will, of course vary, depending on the specific legislation.

The availability of new information at the time a second vote on a statute is required for it to remain in force does not, of course, guarantee that legislators will engage in a serious reassessment, rather than a pro forma review (Breyer (1982)). To increase the likelihood that new information will be conscientiously acted upon, there is another component that should be included in a sunset provision. In addition to an expiration date and evaluative

---

\(^2\) For an overview of the use of temporary legislation, of which sunset statutes are one variety, see Gerson (2007). The U.S. income tax code is, in fact, rife with time-delimited provisions, often referred to as “extenders” (because they are typically automatically rolled over), rather than “sunsets.” For a critical appraisal of the political dynamics of tax sunsets, which, being related to evasion of restrictive budgetary rules, is orthogonal to the issues concerning the use of sunsets in this paper’s context of crisis-driven legislation, see Ayar (2005).
Regulating in the dark

criteria, establishment of a sunset review panel to perform the review, along with a timetable for action, is also necessary [Breyer (1982)]. A sunset review panel should be tasked to recommend what action — repeal, re-enactment, revision — Congress should take, and a timetable should set out the interval in which a panel recommendation would be considered by the House and Senate committees with jurisdiction over the legislation, after which the panel's recommendation would be automatically discharged as a bill for a floor vote if the committees do not themselves bring it, or an amended version, to the floor.³

The sunset review panel should consist of independent experts, who are neither government employees nor officials, and be empowered to obtain information from relevant regulatory agencies and firms to undertake its review. The advantage of independent experts is that they tend to self-identify more strongly with professional norms and are more concerned about reputational damage if peers perceive them to be doing the bidding of interest groups or party politics, than government employees who are in a hierarchical chain of command. For the review panel to be both politically accountable and independent, it should be appointed by Congress and the President, paralleling the practice used for creating blue ribbon government panels. Although Congress could establish a standing blue ribbon review panel, which would reduce the cost to future Congresses of forming a panel, reviews would be more effective if undertaken by panels created specifically for the legislation to be evaluated, as the relevant expertise is likely to vary with a statute's focus. For example, expertise in macroeconomics would be pertinent for reviewing much of the Dodd-Frank Act, but not Sarbanes-Oxley.

To ensure that the sunset process is meaningful, the authorizing legislation would need to include adequate funding for a review. Budgets of prior congressionally appointed blue ribbon investigatory panels could be used to provide guidance. Given budgetary concerns, Congress could impose a fee on the relevant sector affected by the legislation to cover a review panel’s operating cost. It could also mandate that governmental research organizations, such as the Congressional Research Service or General Accounting Office, and the relevant regulatory agencies, provide evaluations of the sunsetting regulations to the panel for use in its review. But that would probably not substantially reduce the expense of a sunset review, as the panel would likely want to seek to conduct its own evaluation de novo.

The rationale for this review mechanism, an expert panel and a timetable, is that the threat of a required floor vote on a recommendation made by outside experts would compel a higher-quality reassessment of a statute by all concerned and, in particular, by congressional committee members who know they cannot prevent a vote on a recommendation they might otherwise be able to oppose merely by inaction. It should also better incentivize review panel members, as they would know that a floor vote on their work product is assured. The use of a review panel has a further benefit, of reducing the time required by legislators and their staff to engage in a sunset review, for the panel would collect data and perform the analyses necessary for the legislature's reassessment. It would thereby mitigate a key operational problem experienced by states in their 1970s' sunset reviews and that led several states to abandon the procedure. Legislators, particularly in states where they were part-time, did not have the time or resources to engage in the demanding process of reviewing numerous state agencies [Kearney (1990)].

A variant of legislative sunset, which would further reduce demands placed on Congress of a required review, would be to impose the sunset review on agencies implementing the regulation. In this alternative, crisis-driven financial legislation would mandate agency reassessment of regulations implemented under the statute, with an automatic expiration in five years unless they are found to be cost-effective. With the technical analysis undertaken by independent experts, rather than agency staff, the potential bias that the agency might be too closely involved in the rules it administers to evaluate them objectively would be minimized [Coglianese (2011); Romano (2005)]. Further, to guard against an agency’s inherent bias in interpreting the independent experts’ analysis in support of the regulatory status quo or its agenda, a congressional vote on the agency’s determination should be required in an administrative sunset review regime.

³ I am advocating a modified version of a proposal of Justice (then professor) Breyer (1982), for review of federal regulatory programs for waste and inefficiency. Breyer rejected a sunset approach because he was concerned that a congressional minority could “destroy” an existing program by preventing a bill from coming out of a committee or by filibustering or otherwise blocking a floor vote to reapprove a majority-supported program. His proposal would, therefore, continue a program were Congress not to adopt a recommendation. Breyer’s proposed automatic discharge eliminates the issue of committee blocking, but not, of course, minority blocking on the floor. But sunset could be retained and the latter issue eliminated with a rule for sunset review analogous to the reconciliation process applicable to budget legislation, which limits debate and bypasses filibusters.
Opening financial legislation up to experimentation

An additional means of increasing the available information, and hence improving the quality of crisis-based legislative decision-making, is to incorporate experimentation into financial legislation. It is the genius of the federal organization of the U.S. government that makes it quite amenable to such an approach (Romano (1993)). Moreover, structuring financial regulation to be more hospitable to experimentation is consistent with a contemporary trend in economics to introduce experimentation into policymaking, as the gold standard for policy evaluation (Greenstone (2009)). Greenstone advocates implementing regulatory initiatives through a process that either starts with small-scale randomized experiments or permits states to implement different regulatory approaches. The expectation is that coverage would be expanded nationwide were these initial experiments successful, essentially on a cost-benefit metric. Although this approach, as he notes, is most feasible for environmental, health, labor market, and safety regulations, where discrete programs can be implemented using randomized trial experiments or “quasi” experiments, on the model of Food and Drug Administration testing requirements for new drugs, there is, I think, an analogue in the financial setting. That could be done by providing agencies with expanded exemptive and waiver powers and an accompanying directive to use the authority to permit individual institutions, or classes of institutions, to operate under different regulatory arrangements.\(^4\)

Congress has, in fact, used such an approach in crisis-driven financial legislation, but it has been limited in scope. For example, Sarbanes-Oxley’s mandate of independent audit committees (by requiring the SEC to direct stock exchanges to prohibit the listing of any firm without an independent committee), states that the SEC can establish exemptions to the statutory criteria of director independence (Sarbanes-Oxley Act § 301). Such an approach could be more broadly applied, and agencies instructed to implement rules along the lines of a small scale experiment, with incremental expansion only after a cost-benefit analysis undertaken by independent experts.

One means by which experimentation could be implemented within a waiver setting is by permitting a firm or class of firms to request a regulatory waiver, and by not leaving the matter solely up to an agency’s initiative. The standard for approval of an exemption could be an assessment of minimal adverse impact on the statutory objective (i.e., on systemic risk or financial statement fraud, objectives, respectively, of Dodd-Frank and Sarbanes-Oxley). Because an agency could be expected to be predisposed to believe that whatever regulation exists is good and hence to oppose exemptions, it could be required to accept, or at least to have to rebut in a meaningful way, an analysis of the proposed waiver provided by independent experts. Maintenance of the statutory purpose would be safeguarded by having the agency engage in ongoing monitoring and review of approved waivers, to make sure no adverse impact developed. And paralleling Greenstone’s (2009) contemplated regulatory reform process, were the waivers deemed successful, the agency would be expected to extend them to more, or all, firms or sectors. Where the proposed waiver is a private sector initiative, the firms could be required to cover the agency’s cost of evaluating and administering the experiment.

The interaction between statutory experimentation through waivers and required sunset reviews can, however, be complicated. When exempted firms are nonrandom, one cannot evaluate properly either the impact of the waiver with an eye to generalization or the efficacy of the regulation under sunset review, for the analysis would be subject to selection bias, as covered, and excluded firms would not be comparable. For instance, firms that request a waiver would most likely be those that would be most adversely affected by a rule. This difficulty could be addressed if regulatory waivers were constructed as natural experiments, in which firms receiving a waiver were selected by lot.\(^5\) But such an approach would, in my judgment, in many instances be politically infeasible and inappropriate, as it could seriously interfere with market competition, where the exempted firms’ operating cost would be less than the regulated firms’. In addition, if the exemption was for a limited time frame – for instance, until the “experiment” would be evaluated by

---

4 This essay advocates congressional directives to agencies to use existing exemptive authority in implementing emergency regulation. I consider explicit instruction necessary because history teaches us that the SEC will not voluntarily use its exemptive powers to remedy flawed rules that it has adopted in implementing emergency legislation. The well-known cognitive bias to favor the status quo (Samuelson and Zeckhauser (1988)) aids in explaining why agency exemptive power alone is not an effective means of revising flawed legislation and that sunsetting is critical.

5 The Securities and Exchange Commission (SEC) undertook a random experiment to investigate the effect of relaxing restrictions on short selling in 2004 (SEC Office of Economic Analysis (2007)). The experimental results led the agency to repeal the uptick rule restricting short sales (SEC (2007)). Despite the change in policy’s grounding in a “gold standard” natural experiment, in the wake of the financial crisis, pressed by opponents of the rule change, the SEC reinstated a limited version of the rule (SEC (2010)).
the agency for its effectiveness – then firms’ behavior may not represent how they would respond to a permanent rule, as they strategize to affect the outcome. In short, there is an inherent tension between sunset reviews and experimentation. But I do not believe that the potential conflict is sufficient to reject the proposed dual-pronged regulatory approach.6 Given the sunset review panel’s expertise, it should be well attuned to the selection issue and able to recalibrate the analysis when undertaking its regulatory evaluation in the context of experimental data.

Conclusion
Congress typically legislates on financial matters in a crisis environment, which is not conducive to high-quality decision-making. Although Congress is not about to restrain itself from acting in a crisis and increasing financial regulation, there are procedural mechanisms that can be systematically employed to mitigate the unintended consequences likely to accompany those legislative decisions.

First, sunsetting is a useful legislative tool that could mitigate legislative failure in the field of financial regulation. Second, one could also make headway in improving the quality of decision-making by incorporating experimentation and regulatory diversity into the legislative process by having Congress direct agencies to use regulatory exemptive and waiver powers to foster such objectives. In tandem with sunsetting, the greater flexibility arising from the use of such tools would facilitate timely updating of the legislative and regulatory architecture, offsetting the woefully inadequate information environment in which crisis-based legislation is enacted.

References
Coglianese, C., 2011, “Let’s review the rules,” Los Angeles Times, 29 April, A23
Weaver, S., 2011, “Sunset review in the states,” Report for the California Legislature Joint Sunset Review Committee, 8 June

Statutes cited

6 Greenstone (2009), it should be noted, recommends automatic sunsets along with experimentation in his regulatory reform agenda and does not view them to be in tension. This is most likely because he envisions experiments undertaken on a randomized, small-scale basis, which would not be likely to interfere but rather would assist in the cost-benefit evaluation of the sunset review he contemplates. In addition, he advocates automatic sunset for all regulations, many of which would not have been subjected to experimentation.
Impact of foreign banks

Stijn Claessens
Assistant Director in the Research Department, and Head of Financial Studies Division, International Monetary Fund, Professor of International Finance Policy, University of Amsterdam, and Fellow, CEPR

Neeltje Van Horen
Senior Economist, Research Department, De Nederlandsche Bank (DNB), and Researcher, Economics Department, University of Amsterdam

Abstract
This paper provides a critical assessment of the costs and benefits of foreign bank ownership. It reviews the extensive literature on the impact of foreign banks and uses a unique database on bank ownership, covering 129 countries, to (re-)examine a number of the issues discussed. It documents (changes in) foreign bank presence between 1995 and 2009, highlighting important differences across host and home countries and strong bilateral patterns. It finds that foreign banks tend to outperform domestic banks in developing countries, countries with weak institutions, and where foreign banks do not play a major role.

In addition, being from a geographically close home country increases the profitability of foreign banks. In terms of impact, it shows that foreign banks can deter domestic financial sector development in developing countries, countries with weak institutions, and where foreign banks play a minor role. Examining the impact of foreign banks on financial stability, it finds that, during the global crisis, foreign banks reduced credit more compared to domestic banks in countries where they had a small role, but not when dominant or funded locally. These findings show that, when analyzing the impact of foreign bank presence, accounting for heterogeneity, including bilateral ownership, is crucial.

JEL Classification Codes: F21, F23, G21
Keywords: foreign direct investment, international banking, distance

1 The authors would like to thank Yiyi Bai and Lindsay Mollineaux for excellent research assistance. The views expressed in this paper are those of the authors and do not necessarily represent those of the institutions with which they are or have been affiliated.
Impact of foreign banks

Introduction
The global financial crisis reinvigorated the debate on the cost and benefits of financial integration in general and foreign bank ownership in particular. While before the crisis the general consensus was that the benefits of foreign bank ownership (greatly) outweigh its costs, especially in emerging markets and developing countries, in the wake of the crisis views have become more diverse. Concerns have been raised about the impact of foreign banks on financial stability as foreign parent banks can act as international transmitters of shocks. Consequently, a critical review of our knowledge with respect to the impact of foreign bank ownership, highlighting the many dimensions through which it affects domestic financial systems, is valuable. This is what this paper intends to do.

Banks have enlarged their global footprint in the past two decades using two avenues. On the one hand, they increased the provision of cross-border credit to other banks and to corporates. On the other hand, they expanded their global reach by establishing a presence abroad through a branch or a subsidiary. Indeed, over the period 1995-2009, 560 foreign investments took place, increasing the average share (in terms of numbers) of foreign banks in local banking systems from 20 to 34 percent.\(^2\) Having established a physical presence allowed them to lend to host country firms and households using funds raised locally or from international capital markets, including through funding from the parent. While foreign bank presence increased sharply, and not many countries were left out, much variation still exists among countries. In some countries, foreign banks today comprise virtually the whole financial system, but in others they play no, or only a very marginal, role.

The presence of foreign banks, and increases therein, can be beneficial for the host market in several ways. First, foreign bank presence tends to lower the cost of financial intermediation and increases its quality. Second, it increases access to financial services for (certain types of) firms and households. Third, it enhances the financial and economic performance of borrowers. These benefits result from an increase in competition, product, technology, and know-how spillovers, and an acceleration of domestic reform. Furthermore, international banks that are diversified can more easily absorb shocks occurring in the host markets and therefore can be a more stable source of capital. The magnitude of these benefits, however, depends on the characteristics of the local market and of the foreign banks themselves: in some cases benefits are large, while in others they are only marginal (see review papers by Clarke et al. (2003), Claessens (2006), and Chopra (2007)).

At the same time, foreign bank presence can also involve costs and risks for the host country. If foreign banks tend to select only the best customers, domestic banks can be left with a worsening credit pool which can hurt their profitability and willingness to lend. In some cases, the net impact of foreign bank presence on credit provision can therefore be negative, as their credit provision is more than offset by reduced credit from domestic banks (Detragiache et al. (2008)). In addition, foreign banks can be a channel through which shocks in one country, e.g., the home market, are transmitted and affect the supply of credit in another country (Peek and Rosengren (1997, 2000)). Consequently, foreign banks can introduce financial instability.

This paper reviews the debate on the benefits and costs of foreign bank presence. It reviews what is known but also what is not yet known and provides some new insights. The rest of the paper is structured as follows. It starts with an overview of the growth in foreign bank presence over the past two decades and the factors driving the location decision of foreign banks. It highlights differences in foreign bank presence across countries and the heterogeneity among foreign banks. It then examines how foreign banks differ from domestic banks and what conditions affect foreign banks’ performance. The paper then continues with a critical assessment on how foreign banks impact financial development and financial stability. It finishes with some final remarks on policy and future research.

Trends in foreign bank presence
The bank ownership database, constructed by, and discussed in detail in, Claessens and Van Horen (2013a), provides very detailed information on the changes in foreign bank ownership that took place between 1995 and 2009.\(^3\) As a summary, Figure 1 provides

---

\(^2\) These numbers are based on the bank ownership database of Claessens and Van Horen (2013a) and therefore only reflect foreign investments where the foreign affiliate is included in Bankscope as a separate identity. As such, some investments through branches and, especially, through representative offices, are not captured.

\(^3\) The database contains ownership information of 5,324 banks over the period 1995-2009. For each year a bank is active during the sample period its ownership, domestic or foreign, is determined. When foreign-owned, the home country of the largest foreign shareholder is provided. The database includes 137 countries, thus covering high-income, emerging markets, and developing countries. For each country, the coverage is comprehensive, with banks included roughly accounting for 90 percent or more of banking system assets. The database includes 8 offshore countries, these are excluded from the analyses in this paper.
an overview of foreign bank penetration in 1995 compared to 2009. It is evident that many countries experienced a sharp increase in the share of foreign banks over the period analyzed, with relative importance, in terms of numbers, increasing from 20 percent in 1995 to 34 percent in 2009. This increase results from two counteracting trends. On the one hand, due to consolidation driven by technological changes and deregulation as well as the occurrence of financial crises, the number of domestic banks decreased. On the other hand, foreign bank presence increased through many foreign direct investments.

Figure 1 also makes clear that large differences exist across countries. Penetration has been rising especially fast over the sample period in emerging markets and developing countries. While in 1995 market shares were 18 and 24 percent in emerging markets and developing countries respectively, in 2009 they had increased to 36 and 45 percent in these groups (Figure 2). As a result, foreign banks now play important roles in financial intermediation in many of these countries, with average loan, deposit, and profit shares between 42 and 50 percent. In contrast, in OECD countries, financial intermediation remains mostly the prerequisite of domestic banks, with average foreign bank loan, deposit, and profit shares of about 20 percent.

Within the emerging market and developing country group, substantial regional differences exist though. Foreign bank ownership expanded fastest in Eastern Europe and Central Asia, with penetration in this region now second largest, at 47 percent. Growth was also strong in South Asia (increase of 120 percent), but as the base was very low, penetration in this region remains relatively limited, only 14 percent. Latin America saw very strong growth early in the period, but after 1999, in the aftermath of the Brazilian and Argentine crises, many foreign banks exited the region and new entries remained limited until investment picked up again in 2006. Still, for Latin America, the share went up considerably, from 25 to 39 percent. Foreign bank penetration in Sub Saharan Africa, already high in 1995 at 31 percent, in part due to colonial links, rose further over the period and in 2009 was over 50 percent.

What factors, besides entry restrictions, can explain why foreign bank penetration is high in some countries but low in others? The literature has identified several factors. Earlier studies found evidence that foreign bank investment tends to correlate with trade and general FDI flows, indicating that foreign banks tend to follow their customers [Grosse and Goldberg (1991), Brealey and Kaplanis (1996)]. Host country expected economic growth and local bank inefficiencies [Focarelli and Pozzolo (2000)], and low costs and efficient regulations [Buch and DeLong (2004)] have also been found to be important drivers.

Besides having access to clients with growth potential and an institutional environment where claims can be legally enforced, being able to acquire and use information efficiently is important...
as well. A bank collects and processes information to screen and monitor borrowers and projects for creditworthiness and riskiness, and, more generally, to reduce agency issues. This way a bank can offer its customers financial services on better terms than other providers may [Rajan (1992)]. To obtain this advantage, the bank needs to be able to acquire, use, and exploit information about local clients. Theory indicates that this is easier when the loan officer is close to the client and when distance between the (foreign) affiliate and headquarters is short [Aghion and Tirole (1997), Petersen and Rajan (2002), Stein (2002)]. This suggests that distance is an important factor driving location decisions. Indeed, a number of empirical studies have shown that the probability of investment is greater when countries are geographically, culturally, or institutionally close [Galindo et al. (2003), Berger et al. (2004), Focarelli and Pozzolo (2005), Claessens and Van Horen (2013b)].

Given the findings of the importance of “distance,” it is not surprising to see some strong bilateral patterns in our data. The top panel in Table 1 shows that banks from OECD countries (the biggest group of investors, owning about 70 percent of all foreign banks) tend to invest mostly in emerging markets or other OECD countries. And banks from emerging markets tend to invest in developing countries or emerging markets, while banks from developing countries tend to invest in other developing countries or emerging markets. So banks seem to seek out those host countries that are relatively similar in income levels and institutional development to their home market, presumably after taking into account the degree of competition and growth opportunities in the host market.

Furthermore, foreign bank presence tends to be regionally concentrated (bottom panel Table 1). Splitting countries in four broad geographical regions that cut across income groups (America, Asia, Europe, and Middle East and Africa), shows that the share of foreign banks coming from countries within the region is always more than 50% (the diagonal in Table 1). The highest intraregional share in 2009 is, maybe surprisingly, found for Middle East and Africa, more than 70%.

Performance of foreign banks
The previous section showed that foreign banks are important in many countries. But how does the behavior of foreign banks compare to that of domestic banks? It is obviously very difficult to capture the (different) business models of banks and how these compare between foreign and domestic banks. Table 2 nevertheless provides some insights using information from financial statements. It shows the mean values of a number of balance sheet variables and tests whether these differ significantly between foreign and domestic banks, with significance at the 10 percent level indicated by bold marked

---

**Figure 2: Share of foreign banks by income level and region, 1995 and 2009**

Source: Claessens and Van Horen (2013a)

Note: OECD includes all core OECD countries. Other high-income countries includes all countries classified as high-income by the World Bank in 2000 but not belonging to the OECD. Emerging markets includes all countries that are included in the Standard and Poor’s Emerging Market and Frontier Markets indexes and that were not high-income countries in 2000. Developing countries includes all other countries. The regions represent the regional classification as used by the World Bank.
coefficients. It turns out that for almost all indicators, foreign banks differ from domestic banks.

The table first examines the loan to asset ratio to see how foreign and domestic banks differ with respect to their asset side, i.e., their lending activities. Taking all countries together, foreign banks on average have lower loan to asset ratios than domestic banks do. This suggests that foreign banks are less involved in traditional forms of financial intermediation (i.e., lending) than domestic banks and more in other, less-traditional forms. However, in emerging markets the reverse is true. In these countries, foreign banks tend to be more active in lending (the highest for all income groups), even more so than domestic banks. A possible explanation is that in these countries foreign banks tend to be more committed, as they often entered a market through the acquisition of a (large) domestic bank with a large number of retail clients.

The table next shows the loan to deposits ratio. This ratio proxies both for the degree to which the bank is active in lending and for its dependency on wholesale funding relative to (local) deposit funding. The ratio is on average higher for domestic banks than for foreign banks, consistent with the notion that foreign banks are relatively less active in lending. This is especially so for the group of developing countries. However, in emerging markets, foreign banks tend to have higher loan to deposits ratios than domestic banks do. This suggests that in these countries foreign banks are relatively more active in lending and are also able to attract non-deposit sources of funding (including from their parent banks).

Third, we examine the liquidity position of foreign compared to domestic banks. In all countries (except for the group of other high-income countries), foreign banks have significant more liquid assets than domestic banks do. The difference is especially high in OECD countries, and less so in emerging markets. On the one hand, this suggests that foreign banks operate more conservatively compared to domestic banks, as they have greater liquidity buffers. On the other hand, since this liquidity measure also includes tradeable securities, varying from government

---

<table>
<thead>
<tr>
<th>Number and share of foreign banks from home country present in host country</th>
<th>Host income group</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OECD</td>
<td>OHI</td>
<td>EM</td>
<td>DEV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OECD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other high-income</td>
<td>10</td>
<td>0.14</td>
<td>4</td>
<td>0.06</td>
<td>38</td>
<td>0.54</td>
<td>18</td>
</tr>
<tr>
<td>Emerging markets</td>
<td>40</td>
<td>0.15</td>
<td>17</td>
<td>0.06</td>
<td>94</td>
<td>0.35</td>
<td>117</td>
</tr>
<tr>
<td>Developing countries</td>
<td>7</td>
<td>0.07</td>
<td>2</td>
<td>0.02</td>
<td>25</td>
<td>0.26</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Number and share of foreign banks from home to host income group and region (2009)
Source: Claessens and Van Horen (2013a)
Note: In the top panel of the table OECD includes all core OECD countries. Other high-income countries includes all countries classified as high-income by the World Bank in 2000 but not belonging to the OECD. Emerging markets includes all countries that are included in the Standard & Poor’s Emerging Market and Frontier Markets indexes and that were not high-income countries in 2000. Developing countries includes all other countries. In the bottom part of the table, countries are grouped in four geographical regions irrespective of the income level of the countries. “America” includes Canada, U.S. and all countries in Latin American and the Caribbean, “Asia” includes all countries in Central, East and South Asia and the Pacific countries including Japan, Australia and New Zealand. “Europe” includes all Western and Eastern European countries “MEA” includes all countries in the Middle East and North and Sub-Saharan Africa.
bonds to asset-backed securities, it probably also reflects foreign banks’ general greater activity in capital markets. Since some of these activities were important triggers for the recent financial crisis, the overall meaning of the higher liquidity ratios of foreign banks for financial stability is not so clear.

When comparing leverage, that is the ratio of equity to (unweighted) assets, foreign banks tend to be less leveraged than domestic banks, especially in OECD and developing countries, with less of a difference in other high-income countries and emerging markets. Also, in terms of capital adequacy, that is the ratio of equity to risk-weighted assets, foreign banks tend in general to have higher ratios than domestic banks, with differences across income groups similar to those for leverage. The one exception is that in emerging markets, foreign banks have similar leverage but higher capital adequacy ratios, which implies that foreign banks hold portfolios with lower risk weights. Finally, in other high-income countries and emerging markets, foreign banks tend to provision less for non-performing loans, maybe because they tend to target better quality firms. Together, this suggests that foreign banks are in general more conservative than domestic banks with respect to their asset composition and capital buffers.

This comparison shows that foreign and domestic banks differ substantially in terms of their asset mix and funding structures. How does this translate in the performance of foreign banks compared to domestic banks? Berger and Mester (1997) show that the strategic focus of a bank, and its funding and asset mix affect its performance. However, other factors likely matter as well since foreign banks can have several advantages over domestic banks. By servicing clients in more than one country, they can achieve efficiency and scale gains [see Berger (2007) for a review of the literature on economies of scale]. By being larger, for example, they may be able to afford more sophisticated models giving them superior risk management skills. Furthermore, as they can diversify risks better, they may be able to undertake investments with higher idiosyncratic risk, but also higher return. And by spreading best-practice policies and procedures over more than one country, they may achieve productivity gains. In addition, funding costs may be lower for foreign banks as they tend to have more diversified funding bases, including access to liquidity from their parent banks.

At the same time, foreign banks are likely to incur additional costs and face greater barriers in financial services provision than domestic banks. They may have less information compared to local banks on how to do business in the host country, putting them at a disadvantage, at least until they have been in the country for some time. Foreign banks might be exposed to unfair treatment by host country government and customers. And diseconomies might arise because of difficulties operating and monitoring from a distance or having to work in an institutional environment that is culturally or otherwise different. Depending on which effects dominate, which will vary by bank and host, and home characteristics, foreign banks may perform better or worse in the host country than domestic banks.

In practice indeed, studies on the performance of foreign banks find different results. Studies focusing on the U.S. find that foreign-owned banks perform significantly worse than domestic banks [see, among others, DeYoung and Nolle (1996), and

<table>
<thead>
<tr>
<th>All countries</th>
<th>Loan to assets</th>
<th>Loan to deposits</th>
<th>Liquid assets to total assets</th>
<th>Equity to assets</th>
<th>Capital ratio</th>
<th>LLR to assets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic</td>
<td>0.58</td>
<td>1.17</td>
<td>0.22</td>
<td>0.12</td>
<td>0.17</td>
<td>0.02</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.49</td>
<td>1.10</td>
<td>0.33</td>
<td>0.15</td>
<td>0.22</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Income groups</th>
<th>OECD</th>
<th>Other high-income</th>
<th>Emerging markets</th>
<th>Developing countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic</td>
<td>0.65</td>
<td>0.54</td>
<td>0.51</td>
<td>0.54</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.43</td>
<td>0.52</td>
<td>0.54</td>
<td>0.49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>All countries</th>
<th>Loan to assets</th>
<th>Loan to deposits</th>
<th>Liquid assets to total assets</th>
<th>Equity to assets</th>
<th>Capital ratio</th>
<th>LLR to assets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic</td>
<td>0.58</td>
<td>1.17</td>
<td>0.22</td>
<td>0.12</td>
<td>0.17</td>
<td>0.02</td>
</tr>
<tr>
<td>Foreign</td>
<td>0.49</td>
<td>1.10</td>
<td>0.33</td>
<td>0.15</td>
<td>0.22</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 2: Differences in balance sheet between domestic and foreign banks (2007)
Source: Claessens and Van Horen (2013a) and Bankscope
Note: OECD includes all core OECD countries. Other high-income countries includes all countries classified as high-income by the World Bank in 2000 but not belonging to the OECD. Emerging markets includes all countries that are included in the Standard & Poor's Emerging Market and Frontier Markets indexes and that were not high-income countries in 2000. Developing countries includes all other countries. All balance sheet variables are measured in 2007.
Mahajan et al (1996)). For other industrialized countries, however, studies find that foreign banks perform better [Sturm and Williams (2004)] or not different from domestic banks [Vander Vennet (1996)]. When studying foreign banks in developing countries, many studies find that foreign banks outperform domestic banks [Grigorian and Manole (2006), Berger et al. (2009)]. Others, however, find the opposite [Nikiel and Opiela (2002), Yildirim and Philippatos (2007)] or no significant difference between domestic and foreign banks [Crystal et al. (2001), Mian (2003)].

Claessens and Van Horen (2012) have tried to reconcile these different findings by studying the performance of foreign relative to domestic banks in 51 developing countries and emerging markets over the period 1999 to 2006, accounting for differences in host and home country characteristics. Their measure of bank performance is profitability (measured as profit before taxes divided by total assets, ROA), a comprehensive reduced form measure. They find that foreign banks tend to perform better when larger and having a bigger market share. Foreign banks also perform better when from a high income country, when regulations in the host country are relatively weak, and when from home countries with the same language and similar regulation as the host country.

We extend the analysis of Claessens and Van Horen (2012) here using the same empirical model but with a larger set of 74 countries, including also advanced countries.4 To avoid the impact of the recent financial crisis, performance is measured over the period 1999 to 2007. The variable of interest is a foreign ownership dummy which is one if the bank is foreign owned in a particular year [based on information in the bank ownership database of Claessens and Van Horen (2013a)]. The analysis further controls for a number of commonly used bank characteristics, including three variables that capture risk taking of the bank: ROA volatility, average growth of the bank's loans, and bank leverage (equity to asset ratio). In addition, it includes the market share of the bank, its loan to asset ratio, the ratio of deposit and short-term funding to total liabilities, a dummy capturing whether the bank exited the market within 4 years after entry, and the number of years it has been active in the country. In addition to these bank characteristics, the regression includes country-year fixed effects to control for any unobserved country characteristics, including those that varied over time. The model is estimated using OLS regressions with robust standard errors, where observations are weighted with the inverse of the number of banks in the host country, to prevent any bias due to differences in market size.5

A visual presentation of the findings is provided in Figure 3 which shows for each regression the point estimate of the foreign ownership dummy, with its 5 and 95 percent confidence intervals. The results show that for the full set of countries, foreign banks on average outperform domestic banks. Profitability of foreign banks is 0.3 percentage points higher than that of domestic banks. This is economically meaningful given that mean profitability is 1.6. However, when the sample is split into developing countries, emerging markets and high-income countries, results show striking differences. Confirming the findings of Claessens and Van Horen (2012), foreign banks are especially profitable compared to domestic banks in developing countries. This is also the case for emerging markets but the difference is much less pronounced. In contrast, foreign banks tend to underperform domestic banks in high income countries, a finding in line with the results of DeYoung and Nolle (1996), and Mahajan et al. (1996).

Next, we proceed by splitting the sample of host countries along different dimensions. A number of studies have found evidence that the behavior of foreign banks might relate to their relative importance in the host country [see, e.g., Claessens and Lee (2003), and Claessens and Van Horen (2013a)]. To examine whether this also has an impact on the profitability of foreign banks, the sample is split between those countries where foreign banks control less than half of the assets and countries in which they hold more than half. This shows clearly that foreign banks are especially profitable in countries where they do not dominate. In countries where they dominate, their performance on average does not differ from that of domestic banks.

---

4 To limit endogeneity bias, only countries that are sufficiently open for foreign entry are included (at least 3 foreign banks are active over the entire sample period) and for which there is a large enough control group of domestic banks (at least 3 domestic banks are active over the entire sample period).

5 Since, in the first years after starting up a bank or acquiring an existing bank, profitability is likely affected by start-up costs, observations for the first 2 years the (foreign or domestic) bank is active or acquired are excluded.
An additional question is whether institutional characteristics matter in how well foreign banks tend to perform. Using two World Bank Doing Business indicators, the sample is split between countries where costs of contract enforcement are high (above median costs in our sample) and where they are low (below median), and between countries where credit information is low (below median) and where it is high (above median). In countries where the cost of contract enforcement is high or the availability of credit information is low, foreign banks tend to be more profitable than domestic banks. Although it is difficult to pinpoint exactly what factors drive these results, they suggest that foreign banks tend to “cherry-pick” their customers and only provide financial services to high-margin clients in countries where institutions are weak and where they only have a relatively small share of the market. The next section looks deeper into this issue by examining the link between foreign bank presence and private credit creation.

Finally, there is the question of whether distance matters for performance. Claessens and Van Horen (2012) find that foreign banks from home countries with the same language and similar regulations as in the host country perform better. Their results suggest that being close to a host country is important for foreign banks to perform well, potentially because it eases the collection of soft information and/or transmission of this information to the headquarters of the bank (for theory and empirical evidence see, for example, Aghion and Tirole (1997), Stein (2002), Berger et al. (2005), Liberti (2005), Liberti and Mian (2009), Cerqueiro et al. (2009)). This issue is re-examined here by splitting the sample in two groups of host countries: those with distant foreign banks and those with close foreign banks, based on the average kilometer distance between the host and home country of all foreign banks active in the host country. It shows, as theory would predict, that foreign banks tend to perform better (relative to domestic banks) when headquartered in countries relatively close (average distance is below the median distance in our sample of countries) but not when they are relatively distant (average distance is above the median distance).

Foreign banks’ impact on domestic banking systems

Many studies have examined the consequences of foreign bank ownership on domestic banking systems. Before the crisis, the general consensus was that the benefits of foreign banks greatly outweigh costs in many dimensions (see review papers by Clarke et al. (2003), Claessens (2006), and Chopra (2007)). Since the global financial crisis there has been some revision of this view. This section reviews ideas put forward in the literature and provides some new evidence on the local market impact of foreign banks under two headings: access to financial services and financial stability.

Access to financial services

In general, studies (Claessens et al. (2001) and later studies, e.g., Mian (2003), and Berger et al. (2005)) have found that greater foreign bank presence coincides with lower overall costs of financial intermediation (measured by, among others, margins, spreads, and overheads). Also, evidence exists of better-quality financial intermediation with more foreign entry, e.g., lower loan-loss provisioning and better economic performance of borrowers (Martinez et al. (2004)). A number of factors are thought to be behind these effects. First, foreign bank presence can increase competition in the host country. Second, the entry of foreign
banks can lead to the introduction of new, more diverse products, greater use of up-to-date technologies, and know-how. In addition, foreign banks can pressure governments to improve regulation and supervision, increase transparency and, more generally, catalyze domestic reform [Levine (1996), Mishkin (2007)].

However, the literature also shows that these effects tend to depend on some conditions. For example, limited general development and entry barriers seem to hinder the effectiveness of foreign banks [García-Herrero and Martínez Pería (2007), Demirgüç-Kunt et al. (2004)]. Also, the relative size of foreign banks’ presence appears to matter: with more limited entry (as a share of the total host banking system) fewer spillovers arise, suggesting the existence of a threshold effect [Claessens and Lee (2003)]. Furthermore, as shown in the previous section, important interplays exist between host country characteristics, the distance of the home country to the host country, and foreign banks’ performance in the local market.

In terms of access to financial services, greater foreign bank presence seems to help, although here results depend (even) more on individual bank characteristics. Clarke et al. (2002) find that foreign bank entry improves financing conditions for enterprises of all sizes, although larger firms benefit more. Beck et al. (2004) and Berger et al. (2004) instead conclude that a larger foreign presence leads to a greater availability of credit to SMEs. Brown et al. (2011), on the other hand, find evidence of greater access to finance for more transparent firms when more foreign banks are present in a country. Giannetti and Ongena (2012) show that large and foreign firms are more likely to have a relationship with a foreign bank, while small firms tend to be served by private domestic banks. In addition, they find that an increase in foreign bank presence increases the probability that a firm gets access to bank loans, which holds for all types of firms.

The extent to which foreign banks contribute on net to access to finance and financial sector development thus remains debatable. Although some studies have looked at the relationship between private credit and foreign bank ownership, surprisingly little is known about the conditions under which foreign ownership positively relates to private credit and when negatively. Some suggest that foreign banks “cherry-pick” borrowers. If this is the case, foreign bank presence can undermine access to financial services as it worsens the remaining credit pool available for domestic banks, lowering overall financial sector development. This could be especially so in low-income countries where relationship lending is important. Indeed, the results in the previous section are indicative of some cherry-picking behavior among foreign banks in these types of countries. Furthermore, especially examining the relationship between foreign bank presence and financial sector development, Detragiache et al. (2008) show that greater presence of foreign banks in low-income countries is indeed associated with less credit being extended. However, Cull and Martínez Pería (2011) show that this relationship disappears, or even reverses once crisis-induced acquisition of (distressed) banks by foreigners is accounted for.

While providing interesting new insights, these studies only accounted for some of the heterogeneity found among foreign banks. Therefore, in a recent paper, Claessens and Van Horen (2013a) re-examine the relationship between foreign bank presence and levels of private credit to explore in more detail how country characteristics affect this relationship. Following the methodology of Detragiache et al. (2008), they use cross-country regressions over a sample of 111 countries: a sample that is much more diverse than that of Detragiache et al. (2008), since it represent all levels of development, not just low-income countries.

The results of Claessens and Van Horen (2013a) indicate that several host country characteristics, not only the general development of the country, seem to matter. They find for the whole sample a negative relationship between foreign bank presence and growth in private credit to GDP. However, when splitting the sample across income groups, they find no statistically significant relationship between private credit and foreign bank presence for emerging markets and high-income countries. For the group of developing countries, on the other hand, they find a negative relationship, in line with the findings of Detragiache et al. (2008). A one standard deviation increase in foreign presence is associated with a decline in private credit of 5 percentage points for this group of countries (compared to a mean private credit to GDP ratio of 20 percent). However, splitting the sample across other dimensions, the authors find that foreign banks’ negative impact on credit also occurs when foreign banks have a limited market share, when enforcing contracts is costly, and when credit information availability is limited. These results are very much in line with the previous section and suggest that in certain types of markets foreign banks are more likely to be niche players that “cherry-pick” their customers. This benefits their own profitability, but negatively affects credit provision to the private sector as a whole.
Financial stability
The role of foreign banks with respect to lending stability has been a topic high on the policy and research agenda since the start of the global financial crisis. As a starting point, it is important to realize that foreign banks can offer valuable diversification services and can absorb shocks occurring in the host market. Several papers have highlighted how foreign banks can enhance financial stability when crises occur in the host country. Studying crisis episodes in (mainly) emerging markets and developing countries, several studies show that, due to support of the parent banks, foreign affiliates do not need to rein in their credit supply during a financial crisis in the host country, while domestic banks do have to contract their lending [see, among others, Crystal et al. (2001), and De Haas and Van Lelyveld (2006, 2010)].

At the same time, and especially after the recent crisis, concerns have been raised that foreign banks can be a source of contagion. When faced with capital or funding shocks at home, foreign banks might withdraw from cross-border banking activities to redirect lending at home. This can translate into a reduction in capital that parents lend to their foreign affiliates, which in turn can have a negative impact on the supply of credit by these affiliates in the host market. The seminal studies of Peek and Rosengren (1997, 2000) show indeed that (funding) shocks to parent banks negatively affected local lending by their foreign affiliates. This transmission of bank distress to firm access to credit also took place when shocks occurred to the balance sheets of the parents of the foreign-owned subsidiaries.

Other recent studies suggest, however, that, also with respect to financial stability, one cannot look at foreign subsidiaries as one homogeneous group. Cull and Martínez Pería (2012) show that in eastern Europe, loan growth by foreign banks fell more than that of domestic private banks during the crisis, but that in Latin America foreign banks did not contract their loans at a faster pace. The distinction between the two continents seems to be driven by the fact that foreign banks in Latin America were mostly funded through a domestic deposit base, with most of the lending denominated in domestic currency, in part forced by regulatory requirements. This allowed them to maintain lending even when parent banks were hit by a funding shock. Furthermore, Cetorelli and Goldberg (2012) show that global banks actively manage their inter-office positions: when faced with a funding shock, they tend to reallocate capital within the holding towards “important” subsidiaries. While they do not study the lending behavior of subsidiaries, their results do suggest that some affiliates might be forced to curb lending due to a reduction in funding from the parent, whereas other affiliates do not feel this pressure or might even be in a better position compared to domestic banks to continue to extend credit.

It is, therefore, insightful to investigate under which conditions funding shocks to parent banks negatively affected local lending by subsidiaries during the global financial crisis and when not. Using the extensive database on bank ownership of Claessens and Van Horen (2013a) allows us to not only compare many domestic and foreign banks active in various regions, but also the behavior of different types of foreign banks in different types of host countries. Specifically, we examine how differences in levels of development, relative market share, distance between home and host country, and having access to local deposits affected the stability of lending of foreign banks during the global financial crisis.

Using a sample of 3,615 banks, of which 1,198 are foreign-owned, active in 117 countries over the period 2005-2009, we examine whether loan growth of foreign banks differed from that of domestic banks in 2009. The main variable of interest is the interaction between a dummy which is one if the bank is foreign-owned and a dummy which is one if the year equals 2009. To
control for different bank business models, funding structures, and strategies, the regressions also include a number of bank level variables, such as bank size, leverage, liquidity and deposit structures, whose impacts are allowed to differ in 2008 and 2009, and, in addition, bank fixed effects. Furthermore, all regressions include country-year fixed effects to control for (time-varying) differences in credit demand across countries. The model is estimated using fixed-effects OLS, clustering observations at the bank level, and excluding observations below the 1st and above the 99th percentile of loan growth, to reduce the impact of possibly influential outliers.7 The methodology is the same as used in the Working Paper version of Claessens and Van Horen (2013a), where more details and additional regressions can be found.8

A visual presentation of the findings is provided in Figure 4 which shows for each regression the point estimate of the parameter of interest and its 5 and 95 percent confidence interval. A negative value indicates that foreign banks reduce credit more compared to domestic banks in 2009. The results show for the full set of countries that foreign banks indeed reduced credit by 6 percentage points more in 2009 compared to domestic banks. This difference is large as the mean credit growth in 2009 was only 5 percent. However, when we split the sample again with respect to income levels, it is evident that only in developing countries and emerging markets does there exist a significant difference between foreign and domestic banks, but not in high-income countries.

A distinct difference is visible when the sample is split into countries where foreign banks hold less than 50 percent of the domestic assets and countries in which they hold more than 50 percent. In the first group, the loan growth of foreign banks in 2009 is 7 percentage points less than that of domestic banks. However, in countries where foreign banks dominate, they actually do not show any difference in their loan growth in 2009 compared to domestic banks. This result clearly indicates how important it is to allow for heterogeneity across foreign banks when examining their (crisis) behavior.

Next, it is tested whether the impact of foreign ownership on lending differs in countries where foreign banks are distant and in those where they are close: based on the average kilometer distance between the host and home country of all foreign banks active in the host country. Results indicate that the average distance of the foreign banks located in the host country does not have an impact on lending stability during the global financial crisis.

Finally, it is examined whether having access to local deposits is important for the stability of lending by foreign banks during a financial crisis. In principle, the relationship can go two ways. On the one hand, foreign banks that are large local deposit takers might be less affected by shocks to their parent’s balance sheets. On the other hand, as shown by Cetorelli and Goldberg (2012), parent banks faced with funding shocks might be inclined to transfer funds from those subsidiaries that are more active in local deposit taking (their core funding markets). Splitting countries into host countries where foreign banks on average have a high share of deposits and countries where they have a low share shows that, in both groups of host countries, foreign banks were reducing credit in 2009 by about 6 percentage points more than domestic banks.9 On a host country level, therefore, we find no difference with respect to having access to local deposits. However, interacting the foreign ownership dummy with the ratio of deposits to liabilities at the bank level (results shown in the Working Paper version of Claessens and Van Horen (2013a)), it is found that foreign banks that have a strong deposit base reduce credit significantly less than foreign banks that are funded less by deposits. This shows that the funding structure of foreign banks is very important with respect to the stability of credit provisioning when parent banks are hit by a shock.

Summarizing, the results show that, on average, foreign banks reduced lending more than domestic banks during the global crisis, but only in emerging markets and developing countries and in countries where they capture less than half of the banking sector and when lacking local funding. Foreign banks, therefore, seem to have contributed to financial instability but only under certain circumstances. However, other recent studies show that it is important to bring some further nuances to this

---

7 Results are very similar when adding a lagged dependent variable and estimating using OLS or difference GMM.
9 To split the sample the average deposit to liability share of all foreign banks in a particular host country in 2007 is calculated. A host country is included in the group of low-deposit countries when this average share is below the median of all host countries and is included in the group of high-deposit countries when above the median.
The main conclusion from a critical review of the literature is that when studying foreign banks’ behavior they should not be looked upon as a homogeneous group. Rather, when faced with growth opportunities or shocks, foreign banks from different home countries and with different characteristics will tend to operate and act in different ways in different host countries. To show this point, the paper first reviewed the factors mentioned in the literature that explain foreign bank entry. It then reviewed studies on the relative performance of foreign banks, highlighting substantial differences in findings. It showed that these differences in part arise because of variations in the level of development and the quality of institutions of the host country, the size and monopoly power of foreign banks active in the country, and the distance between home and host country. These factors together were shown to importantly determine the profitability of foreign banks, also relative to domestic banks.

In terms of the impact that foreign banks have on domestic financial systems, the literature also found varying results, with some arguing that foreign banks decrease domestic credit and introduce financial instability, while others find the opposite. The paper showed that many of the factors driving entry and relative performance influence the impact foreign banks have on the development of the domestic financial sector and on financial stability. For example, only in low-income countries and in countries where foreign banks play a minor role or when institutions are weak, is a larger foreign bank presence associated with lower private credit. Furthermore, while foreign banks on average reduced credit more compared to domestic banks during the recent global financial crisis, they did not so when important in the host country and when they funded themselves mainly through local deposit taking.

All this research of foreign banks has great policy relevance. For example, the finding that foreign banks that are funded by deposits tend to pose fewer risks, while banks that rely more on short-term, wholesale funding represent a greater risk, has implications for regulation and supervision. A case can be made

### Figure 4: Credit growth during the global financial crisis, foreign versus domestic banks

Note: The figure shows the point estimates and 5 and 95 percent confidence intervals of how the foreign ownership dummy interacted with a dummy which is one if the year is 2009 in a regression estimated using different country samples. All regressions include several bank level controls (see main text), bank and country-year fixed effects. The model is estimated using OLS with standard errors clustered at the bank level.
for requiring foreign-owned banks to hold certain amounts of deposits, as is already mandatory for subsidiaries operating in many Latin American countries. Another important lesson is that banks operating across national borders can pose risks to the host country. With rules and responsibilities for cross-border regulation and supervision not well defined, there is a clear danger that measures taken in the home country are in that nation’s (short-term) interest, but they might not be optimal from a supranational perspective. Consequently, common rules on interventions and resolution to prevent spillovers, including those arising from bank failures, are ideally introduced globally.

The findings also have implications for the shape of the world’s financial stability? Few of these topics have yet been studied. Financial systems located far from major home countries, but financial sector going forward, especially in light of the recent crisis with a number of emerging markets becoming more similar to high-income countries and realizing that being culturally close is increasingly important role, especially in their own region.

Results also suggest a number of areas for further research. One area is to investigate in more detail why distance matters. What makes distance such a barrier in the acquisition and transfer of information? Is it because of differences in “transportation” costs or because of differences in existing networks for transferring information? Does distance matter equally regardless of how the international bank is internally organized? Is it false that foreign banks closer to their headquarters can make loans easier and better, with lower risks and less need for loan-loss provisioning? Related, how does distance impact a foreign bank’s local operations? Do foreign banks that come from closer home countries make more loans to informationally intensive borrowers, at lower risks? And do banks from closer home countries have a greater impact on domestic financial systems, with less concern for financial stability? Few of these topics have yet been studied.

References
Chopra, A., 2007, “Opening up to foreign banks: harnessing the benefits and avoiding the pitfalls,” Mimeo, International Monetary Fund
Claessens, S., and N. van Horen, 2013b, “Location decisions of foreign banks and competitor remoteness,” Journal of Money, Credit and Banking, forthcoming
Clarke, G., R. Cull, and M. S. Martinez Peria, 2002, “Does foreign bank penetration reduce access to credit in developing countries? Evidence from asking
Impact of foreign banks


Mian, A., 2003, “Foreign, private domestic, and government banks: new evidence from emerging markets,” mimeo, University of Chicago


Foreign-owned banks: (way) underestimated — and volatile — participants in the U.S. banking market

Daniel E. Nolle
Senior Financial Economist, Office of the Comptroller of the Currency

Abstract
Banking and financial industry participants, analysts, and policy makers appreciate the fact that foreign banks play an important role in the U.S. financial system, but many lack a precise understanding of the size, composition, and impact of the foreign banking sector. In large measure, that vagueness is due to the fact that standard statistical reports on the banking industry, and analyses based on data underlying those reports, have traditionally “under-accounted” for foreign-owned banks in the U.S. The purpose of this study is to explain the nature of this statistical “blind spot,” and to suggest how it might best be addressed. The issue is far from trivial. As this study shows, it is crucial from an economic point of view to account fully for foreign-owned banks’ presence and activities in the U.S. banking market. That point is driven home by a consideration of the radical changes in foreign banks’ U.S. operations over the past decade of extreme turbulence in financial markets. In particular, the study shows that despite the well-documented, large swings in the performance and stability of U.S.-owned domestic banks over the decade surrounding the financial crisis, the volatility of foreign-owned banks in the U.S. was even more dramatic.

JEL Classification Codes: G21, G15, F3.
Keywords: Foreign banks, foreign banking, branches and agencies of foreign banks, foreign-owned subsidiary banks, internal capital markets, intra-company funding flows, net due-to balances.

1 The opinions expressed in this paper are those of the author alone, and should not be interpreted as representing those of the Office of the Comptroller of the Currency, or the United States Treasury Department.
Introduction

Financial market participants, analysts, and policy makers in the United States readily acknowledge that foreign-owned banks play an important role in the U.S. banking system, but many are somewhat unclear about the size, composition, and impact of the foreign banking sector. In large measure, that vagueness is due to the fact that standard statistical reports on the banking industry, and analyses based on data underlying those reports, have traditionally “under-counted” foreign-owned banks in the U.S. The first section of this study explains the nature of this statistical “blind spot,” and suggests how it might best be addressed. The issue is far from trivial. As this study shows, it is crucial from an economic point of view to account fully for foreign-owned banks’ presence in the U.S. banking market. That message is driven home by a consideration of the radical changes in foreign banks’ U.S. operations over the past decade of extreme turbulence in financial markets. In particular, the paper shows that despite the well-documented, large swings in the performance and stability of U.S.-owned domestic banks over the decade surrounding the financial crisis, the volatility of foreign-owned banks’ activities was even more dramatic. That greater volatility was manifested in significant ways on both sides of the balance sheet, as the sections below show.

Foreign banks in the U.S.: understanding terms, definitions, and the data

The concepts “international,” “cross-border,” and “foreign” banking are complex by nature because they involve the interactions of entities operating in two or more national financial markets and regulatory systems. Given this inherent conceptual complexity, clarity and transparency in the use of terms and data is especially important. In a recent paper, Goulding and Nolle (2012) construct a clear route through both the conceptual and statistical complexities.2 Their starting point is a discussion of the major concepts making up what has been called “the architecture of global banking.”3 At its foundation, the nature of global banking rests on two basic organizational dimensions: corporate structure and funding and liquidity strategy. Both involve choices along a spectrum, each of which is defined by its end-points. The end-points for the structure spectrum are the “international bank” model and the “multinational bank” model. An international bank operates out of its home country or from a major financial center, and conducts cross-border business—that is, it does not rely on establishing a physical presence in foreign banking markets. Alternatively, a multinational bank establishes a physical presence in foreign markets in the form of branch offices and/or subsidiary banks. The end-points for the strategy spectrum are the centralized strategy and the decentralized strategy. Under the centralized funding/liquidity strategy, a banking company “pools funds at major offices and redistributes them around the banking group.” Under a decentralized funding/liquidity strategy, a banking company “lets affiliates raise funds autonomously to finance assets in each location.”4

Most global banks do not strictly adhere to end-points of the spectrums, in two respects. First, many banks use a mixture of both the multinational and the international bank structural models, emphasizing to different degrees physical presence abroad and cross-border operations. Second, the intersection of the two dimensions introduces complexities. In particular, banking companies that follow (exclusively or predominantly) the multinational model may choose to employ either the centralized funding strategy, the decentralized strategy, or a mixture of both. Multinational banking companies favoring a centralized funding strategy do so by establishing a branch network abroad. Those following a decentralized funding strategy do so by establishing separately capitalized subsidiary banks abroad, which tend to be more adept at attracting local deposits and other funding.5

Foreign banking companies operate in the U.S. under both structural models and funding strategies. In this paper, foreign banks in the U.S. refers to operations under the multinational model, that is, foreign-owned banks with a physical presence in the U.S., whether in the form of a branch or a subsidiary bank.6

---

2 This section of the current paper draws heavily on sections I.A, I.B, and I.C in Goulding and Nolle (2012).
4 Phrases in quotes come directly from McCauley et al. (2010), p. 28.
5 As Goulding and Nolle (2012) note, foreign branches can and do use local funding to some extent, and subsidiary banks do not always, and everywhere, operate completely autonomously. Significantly, Bank for International Settlements (BIS) statistics show a roughly equal use of branches (1,764) and subsidiary banks (1,874) by banking companies [as pointed out in McCauley et al. (2010, p. 26)].
6 This definition follows the practice under U.S. banking law: specifically, Subpart B of the Federal Reserve’s Regulation K defines a “foreign bank” as “an organization that is organized under the laws of a foreign country and that engages directly in the business of banking outside the U.S.” See Goulding and Nolle (2012) for a detailed discussion of definitions under U.S. banking law.
In actuality, federal and state banking laws permit a wide variety of different legal entities that together constitute a foreign “banking presence in the U.S.” However, the vast majority of banking activity takes place in branches, agencies, and subsidiary banks. For example, as Goulding and Nolle (2012) document, branches and agencies of foreign banks (commonly considered together) accounted for two-thirds of the total $3,356 billion assets held in all foreign-owned offices in the U.S. in 2011q3, while foreign-owned subsidiary banks accounted for 31 percent — almost all of the remainder — of assets in foreign-owned banking offices in the U.S.¹⁴

In recognition of these facts, this paper includes as “foreign banks in the U.S.” only U.S. branches and agencies of foreign banks, designated as FBAs, and foreign-owned, U.S.-chartered subsidiary banks, designated as FSUBs. Branches (whether foreign-owned or U.S.-owned) are not separately capitalized, and do not report their earnings on a stand-alone basis. A key difference between foreign-owned branches and U.S.-owned domestic branches is that foreign-owned branches are, with few exceptions, prohibited from accepting retail deposits from U.S. citizens or residents.⁹ Foreign-owned agencies are very similar to branches, functionally and legally.¹⁰ FSUBs are commercial banks of which more than 25 percent is owned by a foreign banking organization; as with all other commercial banks, FSUBs are separately capitalized entities. Together, FBAs and FSUBs comprise “foreign banks in the U.S.,” designated in this paper by the acronym FBUSA.¹¹ U.S.-owned commercial banks are designated as domestic banks in the remainder of the paper.

Federal bank regulatory authorities routinely publish detailed statistical releases on U.S. banking industry structure and performance. Several of the most widely-cited releases are commonly regarded as covering the entire U.S. banking industry.¹² However, none of those statistical releases provide a comprehensive picture of the foreign bank sector in the U.S. For reasons discussed in detail in Goulding and Nolle (2012), either those releases do not include data for branches and agencies, or they do not separately account for foreign subsidiary banks, or both.

Table 1 clearly illustrates that this statistical “blind spot” is substantial. That table shows how foreign banks in the U.S. have traditionally been accounted for (or not) in widely referenced statistical releases, using as an example the FDIC’s Quarterly Banking Profile (QBP).¹³ It also shows (in the right-hand side columns entitled “Comprehensive account of foreign banks”) how big an impact occurs as a result of including FBAs and separately identifying FSUBs.¹⁴ The first data row in Table 1 (“U.S. branches and agencies of foreign banks [FBAs]”) highlights two major facts: (1) the QBP does not include FBAs in its calculations of banking industry assets or number of institutions; and (2) their inclusion in the Comprehensive data set means that an additional U.S.$2,207 billion dollars in assets in 236 FBAs is added to the overall size of the U.S. banking industry. Moving down two rows, the QBP columns show that, although FSUBs are included as part of the commercial bank sector, they are not separately identified as foreign-owned entities. As noted

---

7 The term “banking presence in the U.S.” is in quotes in order to emphasize that it has a specific meaning under federal banking regulation. See, e.g., Supervisory Letter SF 00-14 (SUP) on Interagency Program for U.S. Operations of Foreign Banks, Board of Governors of the Federal Reserve System (23 October 2000) (“... banking presence in the United States ... can take the form of branches, agencies, Edge and Agreement corporations, commercial lending companies, and subsidiary banks.”)

See the quarterly publication Structure and Share Data for the U.S. Offices of Foreign Banking Organizations, Board of Governors of the Federal Reserve System for a comprehensive list of all such foreign-owned banking offices by entity type, location in the U.S., and name and home country of their foreign “parent” company.

8 For underlying details see Nolle (2012), Table 1 and Box 2.

9 Several foreign-owned branches which had traditionally accepted retail deposits before that activity was prohibited with the enactment of the International Banking Act of 1978 were allowed to continue accepting retail deposits, and to insure those deposits under the FDIC’s deposit insurance system. As of 2011q3, the FDIC’s Quarterly Banking Profile listed nine such insured branches; note that the Federal Reserve’s Structure and Share report for that period lists ten insured branches, one of which reported U.S.$0 assets and was not included in the FDIC’s report.

10 See Nolle (2012), Box 2 for a detailed explanation.

11 Note that this paper’s designation of FBAs as “banks” is not a manifestation of “poetic license.” Rather, following Goulding and Nolle (2012), the application of the term “bank” to branches and agencies accords perfectly with the conceptual literature on the architecture of global banking, usage in federal bank regulatory language, and usage under legal language proscribing significant financial activities in which “banks” may engage.


13 It should be emphasized that the focus in this study on the FDIC’s Quarterly Banking Profile (QBP) is meant as illustrative, not comprehensive. In addition to the QBP, two other widely-cited reports or sources of aggregate data for “the” U.S. banking industry are the so-called “H.8” release (Assets and Liabilities of Commercial Banks in the United States), and the Flow of Funds Accounts of the United States, both published by the Board of Governors of the Federal Reserve System. Goulding and Nolle (2012) provide detailed descriptions of each of these two data sources and the QBP, and explain the ways in which each release “under-counts” the foreign banking sector in the U.S.

14 This dataset is the same one underlying much of the empirical analysis in Goulding and Nolle (2012). For complete underlying details see Nolle (2012), Box 4.
Foreign-owned banks: (way) underestimated – and volatile – participants in the U.S. banking market

in the corresponding right-hand side cells, in 2011q3 there were 60 FSUBs among all commercial banks, accounting for over U.S.$1,000 billion in U.S. banking system assets.

The two rows quantifying the “Total U.S. banking industry” reveal two additional facts. First, FDIC-insured savings institutions are counted in the QBP as part of the grand total of 7,436 “banks,” 456 more institutions than counted in the Comprehensive dataset (which does not include savings institutions among “banks”);15 but even so, at U.S.$15,011 billion in 2001q3, the banking industry – including FBAs – in the Comprehensive dataset is U.S.$1,204 billion larger than is commonly appreciated.

Second, using reports such as the QBP, it is simply not possible to ascertain how large the foreign bank presence is in the U.S.

The bottom section of Table 1 offers further perspective. The exclusion of FBAs means that a sector accounting for an almost 15 percent market share of the total U.S. banking industry is also excluded. Finally, the inclusion of all foreign-owned banks in the U.S., either in the form of branches and agencies or as subsidiary commercial banks, in a manner consistent with economic concepts and regulatory and legal language, means accurately accounting for a sector with more than a 20 percent share of the overall U.S. banking industry.

15 As noted at the bottom of Table 1, the Comprehensive data set uses a slightly broader definition of “commercial bank” than does the QBP.

Table 1 - “Under-counting” the presence of foreign-owned banks in U.S. [2011q3]

<table>
<thead>
<tr>
<th>U.S. banking industry sectors</th>
<th>Traditional coverage: FDIC Quarterly Banking Profile</th>
<th>Comprehensive account of foreign-owned banks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assets [Billion]</td>
<td>Number of banks</td>
</tr>
<tr>
<td>U.S. branches and agencies of foreign banks (FBAs)</td>
<td>not included1</td>
<td>not included1</td>
</tr>
<tr>
<td>Commercial banks</td>
<td>U.S.$12,559.9</td>
<td>6,352</td>
</tr>
<tr>
<td>of which in U.S.-chartered foreign-owned subsidiary banks (FSUBs)</td>
<td>not separately identified1</td>
<td>not separately identified1</td>
</tr>
<tr>
<td>Total U.S. banking industry</td>
<td>U.S.$13,807.72</td>
<td>7,436</td>
</tr>
<tr>
<td>of which in All Foreign-Owned Banks in the U.S. [FBUSA]2</td>
<td>unascertainable</td>
<td>unascertainable</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Foreign banking market share comparisons</th>
<th>Assets1 in foreign offices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreign banking groups</td>
<td>not included</td>
</tr>
<tr>
<td>FBA market share of U.S. banking system assets</td>
<td>not included</td>
</tr>
<tr>
<td>FDBA market share of U.S. banking system assets</td>
<td>unascertainable</td>
</tr>
<tr>
<td>Comprehensive account of</td>
<td></td>
</tr>
<tr>
<td>foreign-owned banks</td>
<td></td>
</tr>
<tr>
<td>Assets1 in foreign offices</td>
<td></td>
</tr>
<tr>
<td>14.7%</td>
<td></td>
</tr>
<tr>
<td>21.6%</td>
<td></td>
</tr>
</tbody>
</table>

Source: Adapted from Goulding and Nolle (2012), Table A.2.
1 - FBUSA = FBAs + FSUBs.
2 - Nine U.S. branches of foreign-owned banks which are permitted to offer FDIC-insured retail deposit accounts, but are not counted in the main industry aggregate tables in the Quarterly Banking Profile. These FBAs have approximately U.S.$30 billion in assets and U.S.$15 billion in insured domestic deposits. All of these FBAs were in existence, and taking insured deposits, prior to 20 December 1991 and hence were grandfathered in under the Foreign Bank Supervision Enhancement Act of 1991 as insured deposit-taking institutions.
3 - The Quarterly Banking Profile includes tables segregating groups of banks into “asset concentration groups;” one such group is labelled “International banks;” but this group should not be confused with FSUBs, as defined in the current paper. Specifically, international banks are “banks with assets greater than U.S.$10 billion and more than 25 percent of total assets in foreign offices.” This group, which could otherwise be characterized as “large, internationally-active banks,” includes both domestic banks (as defined in the current paper) and FSUBs meeting the minimum thresholds (but not separately identified as FSUBs). See, for example, Table V-A (p.10) in the third quarter 2011 issue of the Quarterly Banking Profile.
4 - Includes 6,352 FDIC-insured commercial banks and 1,084 FDIC-insured savings institutions.
5 - Following guidelines for “U.S. commercial banks” in the Federal Reserve Board’s Micro Statistics: Reporting Panels, Identification of Entity Types instructions, includes banks in the 50 states chartered as a commercial bank, a nondeposit trust company, or an industrial bank.
6 - FSUBs are a subset of commercial banks, as defined in the note 5 above, which are 25 percent or more foreign-owned.
Foreign banks’ volatility an even more important factor than sector size: key asset-side activities

The previous section explains how, by excluding U.S. branches and agencies of foreign banks and/or not separately identifying U.S.-chartered foreign-owned subsidiary banks, “traditional” banking statistics and analyses underestimate the size of the U.S. banking industry and obscure important dimensions of its structure. By itself, that observation argues for a careful rethink of standard operating procedures in reporting on, and analysis of, the U.S. banking industry. However, the next two sections of this paper “up the ante” by showing that in addition to its traditionally “under-counted” size, the foreign-owned banking sector in the U.S. warrants close attention because of the truly astonishing swings in foreign banks’ activities over the turbulent first decade of the 21st century. Indeed, despite the volatile behavior of domestic banks over that period, U.S. branches and agencies of foreign banks were measurably more volatile on both sides of the balance sheet.

Long-run trends in assets

Figure 1 provides a starting point for examining assets-side activities. The figure shows trends in total assets over the 2001-2011 period for FBAs, FSUBs, and domestic banks. Because FBAs and FSUBs operate at a smaller scale than domestic banks in aggregate, the figure uses two axes. That artifice makes it easy to identify the main point illustrated in Figure 1: swings and trajectories, both up and down, in total assets for FBAs were proportionally more extreme than for domestic banks. A secondary point illustrated in the figure is that FSUBs’ assets trends were similar overall to those of domestic banks.

Focusing first on domestic banks’ assets trends, the solid line shows a relatively steady pattern of assets growth over the pre-crisis period (roughly, 2002-2007). Domestic banks’ assets dropped with the full eruption of the financial crisis in 2008q3-q4, and then dropped throughout 2009 as the crisis deepened. Domestic banks’ assets growth began a modest recovery early in 2010, and by 2011q3 (the end of the data set constructed for this paper) total assets had exceeded their previous peak.

There are substantial contrasts between domestic banks’ assets growth and FBAs’ trends over almost the entire 2001-2011 period. Unlike the 2002-onward upward trajectory of domestic banks’ total assets, the dashed line in Figure 1 shows that FBAs’ assets did not begin to increase until end-2003/early-2004. Subsequently, FBAs’ assets grew at a more rapid pace than domestic banks until early 2008, at which point assets dropped off more sharply than in the case of domestic banks. Finally, the downward trend in FBAs’ assets continued well into 2010, turning back up abruptly at the end of that year and, finally, surging in 2011H1.

Looking just at long-run assets trends for FBAs as from the onset of the financial crisis can be misleading, however. That is especially true in regard to the inference one might draw from Figure 1 that the late-2010 upward surge in FBAs’ total assets signaled a recovery in credit extension by FBAs. A review of underlying data shows how off-base such an inference is. Figure 2a is a first step toward understanding that point. In the figure, FBAs’ assets are decomposed into three major activities: loans, cash and balances due, and other assets. Looking first at the far right-hand portion of the trend line for FBAs’ loans, it is apparent that the strong overall increase in total assets (shown in Figure 1) was not much aided by a very modest upturn in lending in early-2011. Rather, what stands out is the steep upward trajectory in FBAs’ cash holdings, as well as the enormous dollar value of those holdings, especially as from the beginning of 2011.
The Journal of Financial Perspectives

Foreign-owned banks: (way) underestimated — and volatile — participants in the U.S. banking market

Indeed, from end-2010 through 2011q3, FBAs’ cash balances more than doubled, from U.S.$397 billion to U.S.$803 billion. Furthermore, that surge followed a period of high volatility in FBAs’ cash positions, triggered initially by the global liquidity crunch that characterized the onset of the financial crisis.

The volatility of FBAs’ loan and cash trends is underscored by comparing them with domestic banks’ activities, shown in Figure 2b. Differences between domestic banks’ and FBAs’ trends abound. Domestic banks’ strong loan growth of 45 percent from 2004 to 2008q3, fueled in particular by home mortgage lending, pales in comparison with the 135 percent surge in FBA lending over that period (illustrated in Figure 2a).

On the downside, the 5 percent decline in domestic banks’ lending between 2008q4 and its nadir in 2010q3 was only one-fifth the 25 percent plunge in FBA lending activity over that period. Furthermore, although domestic banks’ cash and balances due spiked amid the rush for liquidity in the second half of 2008, the 86 percent increase between 2008q2 and 2009q1 was less than one-third of the 271 percent explosion in cash balances held by FBAs over that period. As Figure 2b shows, after their late-2008 increase, cash balances at domestic banks remained roughly steady over the next two years before rising 31 percent in 2011q1. Those not inconsequential changes were, however, dwarfed by the great volatility, and the enormous overall surge, in FBAs’ cash balances.

Lending
The comparison with domestic banks’ trends suggests that FBAs’ extraordinary lending and cash balances activities warrant further investigation. Data on the underlying composition of each of those activities is revealing. First, as past research has uniformly shown, C&I lending is by far the biggest component of the loan

16 See Goulding and Nolle (2012) for a detailed look at the underlying composition of the “other assets” category. That study concludes that “loans” and “cash and balances due” are by far the most significant drivers of FBAs’ assets-side activities; in light of that conclusion, the current paper focuses exclusively on those two assets components.

17 In the interest of keeping a complex story as clear as possible, the current paper does not include comparable information on underlying assets trend for FSUBs, a decision motivated by the findings in Goulding and Nolle (2012). That study explains at length how long-run assets trends for FSUBs differ from those of both FBAs and domestic banks but emphasized that, overall, FSUBs’ trends were roughly similar to those of domestic banks.
book for FBAs’ share. Over the period covered by the current study, of the roughly 70-75 percent of FBAs’ lending directed to the nonfinancial sector, the predominant share (two-thirds to three-quarters) is in the form of C&I lending. Indeed, detailed analysis by Goulding and Nolle (2012) of changes in FBAs’ loan book show that C&I lending was “the driver” of loan trends, with other components remaining proportionally stable over the crisis-through-post-crisis period (and, as that study shows, real estate loans accounted for less than 8 percent of total loans). Figure 3 focuses in on the pattern of C&I lending as the main driver of (1) the large decline in FBAs’ lending activity since the onset of the financial crisis, and (2) the exceptionally anemic recovery in FBAs’ lending. In the figure, two trends highlight the gravity of recent C&I lending patterns by FBAs. The solid line shows the pattern of C&I lending by the entire banking system, including, in particular, the precipitous drop from 2008q3 through the first half of 2010. Thereafter, C&I lending, banking industry-wide, has grown modestly but steadily. However, the bars, which show FBAs’ share of all C&I lending, reveal that FBAs’ participation in the recovery of C&I lending has been especially tepid.

Cash and balances due
Turning now to cash activities, Figure 4 decomposes cash and balances due into constituent activities reported by FBAs on their quarterly call reports. The left-most bar in the figure shows the composition of cash and balances due in the 2004q1-2007q4 pre-crisis period, when FBAs placed the majority of their cash balances with banking offices in the U.S., including other FBAs (almost 56 percent of all cash balances), as well as unaffiliated banks and other depositories (13.3 percent of all cash balances). Almost all the rest of their cash balances were placed with unaffiliated banking institutions abroad, including banks in their home country and the home country central bank (5.1 percent of their total cash balances), and unaffiliated banks and central banks in other foreign countries (almost 20 percent). Very small balances (less than 6 percent) were booked at Federal Reserve Banks.

The middle bar in Figure 4 shows the composition of FBAs’ cash positions over the 2008-2009 crisis period. On average, over that period, FBAs’ reserves at Federal Reserve Banks ballooned to more than half (54.5 percent) of their total cash balances which, as illustrated above in Figure 2a, had begun an unprecedented ascent. As this happened, FBAs reduced their reliance on other banks, especially banks abroad (top two segments of each bar): FBAs’ use of home country, and other, foreign banks dwindled from a pre-crisis 25 percent of cash and balances to less than 10 percent.

Although stability returned to the U.S. banking system as the crisis abated, FBAs did not return to, or even toward, their pre-crisis cash assets practices. As the right-hand bar in Figure 4 shows, 90 percent of their by-then record-high levels of cash

---

18 At least since Goldberg (1992), analyses of the importance of foreign banks in the U.S. banking market have focused on foreign banks’ business lending, as measured by commercial and industrial (C&I) loans and/or C&I loans plus commercial real estate loans. That perspective is based on the observation that FBAs in particular do not engage in as wide a range of asset-side activities as do commercial banks, instead focusing on business lending. In the event, foreign banks’ business lending market share has traditionally been half that of their share of total U.S. banking market assets. Indeed, several studies suggest that, from a business-strategy viewpoint, the most accurate measure of foreign banks’ importance relative to U.S. banks is one that includes both the C&I loan figures reported on the regular quarterly call report, and the lending activities (most of which targets business customers) of FBAs’ and internationally-active U.S.-chartered commercial banks’ so-called “Caribbean branches.” Using this more comprehensive measure, focused only on business lending, has shown as much as a 40 percent U.S. banking market share for foreign banks. McCauley and Seth (1992) and Nolle (1994) were the first to emphasize and empirically investigate this issue.

19 See Miu et al. (2012) for a perspective on this focusing on funding pressures faced by banks headquartered in Eurozone countries.
Foreign-owned banks: (way) underestimated — and volatile — participants in the U.S. banking market

The majority of the remaining 10 percent of cash balances were booked at unaffiliated banking offices in the U.S. (mostly with unaffiliated FBAs), while a small proportion (2.7 percent) were placed abroad with unaffiliated banks and central banks. Notably, almost nothing was placed with home country banks or the home central bank (0.25 percent).

Together, the overall surge in FBAs’ cash assets, and the radical changes in the underlying composition of those assets had major impacts on the balance sheet of the Federal Reserve System. Figure 5 illustrates this fact. The farthest left-hand set of bars in the figure uses banks’ average (and very stable) behavior over the 2001q2-2008q2 period as a reference point for gauging the extraordinary changes in FBAs’ cash activities as of the full onset of the financial crisis in 2008q3. Beginning in 2008q3, two sets of trends are plotted in Figure 5. The bold, dotted line in the figure plots the share of total Federal Reserve System liabilities accounted for by reserves at all banks and other depository institutions. Pre-crisis, reserves accounted for 5 percent or less of Federal Reserve liabilities. Figure 5 shows the abrupt change in that situation in the fall of 2008, when reserves surged to more than 50 percent of total Federal Reserve System liabilities. As is well-documented, this increase in reserves is a direct consequence of the massive liquidity provision measures engineered by the Federal Reserve (Fleming (2012), Cecchetti (2009), Willardson and Pederson (2010), and Fleming and Klagge (2011)).

The bars in Figure 5 show the percentage of all reserves at Federal Reserve Banks accounted for separately by domestic banks, FBAs, and FSUBs. Of particular note are the large declines in the share of total reserves at Federal Reserve Banks accounted for by domestic banks (textured bars), and the correspondingly large increases in FBAs’ share (darkest bars). It is now well-known that foreign-owned banks participated heavily in the Federal Reserve’s liquidity provision programs;21 perhaps less well-known is the magnitude of the build-up of foreign-owned banks’ reserves. In particular, as the dark FBA bars show, by 2011q3,

---

20 Note that in 2011q3, the last quarter covered in our dataset, FBAs’ reserves at Federal Reserve Banks accounted for 95 percent of their total cash and balances due.

21 See, for example, Goldberg and Skeie (2011). See also Fleming (2012), who documents the fact that “the 10 most frequent borrowers from the TAF (Term Auction Facility) [were] New York branches of foreign banks” (p. 18, and Table 2 in that paper).
FBAs accounted for almost 50 percent of all reserves, about 10 percentage points higher than domestic banks. Adding the nearly-5 percent share held by FSUBs to the FBA share means that by 2011q3, well over half of all reserves were held by foreign-owned banks in the U.S.

**Foreign banks' volatility: key liabilities-side activities**

The liabilities side of a bank’s balance sheet, broadly speaking, is a record of how credit extension and other asset-side activities are funded. As the financial crisis rapidly deepened in the fall of 2008, serious concerns about how, and even whether, the inter-bank market would continue to function drew attention to banks’ funding practices and liabilities-side activities more broadly. Relatively unremarked at the time were volatile swings in key liabilities-side components for U.S. branches and agencies of foreign-owned banks which, as this section shows, were considerably more extreme than in domestic banks. Also significant is the fact that, while domestic banks’ liabilities-side activities have returned to, or are near, long-run trends, instabilities continue to characterize FBAs’ liabilities-side activities.

**Long-run trends in liabilities**

Figure 6a, showing trends over 2001 to 2011 in the major liabilities-side components of domestic banks, provides a useful context. The substantial drop in borrowed funds that began in 2008q3 reflects the abrupt, substantial, and sustained nature of the turmoil in the inter-bank funding market. In distinct contrast, Figure 6a shows that domestic banks’ deposits growth remained relatively smooth over the entire decade, recording a nearly unbroken upward trend over the entire decade. Worth noting here is that long-run trends in FSUBs’ major liabilities components (not shown) were roughly similar to those of domestic banks. Figure 6b, which decomposes FBAs’ liabilities-side activities into the same major categories, presents a sharply different picture in every respect except for the residual “other liabilities” trends which, as with domestic banks, remained fairly steady, and proportionally small, over the decade. The remainder of this section examines key components underlying the significant changes in deposits, borrowed funds, and net due-to activities of FBAs.

**Deposits**

Retail deposits (i.e., less than U.S.$100,000), despite their
short-term nature, are considered a stable source of funding because they are relatively interest rate inelastic, tend to serve as transactions accounts, and hence are less likely to drain away in such a large and rapid manner that a bank's liquidity position would be imperiled. As noted above, with few exceptions, FBAs do not have access to retail deposits, and as a consequence they have relied more heavily on nonretail deposits, broadly defined as large (U.S.$100,000 and greater) deposits from financial and nonfinancial companies. Such deposits are widely regarded as being more volatile than retail deposits in that they are interest rate-elastic, do not underpin transactions accounts, and hence are more likely to move quickly, in financially meaningfully volumes, in response to actual or anticipated market shocks.

Figure 7 shows the retail-versus-nonretail mix of deposits for FBAs, domestic banks, and FSUBs. The lines in the figure represent the share of total deposits accounted for by large (nonretail) time deposits (U.S.$100,000 or greater) for each of the three banking groups, where such deposits are one commonly used gauge of nonretail, large institutional deposits.22 The top, dashed line shows the traditionally much larger role played by large time deposits for FBAs compared to domestic banks and FSUBs. Starting at the left-hand side of Figure 7, in 1994, large time deposits accounted for about one-third of FBAs' total deposits, a share approximately four times that for both domestic banks and FSUBs.23 Throughout the 1990s and into the early 2000s, amid a banking system-wide struggle for core deposits, large time deposits' share of total deposits rose for all three groups of banks, but the steep upward trajectory for FBAs was magnitudes greater than the modest increases for the other two groups of banks.24 By the beginning of the baseline period, large time deposits' share of total deposits rose for all three groups of banks, but the steep upward trajectory for FBAs was magnitudes greater than the modest increases for the other two groups of banks.24 By the beginning of the baseline period, large time deposits accounted for more than 80 percent of FBA deposits, compared with ratios in the 10-13 percent range for domestic banks and FSUBs. Reliance on large time deposits continued to increase for FBAs through the pre-crisis and crisis periods, while over the same periods domestic banks and FSUBs gradually reduced their large time deposits-to-total deposits shares. By the end of the post-crisis period, domestic banks and FSUBs had reduced those ratios down to, or slightly below, the low levels of the mid-1990s. However, FBAs continued to be overwhelmingly dependent on large time deposits.

Finally, recalling the overall growth of deposits shown in Figure 6a for domestic banks, it is clear that domestic banks' deposit bases became more stable over the post-crisis period.25 FBAs also experienced a rising deposits base, as illustrated in Figure 6b. However, the increase in deposits was made up overwhelmingly of large nonretail rather than retail deposits, and as a consequence, FBAs' deposit base has become substantially less stable in the post-crisis period.

**Borrowed funds**

In addition to deposits, banks fund their activities by borrowing, largely short-term, from other banks and financial institutions. Such wholesale funding has traditionally been more important to FBAs than to commercial banks. Even so, Figure 6b showed that, after increased reliance on borrowed funds during the pre-crisis through crisis periods, during the post-crisis period FBAs...
returned to a lower reliance on borrowed funds, to a degree about equal to that in the more stable early-2000s period.

The top row of Table 2 starts from that relatively benign observation: borrowed funds accounted for 35.46 percent of FBAs’ assets in the baseline period, and 35.57 percent in the post-crisis period. However, underlying details about FBAs’ borrowed funds activities, shown in the remainder of Table 2, paint a less sanguine picture. Below the first line in the table, FBAs’ wholesale funding is divided into two basic categories: funds obtained in the fed funds and the repurchase agreements (repo) markets, and other (largely) short-term funding transactions with banks and other financial institutions. Of these two basic markets for borrowed funds, liquidity and market risk are widely considered to be lower for the fed funds and repo markets. With that in mind, one significant observation about the underlying details in Table 2 is that FBAs reversed their relative reliance on these two basic sources between the baseline and post-crisis period. Specifically, in the relatively stable “baseline” period, fed funds and repos accounted for 70 percent of FBAs’ borrowed funds compared to the 30 percent share provided by other borrowed money. By the post-crisis period, less than half of FBAs’ wholesale funding came from the fed funds and repo markets. Also of note is that, as the use of the fed funds market declined, the share of fed funds supplied by commercial banks operating in the U.S. dropped steeply. The mix of counterparties with whom FBAs dealt in “other borrowed funds” markets also changed in noteworthy ways. Specifically, wholesale funding supplied by other banks dropped decisively, from just over 30 percent of “other borrowed funds” during the baseline period, to less than 8 percent in the post-crisis period, as FBAs became increasingly reliant on nonbank institutions (and, although not shown in Table 2, these nonbanks were located predominantly in foreign countries). Overall, the underlying data on FBAs’ borrowed funds activities strengthens the case for concluding that FBAs’ funding base has become less stable over time.

### Intra-company funding

An individual FBA will engage in a variety of financial transactions with its parent bank and related offices in its home country and other foreign countries. Corporate business strategy determines the exact composition of such intra-company transactions, which include deposits, loans and borrowings, fed funds and repos, claims resulting from clearing activities and foreign exchange transactions, etc. Over a given time period, transactions flows can, and do, go both directions. Conceptually, an FBA’s due-from balance is an extension of credit to its parent company (and/or related offices abroad) and, as with a loan, is booked as an asset; a due-to position represents an amount owed by the FBA to its parent (and/or related offices abroad) and hence is a liability. In their periodic reports, FBAs calculate the net value of these two-way flows over the relevant interval: an individual FBA with net balances due from its parent and related foreign offices will book that value on the assets side of its balance sheet, while an individual FBA with a net due-to position with its parent and related offices abroad records that position on the liabilities side. In a given quarter, an individual FBA records either a net

<table>
<thead>
<tr>
<th>Selected components of wholesale (borrowed) funding</th>
<th>[Beginning of decade] “Baseline” period 2001q2-2003q4</th>
<th>Post-crisis 2010q1-2011q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Borrowed funds as percentage of assets</td>
<td>35.5</td>
<td>35.6</td>
</tr>
<tr>
<td>Fed funds purchased and repos sold: percentage of all borrowed funds</td>
<td>70.4</td>
<td>48.5</td>
</tr>
<tr>
<td>Fed funds purchased with banks in U.S.: as % of all fed funds purchased</td>
<td>37.8</td>
<td>13.8</td>
</tr>
<tr>
<td>Other borrowed funds: percentage of all borrowed funds</td>
<td>29.7</td>
<td>51.5</td>
</tr>
<tr>
<td>Other borrowed funds owed to banks (non-related, in U.S. and abroad): as percentage of total other borrowed funds</td>
<td>30.2</td>
<td>7.8</td>
</tr>
</tbody>
</table>

Table 2: Funding sources for U.S. branches and agencies of foreign banks [FBAs] over a turbulent decade - from wholesale to super-wholesale

Source: FFIEC 002 bank call reports.

1 - Assets exclude due-from position with related foreign offices.

26 For a more detailed breakout, see Goulding and Nolle (2012), Table 9.

27 A comprehensive list of such intra-company transactions can be found in the instructions for Schedule M of the FFIEC 002 call report that FBAs file quarterly.

28 Note that FBAs do not separately break out and report transactions solely with their parent company on their quarterly call reports.
due-from or net due-to balance, but not both. In contrast, in any given quarter, some FBAs have net due-from balances and others have net due-to balances, and hence aggregated industry statistics for FBAs show net balances on both the assets and liabilities sides.32

Figure 8 illuminates these FBAs-in-aggregate concepts by showing the trends in FBAs’ aggregate (net) due-from (dotted line) and (net) due-to (dashed line) positions with their foreign parent companies and related offices abroad. As the top-most box on the right-hand side of Figure 8 explains, as an asset, a due-from balance represents an intra-company flow of funds out of FBAs to parents and related offices abroad. The box just below that explains, in turn, that due-to balances represent intra-company inflows of funds to FBAs from parents and related offices.30

Figure 8 also includes a third line, the solid Net-due To Related Foreign Offices; that line is replicated from Figure 6b above. It is the difference between the two net-due positions for FBAs in aggregate. Although not based on standard accounting practice, the construct follows the analytic convention established by the Federal Reserve System for its weekly H.8 statistical release as a way to quantify and characterize the direction of overall transactions flows between FBAs in aggregate and their related offices abroad.31 By convention, the resultant value (so to speak, the “netting of the two net-dues”) is calculated as FBAs’ aggregate net due-to positions minus their aggregate net due-from positions. Because the order of the “net netting” is liabilities-minus-assets, in the H.8 release “net due-to related foreign offices” is recorded on the liabilities side of the aggregate balance sheet - even if the value is negative.32

The two left-hand side boxes in Figure 8 help explain how to interpret a given value along the solid ‘net due-to’ line. A positive value (i.e., a value above the horizontal axis at 0) along the solid line means that, for the given time period, FBAs’ aggregate due-to positions were greater than their aggregate due-from positions. As the top left-hand box in Figure 8 explains, in terms of funding flows, positive values along the solid ‘net due-to’ line mean that FBAs in aggregate owe more to parents and related offices abroad than those offices owe them, a situation that results in an intra-company funding inflow to FBAs in aggregate from their parents and related offices abroad. Conversely, a negative value along the solid line means that, for the given time period, FBAs’ aggregate ‘due-to’ balances were smaller than their ‘due-from’ balances; that is, FBAs had extended a greater amount of credit to related foreign offices than those offices had extended to the FBAs. The box in the lower left-hand side quadrant indicates that such a result represents a funding outflow from FBAs to parents and related offices abroad.

What particularly stands out in Figure 8 about all three components are their large dollar volumes and erratic behavior, especially in the right-hand “half” of the figure. Starting from the left-hand side and considering all three of the lines together, because ‘due-to’ balances and ‘due-from’ balances registered

---

29 See, for example, any issue of the Federal Reserve System’s quarterly statistical release (4.30) Assets and Liabilities of U.S. Branches and Agencies of Foreign Banks.
30 Both the ‘due-to’ (a liability) and ‘due-from’ (an asset) components are never negative, consistent with standard accounting practices.
31 The underlying data for the H.8 Assets and Liabilities of Commercial Banks in the United States is collected by the Federal Reserve System on its FR 2644 report (Weekly Report of Selected Assets and Liabilities of Domestically Chartered Commercial Banks and U.S. Branches and Agencies of Foreign Banks), submitted by a small, representative set of banks and FBAs in the U.S. Note that in the FR 2644, the accounting entry “net due to related foreign offices” is not the “net-net” concept constructed and employed in the H.8 release for FBAs in aggregate.
32 With few exceptions, standard accounting practice does not allow booking a negative value. A common exception to this principle in bank accounting is that required changes in the “stock” of reserves for expected loan losses, always booked on the assets side, may for a given quarter be negative, in which case the entry is a “contra asset.”
similar values throughout the 1990s, the difference between them, respectively, was small, and hence the solid ‘net due-to’ line was close to the horizontal axis during those years. In the event, for most of that time period, FBAs’ aggregate net ‘due-to’ position was positive, indicating that FBAs as a group saw small but steady funding inflows from parent and other related foreign offices.

The direction of the net funding flows changed decisively in the early 2000s. The negative portions of the solid net ‘due-to’ line indicate that, for a sustained period of time, FBAs and their parents engineered funding outflows to the parents and other related offices abroad. These net outflows grew at an increasing rate through the 2004-2007 pre-crisis period, almost entirely based on steeply rising due-from transactions, as the steeply rising dotted line over that period shows.

As the crisis erupted, these intra-company funding flows changed in several significant ways. Between 2008q2 and 2008q4, net ‘due-to’ balances more than halved, to U.S.$278 billion, from their historical negative “peak” of U.S.$628 billion. This large decrease in funding outflow from FBAs to parents and other related foreign offices was due to large changes in both the asset-side and liabilities-side components of intra-company flows. On their asset side, FBAs sharply reduced funding outflows to parents and other offices abroad, with ‘due-from’ balances declining more than 30 percent over the 2008q2 - 2008q4 period (from U.S.$780 billion to U.S.$538 billion). On their liabilities side, FBAs’ ‘due-to’ balances increased 70 percent, from U.S.$152 billion to U.S.$260 billion, over that same three-quarters period. Hence, as the financial crisis escalated amid a global scramble for liquidity, FBAs as a group not only reduced funding outflows to their parents and related foreign offices, but also received funding inflows from their parents and other offices abroad.

Subsequently, FBAs were able to access liquidity-provision facilities set up by the Federal Reserve System in response to the crisis and, as a consequence, the crisis-induced changes in both components of their intra-company funding flows reversed direction.33 The dotted line in the top portion of Figure 8 shows the increase, over the first part of 2009, in FBAs’ funding outflows to parents and related offices, and the dashed line shows that, at the same time, FBAs somewhat reduced their funding inflows from parents and other offices abroad. The combined effect of those trends was that FBAs’ net ‘due-to’ balances moved back toward pre-crisis levels, as the sharp drop in the solid line over the 2009q1-2009q3 period illustrates.

Thereafter, what might have seemed, in early-2010, to be a return to a traditional, long-run pattern of intra-company funding flows instead turned into an “overshoot” of historic proportions. Two dynamics were at work over the 2010-2011q3 period. First, over the entire period, FBAs reduced their funding outflows to parents and related foreign offices, as the undulating drop in the dotted line at the far right-hand side of Figure 8 shows. Second, beginning in 2011q1, FBAs began to receive historically large intra-company funding inflows, as shown by the upward surge in the dashed line in the figure. The net result of those two trends was that, for the first time in more than a decade, FBAs as a group received net funding inflows from parents and related foreign offices, as the positive values for the solid line indicate. Indeed, as of 2011q3, net funding inflows to FBAs had reached a level double the average characterizing the 1990s.

FSUBs and domestic banks also engage in intra-company funding activities with their offices abroad, and a comparison of their volumes and patterns with those of FBAs is useful. Ahead of such a discussion, it is important to be clear about the perspective conveyed by the call report data for both domestic banks and FSUBs. As a U.S.-chartered commercial bank, an FSUB (or a domestic bank) reports activities on a “consolidated” basis, i.e., for itself plus any branches or subsidiaries it owns, including those located in foreign countries.34 Consolidated reporting in this sense therefore covers “down-stream” operations only. It does not include “up-stream” information on an FSUB as part of a larger corporate network under a foreign-based parent company; i.e., from the point of view of the call report data, FSUBs are the parent bank vis-à-vis their network of branches and subsidiaries, both in the U.S. and abroad, in exactly the same

33 See Goldberg and Skeie (2011) for a description of discount window borrowing by U.S. branches and agencies of foreign banks during the crisis. Cetorelli and Goldberg (2012) look at parent bank-U.S. FBAs funding flows as from the crisis; see also Miu et al. (2012), and Cetorelli and Goldberg (2011).

34 Specifically, instructions for the FFIEC 031 quarterly call report state that it “is to be filed by banks with branches and consolidated subsidiaries in U.S. territories and possessions, Edge or Agreement subsidiaries, foreign branches, consolidated foreign subsidiaries, or International Banking Facilities.”
way as are domestic banks. In the event, FSUBs’ intra-company flows are small in both an absolute sense and in proportion to the total funding picture for FSUBs. Furthermore, FSUBs’ net ‘due-to’ positions follow patterns very similar to those of domestic banks. As shown in Figure 6a, domestic banks’ net ‘due-to’ activities have always been a proportionally small, and relatively stable, component of their liabilities.

Nevertheless, relative to FBAs’ net ‘due-to’ activities, domestic banks’ intra-company funding flows were large, as Figure 9 makes clear. Figure 9 considers the combined impact of the intra-company funding flows of all three groups of banks on U.S. banking system liquidity, as follows. The solid red line in the figure replicates FBAs’ net ‘due-to’ funding flows from Figure 8; the dashed line shows domestic banks’ net ‘due-to’ flows; and the dotted line shows FSUBs’ net ‘due-to’ activities. The bars in Figure 9 show the combined effect of the three intra-company/ cross-border funding flows in terms of net inflows into, or outflows out of, the U.S. banking market. As the box in the upper left-hand quadrant of the figure explains, positive values along any line represent funding inflows to the U.S. banking market and negative values represent funding outflows from the U.S. banking market. Adding the positive and negative values of the three flows together for each time period results in a “net” balance, either positive or negative, for that time period; the bars show the pattern of those net balances over time. The small negative bars for most of the pre-crisis period (roughly, 2001-2007) show small declines in U.S. banking system liquidity, as intra-company funding flows to “parent” domestic banks and FSUBs by their foreign offices were not enough to offset FBAs’ funding outflows from the U.S. market to their parents and related offices abroad. During and after the crisis, the net effects on U.S. banking system liquidity were larger and more erratic, as the bars in the right-hand third reveal. That said, in a broader context, the net impact on overall U.S. banking system liquidity via those intra-company liquidity management activities remained small, averaging less than 1 percent of U.S. banking system assets over almost every time period.

Summary and conclusions

The purpose of this study is to call attention to the nature and importance of the “under-count” of foreign-owned banks’ U.S. presence, and explain how this issue can be addressed. The first section of the paper explained the nature of the under-count, showing that a proper accounting for foreign-owned banks results in an increase of about 20 percent in the size of the U.S. commercial banking market “pie” as compared to traditional, widely referenced statistical releases. That section emphasizes in particular that, on conceptual, economic, regulatory, and legal grounds, U.S. branches and agencies of foreign banks (FBAs) should be fully accounted for in statistical profiles and analyses of “the” U.S. banking industry.

The following sections show why, in addition to their aggregate share of the U.S. banking market, foreign banks in the U.S. have become increasingly important from an economic policy point of view. Those sections identify the most important activities underlying trends on both sides of foreign-owned banks’ balance sheets over the past ten years of historic turbulence in U.S. and

---

35 There is no separate identification of FSUBs’ transactions with their foreign parents in the call report data. Nevertheless, FSUBs’ net due-to data provide some interesting insights in that regard, as Goulding and Nolle (2012) explain. They note, however, that those inferences are not of the first order of magnitude, especially relative to FBAs’ intra-company funding flows. For this reason, the current study does not delve into the details of FSUBs’ net due-to flows.

36 However, see Goulding and Nolle (2012), section V.D for one interesting, although not especially important, exception to that generalization. See also that study’s detailed description of trends in the underlying ‘due-to’ and ‘due-from’ components of domestic banks’ net ‘due-to’ activities.
global financial markets. The clear conclusion emerging from that analysis is that, on both sides of the balance sheet, foreign banks in the U.S. - especially FBAs - were, and to a large extent remain, more volatile than U.S.-owned "domestic banks."

Indeed, the measurably greater volatility characterizing FBAs is the central story line in both the asset-focused and liabilities-focused sections of this paper. On the assets side, in addition to their significantly anemic C&I lending performance, FBAs' cash assets activities are of particular note. The investigation shows that FBAs' extreme emphasis, post-crisis, on increasing cash balances not only dominates the assets side of their aggregate balance sheet, but now exceeds domestic banks' balances. In addition, FBA cash balances are almost entirely composed of excess reserves held at Federal Reserve Banks, such that FBAs accounted for approximately one-third of total liabilities on the Federal Reserve System's balance sheet in 2001q3.

Changes in FBAs' liabilities-side activities are, if anything, even more significant. The nature of these changes can be summarized as follows: FBAs, traditionally more dependent on wholesale funding than domestic banks, are currently reliant on an unprecedented level of funding from wholesale and super-wholesale sources of borrowed funds. At the same time, as a group FBAs are, for the first time this century, net borrowers from their foreign parents and related offices abroad. The overall profile is one of much less stable funding than at any time in the past two decades. Under those circumstances, the importance of the emerging literature on the effects of banking and financial system problems in Eurozone countries on U.S.-based FBAs via their parents looms ever larger.

References
Valuing financial services firms

Aswath Damodaran
Kerschner Family Professor in Finance Education, Professor of Finance, Stern School of Business, New York University

Abstract
Valuing banks, insurance companies, and investment banks has always been a daunting exercise, but the rolling market crises of the last few years have made a difficult job even more so. There are two key measurement problems that you face in valuing financial services firms. The first is that the cash flows cannot be easily estimated, since many of the ingredients needed are not clearly defined. The second is that most financial services firms operate under regulatory frameworks that govern how they are capitalized, where they invest, and how fast they can grow. Consequently, changes in the regulatory environment can create large shifts in value. In this paper, we confront both factors. We argue that financial services firms should be valued using equity valuation models, rather than enterprise valuation models, and with actual or potential dividends used as cash flows. The two key numbers that drive value are the cost of equity, which is a function of the risk that emanates from the firm’s investments, and the return on equity, which is determined by the company’s investment choices and regulatory restrictions.
Valuing financial services firms

Introduction
The principles of valuation are well established. Thus, the value of a business or asset can be estimated in one of two ways. The expected cash flows from owning the business can be discounted back at a risk-adjusted rate to arrive at an intrinsic value. Alternatively, the asset or business can be valued by looking at how the market is pricing similar assets or businesses in a relative valuation. While these principles should apply just as much when you are valuing banks, insurance companies, and other financial services firms as they do when valuing other firms, these firms pose special challenges for an analyst attempting to value them, for three reasons. The first is that the nature of their businesses makes it difficult to define both debt and reinvestment, making the estimation of cash flows much more difficult. The second is that these firms tend to be heavily regulated and changes in regulatory requirements can have a significant effect on value. The third is that the accounting rules that govern accounting for financial services firms have historically been very different from the accounting rules for other firms, with assets being marked to market more frequently.

In this paper, we begin by considering what makes financial services firms unique and ways of dealing with the differences. We then look at how best we can adapt discounted cash flow models to value financial services firms by laying out four alternatives – the classic dividend discount model, a creative version of a cash flow to equity model, an excess return model, and an asset-based model. Using these models, we derive the key drivers of value for a financial services firm, and use them to examine how relative valuation works within financial services firms and what multiples may work best with these firms.

Financial services firms – the big picture
Any firm that provides financial products and services to individuals or other firms can be categorized as a financial services firm. We would categorize financial services businesses into four groups from the perspective of how they make their money. A bank makes money on the spread between the interest it pays to those from whom it raises funds and the interest it charges those who borrow from it, and from other services it offers its depositors and its lenders. Insurance companies make their income in two ways. One is through the premiums they receive from those who buy insurance protection from them and the other is income from the investment portfolios that they maintain to service the claims. An investment bank provides advice and supporting products for other firms to raise capital from financial markets or to consummate deals such as acquisitions or divestitures. Investment firms provide investment advice or manage portfolios for clients. Their income comes from advisory fees for the advice and management and sales fees for investment portfolios. With the consolidation in the financial services sector, an increasing number of firms operate in more than one of these businesses. For example, many money center banks like Bank of America and Citigroup operate in all four businesses. At the same time, however, there remain a large number of small banks, boutique investment banks, and specialized insurance firms that still derive the bulk of their income from one source.

How big is the financial services sector in the United States? We would not be exaggerating if we said that the development of the U.S. economy would not have occurred without banks providing much of the capital for growth, and that insurance companies predate both equity and bond markets as pioneers in risk sharing. Financial services firms have been the foundation

<table>
<thead>
<tr>
<th>Financial sector</th>
<th>Number</th>
<th>Market capitalization</th>
<th>% of overall market</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banks</td>
<td>471</td>
<td>$1,068,475</td>
<td>5.28%</td>
</tr>
<tr>
<td>Financial services</td>
<td>225</td>
<td>$593,952</td>
<td>2.93%</td>
</tr>
<tr>
<td>Insurance (life)</td>
<td>30</td>
<td>$161,374</td>
<td>0.80%</td>
</tr>
<tr>
<td>Insurance (property/casualty)</td>
<td>49</td>
<td>$174,876</td>
<td>0.86%</td>
</tr>
<tr>
<td>Securities brokerage</td>
<td>28</td>
<td>$156,940</td>
<td>0.77%</td>
</tr>
<tr>
<td>Thrift</td>
<td>148</td>
<td>$38,342</td>
<td>0.19%</td>
</tr>
<tr>
<td>All financial services firms</td>
<td>951</td>
<td>$2,193,959</td>
<td>10.83%</td>
</tr>
</tbody>
</table>

Table 1: Financial services firms – market capitalizations on January 1, 2012 (in millions)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Banks</td>
<td>4.68%</td>
<td>4.80%</td>
<td>5.16%</td>
<td>5.51%</td>
<td>5.28%</td>
</tr>
<tr>
<td>Financial services</td>
<td>4.50%</td>
<td>3.04%</td>
<td>2.63%</td>
<td>2.60%</td>
<td>2.93%</td>
</tr>
<tr>
<td>Insurance (life)</td>
<td>1.33%</td>
<td>0.35%</td>
<td>1.08%</td>
<td>1.10%</td>
<td>0.80%</td>
</tr>
<tr>
<td>Insurance (property/casualty)</td>
<td>6.28%</td>
<td>7.40%</td>
<td>0.97%</td>
<td>0.95%</td>
<td>0.86%</td>
</tr>
<tr>
<td>Securities brokerage</td>
<td>1.42%</td>
<td>0.87%</td>
<td>1.16%</td>
<td>1.07%</td>
<td>0.77%</td>
</tr>
<tr>
<td>Thrift</td>
<td>0.30%</td>
<td>0.36%</td>
<td>0.26%</td>
<td>0.32%</td>
<td>0.19%</td>
</tr>
<tr>
<td>All financial services firms</td>
<td>18.51%</td>
<td>16.81%</td>
<td>11.26%</td>
<td>11.55%</td>
<td>10.83%</td>
</tr>
</tbody>
</table>

Table 2: Financial services firms – market capitalization as percentage of the U.S. market (1 January 2008 – 1 January 2012)
of the U.S. economy for decades and the results can be seen in many measures. Table 1 summarizes the market capitalization of publicly traded banks, insurance companies, brokerage houses, investment firms, and thrifts in the U.S. at the end of 2011 and the proportion of the overall equity market that they represented at the time. In addition, the financial services sector, in the 2002 economic census, accounted for 6% of all full-time employees in the U.S.

The last 5 years have been tumultuous years for the overall market, and even more so for financial services firms. To the extent that the crisis of 2008 can be traced to the failures of banks and other financial services firms, it is worth looking at how the market capitalization of these firms has changed between 2007 and 2011, relative to the market, at least in the U.S. (Table 2).

Financial services firms have become a smaller proportion of the overall market, but the bulk of the change has occurred in the property/casualty insurance companies, where the restructuring of AIG has had a dramatic impact.

What about outside the U.S.? To answer this question, we looked at the proportion of overall market value accounted for by financial services firms globally, as well as just in emerging markets, in Figure 1.

In emerging markets, banks account for a larger proportion of the overall market value than they do in developed markets. In addition, financial services firms range the spectrum, from small to large, mature to growing, in developed and in emerging markets, and it is quite clear that no one template will value all financial services firms and that we have to be flexible in our use of valuation models.

Characteristics of financial services firms
In this section, we will focus on four key differences between financial services firms and the rest of the market, and look at why these differences can create estimation issues in valuation. The first is that many categories (albeit not all) of financial services firms operate under regulatory constraints on how they run their businesses and how much capital they need to set aside to keep operating. The second is that accounting rules for recording earnings and asset values at financial services firms are at variance with accounting rules for the rest of the market. The third is that debt for a financial services firm is more akin to raw materials than to a source of capital; the notion of cost of capital and enterprise value may be meaningless as a consequence. The final factor is that the defining reinvestment (net capital expenditures and working capital) for a bank or insurance company may be not just difficult, but impossible, and cash flows cannot be easily computed.

The regulatory overlay
Financial services firms are heavily regulated all over the world, though the extent of the regulation varies from country to country. In general, these regulations take three forms. First, banks and insurance companies are required to maintain regulatory capital ratios, based upon their risk exposure, to ensure that they do not expand beyond their means, and put

---

**Figure 1: Financial services firms versus rest of the market: global – January 2012**

- Global:
  - Banks 9%
  - Financial services 2%
  - Insurance companies 2%
  - Securities brokerage 1%
  - Rest of market 86%

- Emerging markets:
  - Banks 15%
  - Financial services 1%
  - Insurance companies 3%
  - Securities brokerage 1%
  - Rest of market 80%
their claimholders or depositors at risk. Second, financial services firms are often constrained in terms of where they can invest their funds. For instance, until a decade ago, the Glass-Steagall Act in the U.S. restricted commercial banks from investment banking activities as well as from taking active equity positions in nonfinancial services firms. Third, the entry of new firms into the business is often controlled by the regulatory authorities, as are mergers between existing firms.

Why does this matter? From a valuation perspective, assumptions about growth are linked to assumptions about reinvestment. With financial services firms, these assumptions also have to be scrutinized to ensure that they pass regulatory constraints. There might also be implications for how we measure risk at financial services firms. If regulatory restrictions are changing or are expected to change, it adds a layer of uncertainty (risk) to the future, which can have an effect on value. Put more simply, to value banks, insurance companies, and investment banks, we have to be aware of the regulatory structure that governs them.

Differences in accounting rules
The accounting rules used to measure earnings and record book values are different for financial services firms than for the rest of the market, for two reasons. The first is that the assets of financial services firms tend to be financial instruments (bonds, securitized obligations) that are often traded. Not surprisingly, marking assets to market value has been an established practice in financial services firms, well before other firms even started talking about fair value accounting. The second is that the nature of operations for a financial services firm is such that long periods of profitability are interspersed with short periods of large losses; accounting standards have been developed to counter this tendency and create smoother earnings.

Mark to market – if the new trend in accounting is towards recording assets at fair value (rather than original costs), financial services firms operate as a long-standing laboratory for this experiment. After all, accounting rules for banks, insurance companies, and investment banks have required that assets be recorded at fair value for decades, based upon the argument that most of a bank’s assets are traded, have market prices, and therefore do not require too many subjective judgments on the part of accountants. To the extent that some, or a significant portion, of the assets of financial services firms are marked to market, and the assets of most nonfinancial services firms are not, we face two problems. The first is in comparing ratios based upon book value (both market to book ratios like price to book and accounting ratios like return on equity) across financial and nonfinancial services firms. The second is in interpreting these ratios, once computed. While the return on equity for a nonfinancial services firm can be considered a measure of return earned on equity invested originally in assets, the same cannot be said about the return on equity at financial services firms, where the book value of equity measures not what was originally invested in assets but an updated market value (or at least the accountant’s measure of that market value).

Loss provisions and smoothing out earnings – consider a bank that makes money the old-fashioned way – by taking in funds from depositors and lending these funds out to individuals and corporations at higher rates. While the rate charged to lenders will be higher than that promised to depositors, the risk that the bank faces is that lenders may default, and the rate at which they default will vary widely over time – low during good economic times and high during economic downturns. Rather than write off the bad loans, as they occur, banks usually create provisions for losses that average out losses over time and charge this amount against earnings every year. Though this practice is logical, there is a catch, insofar as the bank is given the responsibility of making the loan loss assessment. A conservative bank will set aside more for loan losses, given a loan portfolio, than a more aggressive bank, and this will lead to the latter reporting higher profits during good times.

Debt and equity
There are only two ways to raise funds to finance a business – debt and equity. While this is true for all firms, financial services firms differ from nonfinancial service firms on three dimensions:

Debt is raw material, not capital – when we talk about capital for nonfinancial services firms, we tend to talk about both debt and equity. A conventional business raises capital from both equity investors and bondholders (and banks) and uses these funds to finance its investments. When we value the firm, we value the assets owned by the firm, rather than just the value of its equity. With a financial services firm, debt has a different connotation. Rather than viewing debt as a source of capital,
most financial services firms view it as a raw material. In other words, debt is to a bank what steel is to an automobile company, something to be molded into products which can then be sold at a higher price and yield a profit. Consequently, capital at financial services firms is narrowly defined as including only equity capital. This narrow definition of capital is reinforced by the regulatory authorities, whose core measures of regulatory capital are built around equity.

- **Defining debt** – the definition of what comprises debt is also murkier with a financial services firm than it is with other types of firms. For instance, should deposits made by customers into their checking accounts at a bank be treated as debt by that bank? Especially on interest-bearing checking accounts, there is little distinction between a deposit and debt issued by the bank. If we do categorize this as debt, the operating income for a bank should be measured prior to interest paid to depositors, which would be problematic since interest expenses are usually the biggest single expense item for a bank.

- **Degree of financial leverage** – even if we can define debt as a source of capital and can measure it precisely, there is a final dimension on which financial services firms differ from other firms. They tend to use more debt in funding their businesses and thus have higher financial leverage than most other firms. While there are good reasons that can be offered for why they have been able to do this historically — more predictable earnings and the regulatory framework are two that are commonly cited — there are consequences for valuation. Since equity is a sliver of the overall value of a financial services firm, small changes in the value of the firm’s assets can translate into big swings in equity value.

**Estimating cashflows is difficult**

We noted earlier that financial services firms are constrained by regulation in terms of both where they invest their funds and how much they invest. If we define reinvestment as necessary for future growth, there are problems associated with measuring reinvestment for financial services firms. Note that we consider two items in reinvestment — net capital expenditures and working capital, and measuring either of these items at a financial services firm can be problematic.

Consider net capital expenditures first. Unlike manufacturing firms that invest in plant, equipment, and other fixed assets, financial services firms invest primarily in intangible assets such as brand name and human capital. Consequently, their investments for future growth are often categorized as operating expenses in accounting statements. Not surprisingly, the statement of cashflows for a bank shows little or no capital expenditures and correspondingly low depreciation. With working capital, we run into a different problem. If we define working capital as the difference between current assets and current liabilities, a large proportion of a bank’s balance sheet would fall into one or the other of these categories. Changes in this number can be both large and volatile and may have no relationship with reinvestment for future growth.

As a result of this difficulty in measuring reinvestment, we run into two practical problems in valuing these firms. The first is that we cannot estimate cashflows without estimating reinvestment. In other words, if we cannot identify how much a company is reinvesting for future growth, we cannot estimate its free cash flows today. The second is that estimating expected future growth becomes more difficult if the reinvestment rate cannot be measured.

**The intrinsic value of a bank**

In a discounted cash flow model, we consider the value of an asset to be the present value of the expected cashflows generated by that asset. In this section, we first lay out the argument that financial services firms should be valued on an equity basis, rather than on a firm basis, and that dividends, for better or worse, are often the only tangible cashflow that we can observe or estimate. Consequently, our focus will be on variants of the equity valuation models and how they can best be used in valuing banks, investment banks, and insurance companies.

**Equity versus firm valuation**

Note the distinction between valuing a firm and valuing the equity in the firm. We value firms by discounting expected after-tax cash flows prior to debt payments at the weighted average cost of capital. We value equity by discounting cashflows to equity investors at the cost of equity. Estimating cashflows prior to debt payments at a weighted average cost of capital is problematic when debt and debt payments cannot be easily identified, which, as we argued earlier, is the case with financial services firms. Equity can be valued directly, however, by discounting cashflows to equity at the cost of equity. Consequently, we would argue for the latter approach for financial services firms.
Even with equity valuation, we have a secondary problem. To value the equity in a firm, we normally estimate the free cashflow to equity, defined as follows:

Free cashflow to equity = net income - net capital expenditures - change in non-cash working capital - (debt repaid - new debt issued)

If we cannot estimate net capital expenditures or non-cash working capital, we clearly cannot estimate the free cashflow to equity. Since this is the case with financial services firms, we have four choices. The first is to use dividends as cashflows to equity and assume that firms, over time, pay out their free cashflows to equity as dividends. Since dividends are observable, we therefore do not have to confront the question of how much firms reinvest.

The second is to adapt the free cashflow to equity measure to allow for the types of reinvestment that financial services firms make. For instance, given that banks operate under a regulatory capital ratio constraint, it can be argued that these firms have to increase regulatory capital in order to make more loans in the future. The third is to keep the focus on excess returns, rather than on earnings, dividends, and growth rates, and to value these excess returns. In the final approach, we value financial services firms based upon net asset values, where we value the assets today and subtract out debt.

Dividend discount models

In the basic dividend discount model, the value of a stock is the present value of the expected dividends on that stock. While many analysts view the model as old-fashioned, it retains a strong following among analysts who value financial services companies, because of the difficulties we face in estimating cashflows. In this section, we will begin by laying out the basic model and then consider ways in which we can streamline its usage, when valuing financial services companies.

The standard model

If we start with the assumption that equity in a publicly traded firm has an infinite life, we arrive at the most general version of the dividend discount model:

\[
\text{Value per share of equity} = \sum_{t=1}^{\infty} \frac{DPS_t}{(1 + k_e)^t}
\]

where,

- \(DPS_t = \text{Expected dividend per share in period } t\)
- \(k_e = \text{Cost of equity}\)

In the special case where the expected growth rate in dividends is constant forever, this model collapses into the “Gordon growth model.”

\[
\text{Value per share of equity in stable growth} = \frac{DPS_1}{k_e - g}
\]

In this equation, \(g\) is the expected growth rate in perpetuity and \(DPS_1\) is the expected dividends per share next year. In the more general case, where dividends are growing at a rate that is too high to be sustainable in the long term (called the extraordinary growth period), we can still assume that the growth rate will become sustainable (and constant) at some point in the future. This allows us to then estimate the value of a stock, in the dividend discount model, as the sum of the present values of the dividends over the extraordinary growth period and the present value of the terminal price, which itself is estimated using the Gordon growth model.

Value per share of equity in extraordinary growth =

\[
\sum_{t=1}^{n} \frac{DPS_t}{(1 + k_e)^t} + \frac{DPS_{n+1}}{(1 + k_e)^n (1 + k_{e,n+1})}
\]

The extraordinary growth is expected to last \(n\) years, \(n+1\) is the expected growth rate after \(n\) years and \(k_e\) is the cost of equity (hg: high growth period and st: stable growth period).

While the dividend discount model is intuitive and has deep roots in equity valuation, there are dangers in using the model blindly. There are many analysts who start with the current dividends as a base, apply a growth rate to these earnings, based on either history or forecasts, and compute a present value. For the model to yield a value that is reasonable the assumptions have to be internally consistent, with the expected growth rate numbers gelling with the dividend forecasts and risk measures. In addition, we are assuming that the current dividends reflect what the firm has available to pay out, rather than the whims of management.

A consistent dividend discount model

Looking at the inputs into the dividend discount model, there are three sets of inputs that determine the value of equity. The first is the cost of equity that we use to discount cashflows, with the possibility that the cost may vary across time, at least for some firms. The second is the proportion of the earnings that we assume will be paid out in dividends: this is the dividend payout.
ratio and higher payout ratios will translate into more dividends for any given level of earnings. The third is the expected growth rate in dividends over time, which will be a function of the earnings growth rate and the accompanying payout ratio; in general, the more you pay out in dividends, the lower your expected growth rate will tend to be. In addition to estimating each set of inputs well, you also need to ensure that the inputs are consistent with each other.

Risk and cost of equity
As with any publicly traded company, the cost of equity for a financial services firm has to reflect the portion of the risk in the equity that cannot be diversified away by the marginal investors in the stock. This risk can be estimated using a beta (in the capital asset pricing model) or betas (in a multi-factor or arbitrage pricing model). The broad principles on estimating cost of equity are simple:

- **Reflect risk of business** – the cost of equity should reflect the riskiness of the business or businesses that a financial services firm derives its revenues from. Thus, the cost of equity for a bank that chooses to lend to “riskier” customers should be higher than for one that lends only to “safe” customers. In a similar vein, the cost of equity for a bank that derives more of its revenues from proprietary trading should be higher than the cost of equity for one that gets all its revenues from conventional banking.

- **Can be correlated with growth** – financial services firms that push for more growth often have to enter riskier businesses. Consequently, you would expect the cost of equity for higher growth banks and insurance companies to be higher than for more mature companies in the same space.

- **Can change over time** – if risk is a function of your business mix and expected growth, it follows that the risk of a financial services firm should change over time, as its business mix changes and its growth potential subsides.

There are clearly regulatory and measurement issues that are specific to financial services firms and we will return to address those later in the paper.

Growth and payout
There is an inherent tradeoff between dividends and growth. When a company pays a larger segment of its earnings as dividends, it is reinvesting less and should thus grow more slowly. With financial services firms, this link is reinforced by the fact that the activities of these firms are subject to regulatory capital constraints: banks and insurance companies have to maintain equity (in book value terms) at specified percentages of their activities. When a company is paying out more in dividends, it is retaining less in earnings; the book value of equity increases by the retained earnings. In recent years, in keeping with a trend that is visible in other sectors as well, financial services firms have increased stock buybacks as a way of returning cash to stockholders. In this context, focusing purely on dividends paid can provide a misleading picture of the cash returned to stockholders. An obvious solution is to add the stock buybacks each year to the dividends paid and to compute the composite payout ratio. If we do so, however, we should look at the number over several years, since stock buybacks vary widely across time—a buyback of billions in one year may be followed by three years of relatively meager buybacks, for instance.

To ensure that assumptions about dividends, earnings, and growth are internally consistent, we have to bring in a measure of how well the retained equity is reinvested: the return on equity is the variable that ties together payout ratios and expected growth:

\[
\text{Expected growth in earnings} = \text{return on equity} \times (1 - \text{dividend payout ratio})
\]

For instance, a bank that pays out 60% of its earnings as dividends and earns a return on equity of 12% will have an expected growth rate in earnings of 4.8%. However, firms can deliver growth rates that deviate from this expectation, if the return on equity is changing.

\[
\text{Expected growth}_{\text{EPS}} = (1 - \text{Payout ratio})(\text{ROE}_{t+1}) + \frac{\text{ROE}_{t+1} - \text{ROE}_t}{\text{ROE}_t}
\]

Thus, if the bank is able to improve the return on equity on existing assets from 10% to 12%, the efficiency growth rate in that year will be 20%. However, efficiency growth is temporary and all firms will ultimately revert to the fundamental growth relationship.
The linkage between return on equity, growth, and dividends is therefore critical in determining value in a financial services firm. At the risk of hyperbole, the key number in valuing a bank is not dividends, earnings, or growth rate, but what we believe it will earn as “return on equity in the long term.” That number, in conjunction with payout ratios, will help in determining growth. Alternatively, the return on equity, together with expected growth rates, can be used to estimate dividends. This linkage is particularly useful, when we get to stable growth, where growth rates can be very different from the initial growth rates. To preserve consistency in the valuation, the payout ratio that we use in stable growth, to estimate the terminal value, should be:

\[
Payout \text{ ratio in stable growth} = 1 - \frac{g}{\text{ROE}_{\text{stable growth}}}
\]

The risk of the firm should also adjust to reflect the stable growth assumption. In particular, if betas are used to estimate the cost of equity, they should converge towards one in stable growth.

**Cashflow to equity models**

At the beginning of this discussion, we noted the difficulty in estimating cashflows when net capital expenditures and non-cash working capital cannot be easily identified. It is possible, however, to estimate cashflows to equity for financial services firms if we define reinvestment differently. The cashflow to equity is the cashflow left over for equity investors after debt payments have been made and reinvestment needs met. With financial services firms, the reinvestment generally does not take the form of plant, equipment, or other fixed assets. Instead, the investment is in regulatory capital: this is the capital as defined by the regulatory authorities, which, in turn, determines the limits on future growth. The key to using this model then becomes an understanding of the regulatory structure governing financial services firms.

**Regulatory capital**

One of the legacies of the Great Depression was the introduction of a regulatory overlay to prevent banks from collapsing and the social costs from the ensuing bank runs. As part of that overlay, banks and other financial services companies have been required to hold “equity” capital to cover potential losses and shortfalls from their operations, with the required holding being a function of the scale and risk of their business. In the decades since, some firms have tested the regulatory constraints by increasing the risk exposure of their businesses without increasing their regulatory capital holdings and the regulatory authorities have responded to the inevitable “crisis” by increasing capital requirements and/or oversight of business risk.

The framework for capital regulation is contained in the Basel accords, though individual countries have their own supplemental regulations. These regulations are built around two measurement principles:

- **Risk-adjusted assets** – building on the proposition that the capital set aside has to be greater for banks that hold riskier assets, bank regulators have created adjustment mechanisms that try to take risk into consideration when measuring assets. Rather than set different capital ratios for asset holdings with different risk levels, regulations have been built around adjusting the value of the assets for risk, with higher risk translating into higher risk-adjusted values.

- **Regulatory capital** – banks are required to maintain minimum capital to sustain their operations, and there are two measures of capital. Tier 1 capital is the narrower measure and is composed primarily of common equity but also includes non-cumulative preferred stock. Tier 2 capital is a broader measure of capital that includes subordinated debt and cumulative preferred stock.

Individual banks are free to hold more capital, if they so desire, and more conservative banks therefore will set their regulatory capital ratios at above the regulatory minimum.

If regulatory capital has to be maintained at a percentage of risk-adjusted assets or some measure of operations, there are three factors that determine how much a company will have to invest in regulatory capital in the future. The first is the current level of regulatory capital relative to a target capital ratio; this target will reflect not only regulatory requirements but also the degree of risk aversion among the firm’s managers. If a financial services firm has too little regulatory capital relative to its target, it will have to set aside more of its earnings into regulatory capital, thus leaving less to be paid out in dividends. In contrast, a bank that is over-capitalized may be able to pay much higher dividends, given its earnings, as the drawdown of regulatory capital will release more cash for stockholders. The second is the expected growth in operations over time; even a bank that is at its desired
The regulatory capital ratio will have to reinvest more in regulatory capital if it expects growth of 10% a year for the next few years. The third factor is the degree of risk that the firm chooses to expose itself to: a bank or investment bank that chooses to enter riskier businesses (perhaps in search of profitability) will find itself needing to invest more in regulatory capital to reflect the higher risk.

When valuing a bank using the FCFE model, you have to estimate a regulatory capital ratio for the bank and you can use the model on the following approaches:

- Leave the regulatory capital ratio at the current level, on the assumption that barring information to the contrary, this is your best estimate for the future. Thus, the reinvestment in regulatory capital will track the expected growth in risk-adjusted assets over time.
- Obtain a target capital ratio from the bank’s management (and some banks are more transparent than others), based upon their expectations of regulatory changes and their desire to maintain a buffer. You can then make a judgment of the time period over which the firm will move from the existing ratio to the target ratio and incorporate the resulting reinvestment into your cashflow forecasts.
- If you do not feel comfortable staying with the existing regulatory capital ratio and have no indication about a target ratio from the management, you can assume that the firm will move towards the industry norm. This industry norm can be defined as the median regulatory capital ratio across all banks or, at least, the banks that would be viewed as being part of the peer group.

What is a normal regulatory capital ratio? To answer this question, we look at the distribution of tier 1 capital, as a percentage of risk adjusted assets, at U.S. banks, in November 2012, in figure 2.

Note the variation in capital holdings across banks, with some holding more regulatory capital than others, as a percent of risk adjusted assets. The median across all U.S. banks in November 2012 was 14.52% but the highest tier 1 capital ratio was 33.4% and the lowest was in the single digits.

There is also a secondary factor at play that stems from what the regulatory authorities count towards “regulatory capital.” For instance, capital raised from non-cumulative preferred stock can be counted towards tier 1 capital and a bank that issues preferred stock may not need to invest as much of its common earnings back into the business. This benefit, though, has to be offset against the preferred dividends that will have to be paid each period, which will lower the net income. In Figure 3, we look at a...
more primitive measure of capital, book value of common equity, as a percentage of risk-adjusted assets at U.S. banks in November 2012. For most banks, the core portion of tier 1 capital comes from common equity, though it is supplemented with non-cumulative preferred stock and other non-common equity capital, to different degrees, by different banks.

Implementing an FCFE model
To implement an FCFE model, you need two ingredients. The first is the expected net income over time, which will be a function not only of the profitability of the businesses that the financial services firm is involved in but will also be determined by the cash flow claims of lenders, preferred stockholders, and other non-common claimholders. The second is the investment in regulatory capital, which will be a function of both the degree to which the financial services firm is under or over-capitalized to begin the process and the expected growth rate in its risk-adjusted assets.

$\text{FCFE}_{\text{financial services firm}} = \text{net income} - \text{reinvestment in regulatory capital}$

Note that the reinvestment in regulatory capital can exceed the net income under two scenarios. The first is if you have a severely under-capitalized bank that has to replenish its regulatory capital to meet a standard; the undercapitalization itself may have been triggered by losses on existing assets (loans or security holdings). The second is a high-growth bank that has to keep its regulatory capital growing at a high rate to sustain its asset growth. In either case, the negative FCFE will have to be covered with new equity issues over time, thus depressing what you are willing to pay for the common stock today. If the net income is greater than the reinvestment, the FCFE will be positive and can be viewed as potential dividends. By discounting these FCFE, you are, in effect, laying claim to these cashflows, even if the bank does not pay them out as dividends.

Reconciling dividend and FCFE models
If a bank can be valued using both a dividend discount model and an FCFE model, should you get the same value using both approaches? No, and the differences between the two approaches can be summarized as follows:

- In the dividend discount model, you are assuming that what gets paid out as dividend is a good measure of what could have been paid out. Thus, you are assuming that the managers of financial services companies are sensible people who pay dividends only if they can and do not hold back cash. To the extent that firms do sometimes pay out more than they can afford to or hold back cash, the dividend discount model can overestimate or underestimate value.
- If you have negative FCFE, the FCFE model automatically incorporates the effect of the dilution that will arise (from the new stock issue) into the present value; the negative FCFE in the early years reduce the present value of the overall cash flows. In the dividend discount model, you have to explicitly adjust the number of shares for expected future dilution. If you do not do so, the dividend discount model will overestimate the value of equity in financial services firms with negative free cashflows to equity.

Excess return models
The third approach to valuing financial services firms is to use an excess return model. In this model, the value of a firm can be written as the sum of capital invested currently in the firm and the present value of excess returns that the firm expects to make in the future. In this section, we will consider how this model can be applied to valuing equity in a financial services firm.

Basic model
Given the difficulty associated with defining total capital in a financial services firm, it makes far more sense to focus on just equity when using an excess return model to value the firm. The value of equity in a firm can be written as the sum of the equity invested in a firm's current investments and the expected excess returns to equity investors from these and future investments.

Value of equity = equity capital invested currently + present value of expected excess returns to equity investors

The most interesting aspect of this model is its focus on excess returns. A firm that invests its equity and earns the fair market rate of return on these investments should see the market value of its equity converge on the equity capital currently invested in it. A firm that earns a below-market return on its equity investments will see its equity market value dip below the equity capital currently invested.
The other point that has to be emphasized is that this model considers expected future investments as well. Thus, it is up to the analyst using the model to forecast not only where the financial services firm will direct its future investments but also the returns it will make on those investments. To the extent that the firm will generate returns on equity on these new investments that exceed (are less than) the cost of equity, the additional investments (growth) will add to (reduce) the value of equity in the firm.

**Inputs to model**

There are two inputs needed to value equity in the excess return model. The first is a measure of equity capital currently invested in the firm. The second and more difficult input is the expected excess returns to equity investors from new equity investments in future periods.

The equity capital currently invested in a firm is usually measured as the book value of equity in the firm. While the book value of equity is an accounting measure and is affected by accounting decisions, it should be a much more reliable measure of equity invested in a financial services firm than in a manufacturing firm for two reasons. The first is that the assets of a financial services firm are often financial assets that are marked to market; the assets of manufacturing firms are real assets and deviations between book and market value are usually much larger. The second is that depreciation, which can skew book value for manufacturing firms, is often negligible at financial services firms. Notwithstanding this fact, the book value of equity can be affected by stock buybacks and extraordinary or one-time charges. The book value of equity for financial services firms that have one or both may understate the equity capital invested in the firm.

The excess returns, defined in equity terms, can be stated in terms of the return on equity and the cost of equity.

\[
\text{Excess equity return} = (\text{return on equity} - \text{cost of equity}) (\text{equity capital invested})
\]

Here again, we are assuming that the return on equity is a good measure of the economic return earned on equity investments. When analyzing a financial services firm, we can obtain the return on equity from the current and past periods, but the return on equity that is required is the return you will earn on future investments. This requires an analysis of the firm’s strengths and weaknesses, as well as the competition faced by the firm; banks that have strong competitive advantages will generate much higher returns on equity.

**Implications of model**

Linking the value of a financial services firm to its base capital and the excess returns it earns on that capital (and additions to it) provides a useful tool for assessing the market value of the firm, relative to its book value for equity. In particular, if markets expect a financial services firm to generate returns on equity that are less than the cost of equity on its existing assets, they will push the market value of equity below the book value of equity. If this firm then proceeds to invest in growth with the same characteristics, the discount on book value will get larger.

The model can be reversed to estimate imputed returns on equity from the price to book ratio. Thus, in a stable growth firm, the imputed ROE is:

\[
\text{PBV} = \frac{\text{Market value of equity}}{\text{Book value of equity}} = \frac{\text{ROE} - \text{expected growth rate}}{\text{Cost of equity} - \text{Expected growth rate}}
\]

Imputed ROE = PBV (cost of equity – expected growth rate) + expected growth rate

To illustrate, Citigroup was trading at a price to book ratio of 0.70 in November 2012. If we attach a cost of equity of 9% to the company and assume that it is in stable growth (growing at 2% a year), the imputed ROE for Citigroup can be computed as follows:

Imputed ROE for Citigroup = 0.70 (.09 - .02) + .02 = 6.9%

Thus, the market is anticipating that Citigroup will generate a return on equity 2.1% less than its cost of equity on both its existing assets and on new investments for the foreseeable future.

The trend line in price to book ratios across all banks has been negative since the 2008 banking crisis. In Figure 4, we look at the average price to book ratios and returns on equity for U.S. banks going back to 2002. Note the elevated price to book ratios for most of the last decade, as returns on equity stayed high. Note also the precipitous collapse in 2009, after returns on equity crashed in the aftermath of the 2008 crisis. While price to book ratios recovered...
in 2010, with the partial bounce back in return on equity, markets have become decidedly more negative in their outlook again in 2011, even as returns on equity continue to improve.

**Asset based valuation**

In asset-based valuation, we value the existing assets of a financial services firm net of debt and other outstanding claims and report the difference as the value of equity. For example, with a bank, this would require valuing the loan portfolio of the bank (which would comprise its assets) and subtracting outstanding debt to estimate the value of equity. For an insurance company, you would value the policies that the company has in force and subtract out the expected claims resulting from these policies and other debt outstanding, to estimate the value of the equity in the firm.

**Valuing assets**

How would you value the loan portfolio of a bank or the policies of an insurance company? One approach would be to estimate the price at which the loan portfolio can be sold to another financial services firm, and the better approach is to value it based upon the expected cash flows. Consider, for instance, a bank with a U.S.$1 billion loan portfolio with a weighted average maturity of 8 years, on which it earns interest income of U.S.$70 million. Furthermore, assume that the default risk on the loans is such that the fair market interest rate on the loans would be 6.50%; this fair market rate can be estimated by either getting the loan portfolio rated by a ratings agency or by measuring the potential for default risk in the portfolio. The value of the loans can be estimated.

\[
\text{Value of loans} = \text{U.S.}\$70 \text{ million (PV of annuity, 8 years, 6.5%) +} \\
\frac{\text{U.S.}\$1,000 \text{ million}}{1.065} = \text{U.S.}\$1,030 \text{ million}
\]

This loan portfolio has a fair market value that exceeds its book value because the bank is charging an interest rate that exceeds the market rate. The reverse would be true if the bank charged an interest rate that is lower than the market rate. To value the equity in this book, you would subtract out the deposits, debt, and other claims on the bank.

**Limitations of approach**

This approach has merit if you are valuing a mature bank or insurance company with little or no growth potential, but it has two significant limitations. First, it does not assign any value to expected future growth and the excess returns that flow from that growth. A bank, for instance, that consistently is able to lend at rates higher than justified by default risk should be able to harvest value from future loans as well. Second, it is difficult to apply when a financial services firm enters multiple businesses. A firm like Citigroup that operates in multiple businesses would prove to be difficult to value because the assets in each business — insurance, commercial banking, investment banking, portfolio management — would need to be valued separately, with different income streams and different discount rates.

**The drivers of value**

While we presented four different valuation models that we can use to value banks, the drivers of value are the same across these models: the risk in the equity earnings to the firm, the expected growth in these earnings over time, and the quality of the growth. In this section, we take a closer look at each of these determinants.

**Risk**

As with any other set of companies, the primary risk that we worry about when investing in a bank or insurance company is that its earnings will be volatile over time. As investors, we then have to parse this risk to evaluate how much of it we will still be exposed to, in our portfolios, as “diversified” investors. That is the risk we...
capture in the cost of equity in any discounted cashflow model. The basic proposition, then, is a simple one. A financial services firm that is more exposed to risk should be valued less than an otherwise similar financial services firm (in terms of growth level and quality).

**Risk measures**

In our earlier discussion of cost of equity, we suggested that the conventional models for measuring equity risk (using beta or betas) can be adapted to obtain costs of equity for financial services firms. We would argue against the use of a regression beta (with returns on a stock regressed against the market index) because of the noise in the estimates (standard errors) and the possibility that the firm has changed over the period of the regression, and the use of a sector average for the business. While the standard practice with nonfinancial services firms is to unlever these sector average betas and then relever them again, using the company’s debt to equity ratio, we would skip this step for two reasons. First, financial services firms tend to be much more homogeneous in terms of capital structure — they tend to have similar financial leverage primarily due to regulatory restrictions. Second, and this is a point made earlier, debt is difficult to measure for financial services firms. In practical terms, this will mean that we will use the average levered beta for comparable firms as the bottom-up beta for the firm being analyzed.

**Risk variation across businesses**

Valuations would be far simpler if financial services firms operated in a single business, with homogeneous risk. Over the last few decades, a combination of regulatory changes and securitization has made financial services firms much more complex. As a consequence, the large money center banks operate as financial supermarkets, deriving their revenues from many different businesses, with very different risk profiles. In practical terms, this business diversity shows up as different costs of equity for different businesses, with riskier businesses (such as proprietary trading) having much higher risk exposure (and cost of equity) than safer businesses. While you can still compute one beta for a complex financial services firm, reflecting the weights of its different businesses, changes in these weights will translate into changes in the composite cost of equity over time. In fact, a failure to do so will result in the overvaluation of banks that grow their riskier businesses at faster rates than their safer businesses.

<table>
<thead>
<tr>
<th>Category</th>
<th>U.S.</th>
<th>Europe</th>
<th>Emerging markets</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large money center banks</td>
<td>0.77</td>
<td>1.50</td>
<td>0.94</td>
<td>0.82</td>
</tr>
<tr>
<td>Small/regional banks</td>
<td>0.93</td>
<td>0.66</td>
<td>0.44</td>
<td>0.89</td>
</tr>
<tr>
<td>Thrifts</td>
<td>0.71</td>
<td>1.25</td>
<td>1.01</td>
<td>0.99</td>
</tr>
<tr>
<td>Brokerage houses</td>
<td>1.20</td>
<td>0.67</td>
<td>0.96</td>
<td>0.81</td>
</tr>
<tr>
<td>Investment banks</td>
<td>1.30</td>
<td>NA</td>
<td>NA</td>
<td>1.30</td>
</tr>
<tr>
<td>Life insurance</td>
<td>1.58</td>
<td>1.38</td>
<td>0.66</td>
<td>1.10</td>
</tr>
<tr>
<td>Property and casualty insurance</td>
<td>0.91</td>
<td>1.18</td>
<td>0.68</td>
<td>0.77</td>
</tr>
</tbody>
</table>

**Table 3: Betas for financial services businesses**

If we use sector betas and do not adjust for financial leverage, we are in effect using the same beta for every company in the sector. There can be significant regulatory differences across markets, and even within a market, across different classes of financial services firms. To reflect this, we would define the sector narrowly; thus, we would look at the average beta across large money center banks when valuing a large money center bank, and across small regional banks when valuing one of these. We would also argue that financial services firms that expand into riskier businesses — securitization, trading, and investment banking — should have different (and higher betas) for these segments, and that the beta for the company should be a weighted average. Table 3 summarizes the betas for different groups of financial services companies, categorized by region, in January 2012.

**Risk and regulation**

There is one final point that bears emphasizing here. The average betas that we get across financial services firms reflect the regulatory constraints that they operated under during that period. When significant changes are expected to regulation, we should consider the potential impact on costs of equity for firms affected by the regulation. For instance, the crisis of 2008 caused banking regulations to be tightened globally and may very well have pushed up the betas for all banks, at least for the foreseeable future.

The research on the relationship between risk and regulation is still in its infancy and we do not have the answers to several key questions. Is regulatory uncertainty firm-specific or diversifiable...
risk? Should we be incorporating it into betas and costs of equity or should we assume that investors can eliminate it in their portfolios? One way to answer these questions is to look at how the banks are priced across different regulatory regimes, comparing both across time and across countries. If, in fact, regulatory risk is a driver of the cost of equity, we should expect to see banks trade at much lower prices, relative to earnings, dividends, and book value of equity, during periods of high regulatory risk.

**Expected growth and its quality**

In developed markets, we have tended to think of financial services firms as mostly mature businesses, with little potential for explosive growth. That may be appropriate, given how saturated these markets are with financial product offerings, but even in developed markets, some banks find potential for high growth, either by exploiting new markets or competing for market share in existing ones. As a general proposition, it is far more likely that you will find high growth at smaller financial firms in mature markets.

In emerging markets, where access to, and use of, financial products (banking services, mutual funds) is in its infancy, it is still possible to find high growth at financial services firms, even if they are large. This is especially true in the many emerging markets that have been reluctant to open their financial services businesses to foreign competition. Ultimately, though, the common theme across all of the valuation models in the last section is that it is not the growth per se that creates value but whether that growth is accompanied by excess returns, i.e., the difference between the return on equity and the cost of equity. Financial services firms that generate returns on equity that exceed the cost of equity, we argued, will create value from growth. To value a financial services firm, therefore, it behooves us to get better assessments of the return on equity, both on existing assets and new investments.

To get a measure of how much variation there is in return on equity across banks, we looked at the distribution of return on equity for U.S. banks in November 2012 (Figure 5).

Of the 106 banks in the sample, about 70% had returns on equity that were less than the median cost of equity of 9.20% in November 2012 (based upon a risk free rate of 2%, an equity risk premium of 6%, and a beta of 1.20). It is therefore not surprising that so many banks trade at below their book value of equity.

**Relative valuation**

There are a series of multiples that are used to value firms, ranging from earnings multiples to book value multiples to revenue multiples. In this section, we consider how relative valuation can be used for financial services firms.

**Choices in multiples**

Firm value multiples such as enterprise value (EV) to EBITDA or EV to EBIT cannot be easily adapted to value financial services firms because neither value nor operating income can be easily estimated for banks or insurance companies. In keeping with our emphasis on equity valuation for financial services firms, the multiples that we will work with to analyze financial services firms are equity multiples. The three most widely used equity multiples are price earnings ratios, price to book ratios, and price to sales ratios. Since sales are not really measurable for financial services firms, price to sales ratios cannot be estimated or used for these firms.

**Price earnings ratios**

The price earnings ratio for a bank or insurance companies is measured much the same as it is for any other firm.
Price earnings ratio = \(\frac{\text{Price per share}}{\text{Earnings per share}}\)

The price earnings ratio is a function of three variables – the expected growth rate in earnings, the payout ratio, and the cost of equity. As with other firms, the price earnings ratio should be higher for financial services firms with higher expected growth rates in earnings, higher payout ratios, and lower costs of equity.

An issue that is specific to financial services firms is the use of provisions for expected expenses. For instance, banks routinely set aside provisions for bad loans. These provisions reduce the reported income and affect the reported price earnings ratio. Consequently, banks that are more conservative about categorizing bad loans will report lower earnings and have higher price earnings ratios, whereas banks that are less conservative will report higher earnings and lower price earnings ratios.

Another consideration in the use of earnings multiples is the diversification of financial services firms into multiple businesses. The multiple that an investor is willing to pay for a dollar in earnings from commercial lending should be very different than the multiple that the same investor is willing to pay for a dollar in earnings from trading. When a firm is in multiple businesses with different risk, growth, and return characteristics, it is very difficult to find truly comparable firms and to compare the multiples of earnings paid across firms. In such a case, it makes far more sense to break the firm’s earnings down by business and assess the value of each business separately.

Price to book ratios

The price to book ratio for a financial services firm is the ratio of the price per share to the book value of equity per share.

Price to book ratio = \(\frac{\text{Price per share}}{\text{Book value of equity per share}}\)

If anything, the strength of the relationship between price to book ratios and returns on equity should be stronger for financial services firms than for other firms, because the book value of equity is much more likely to track the market value of equity invested in existing assets. Similarly, the return on equity is less likely to be affected by accounting decisions. The strength of the relationship between price to book ratios and returns on equity can be seen when we plot the two on a scatter plot for U.S. commercial banks in the U.S. in November 2012 (Figure 6).

Figure 6: Price to book ratios and returns on equity: U.S. banks in November 2012

*Regression line, with 90% confidence range on estimate*

If anything, the strength of the relationship between price to book ratios and returns on equity should be stronger for financial services firms than for other firms, because the book value of equity is much more likely to track the market value of equity invested in existing assets. Similarly, the return on equity is less likely to be affected by accounting decisions. The strength of the relationship between price to book ratios and returns on equity can be seen when we plot the two on a scatter plot for U.S. commercial banks in the U.S. in November 2012 (Figure 6).

Note the strong link between price to book ratios and returns on equity, with the banks that earn higher returns on equity trading at much higher price to book ratios.

In fact, a key test of this relationship is to see if it holds even in the midst of a crisis. To test that proposition, we went back to February 2009 and graphed price to book ratios against returns on equity for U.S. banks with market capitalizations that exceeded a billion in Figure 7.

In the midst of the biggest crisis in banking since the Great Depression, and in an environment where most analysts have come to the conclusion that investors are in crisis mode and
that equity values in banks reflect the panic and irrationality, it is astounding how close the link is between price to book ratios for banks in February 2009 and the returns on equity, based upon trailing 12-month earnings. Banks that have high price to book ratios tend to have high returns on equity (top right corner of Figure 7), while those that have low returns on equity trade at low price to book ratios (bottom left hand corner of Figure 7). The correlation between price to book ratios and returns on equity is in excess of 0.70. Put another way, there seems to be a fundamental order even in the midst of chaos.

While emphasizing the relationship between price to book ratios and returns on equity, we should not ignore the other fundamentals. For instance, banks vary in terms of risk, and we would expect that, for any given return on equity, riskier banks should have lower price to book ratios. Similarly, banks with much greater potential for growth should have much higher price to book ratios, for any given level of the other fundamentals. Since the banking crisis, one factor that should make a difference is the exposure that different banks have to toxic securities — mortgage backed bonds and collateralized debt obligations (CDOs) — on their balance sheets.

Figure 7: Price to book versus ROE for U.S. Banks: February 2009

Conclusion
The basic principles of valuation apply just as much for financial services firms as they do for other firms. There are, however, a few aspects relating to financial services firms that can affect how they are valued. The first is that debt, for a financial services firm, is difficult to define and measure, making it difficult to estimate firm value or costs of capital. Consequently, it is far easier to value the equity directly in a financial services firm, by discounting cash flows to equity at the cost of equity. The second is that capital expenditures and working capital, which are required inputs to estimating cashflows, are often not observable at financial services firms. In fact, much of the reinvestment that occurs at these firms is categorized under operating expenses. To estimate cashflows to equity, therefore, we either have to use dividends (and assume that what is not paid out as dividend is the reinvestment) or modify our definition of reinvestment.

Even if we choose to use multiples, we run into many of the same issues. The difficulties associated with defining debt make equity multiples such as price earnings or price to book ratios better suited for comparing financial services firms than value multiples. In making these comparisons, we have to control for differences in fundamentals — risk, growth, cashflows, and loan quality — that affect value.

Finally, regulatory considerations and constraints overlay financial services firms’ valuations. In some cases, regulatory restrictions on competition allow financial services firms to earn excess returns and increase value. In other cases, the same regulatory authorities may restrict the potential excess returns that a firm may be able to make, by preventing the firm from entering a business.
Getting risk governance right

Thomas F. Huertas
Partner, Ernst & Young LLP

Abstract
This paper sets out the criteria that boards and supervisors should use to determine whether banks are governing risk correctly. First, boards have to set an overall risk target consistent with the overall return target. Second, the bank's business model has to set a risk capacity and risk appetite that can enable the bank to meet its target for risk and return. Finally, the bank has to ensure that its three lines of defense (management, risk management and compliance, and internal audit) function well, both singly and in combination.

JEL Classification Codes: G21, G28, G30, G32, G38
Introduction
Few would argue with the notion that corporate governance within a majority of financial institutions requires improvement, and that shortcomings in this area played an important role in the recent crisis [Walker (2009)]. In response, policy-makers have implemented extensive reforms to corporate governance within banks and other financial institutions, particularly with respect to risk governance, and supervisors are stepping up their reviews of how effectively banks govern risk [Bank of England Prudential Regulation Authority (2011)].

Shareholders also have an interest in strong risk governance. For banks, risk governance is critical, for banking is the business of taking — and being adequately compensated for — risk. Boards, therefore, need to focus on three questions:

1. Does the bank have the correct overall risk target?
2. Does the bank have the right business model?
3. Is the bank taking risk correctly?

Good risk governance would enable the bank to answer “yes” to each of these questions.

What should be the overall risk target?
The overall risk target must be set in conjunction with the overall return target for the institution. The two targets must be consistent with one another.

Prior to the recent crisis, they were not. Many banks were too ambitious. They targeted Return on Equity (RoE) at 20% or more while stating that they would simultaneously maintain a credit rating of AA or better. Some banks achieved this at the peak of the cycle, but when the crash came, profits plummeted and ratings collapsed. In many cases, the bank in question had to seek public assistance. The combination of high RoE and low risk proved unsustainable (Figure 1).

Supervisors seek to eliminate banks’ ability to pursue high-risk strategies

As a consequence of the recent crisis, regulators (supervisors and central banks) are now focusing on risk significantly more aggressively than they did in the past and the regulatory reform agenda [FSB (2012)] has two major objectives:

1. To reduce the risk that banks will fail. To this end, the Basel III accord [BCBS (2011)] significantly increases the amount and quality of capital that banks must hold. It also institutes, for the first time, a global liquidity standard that will require banks to measure and manage their liquidity risk and to hold a buffer of liquid assets to allow them to meet unexpectedly large outflows of funds.
2. To make banks “safe to fail.” The objective is to enable the authorities to resolve a bank that fails to meet threshold conditions promptly at no cost to the taxpayer and without significant disruption to the financial markets or to the economy at large. To this end, the Financial Stability Board (2011) has developed a list of key attributes that bank resolution regimes should have and these are now being implemented in major jurisdictions around the world.

Together, these regulatory initiatives will have a significant impact on banks, especially the systemically important financial institutions that are at the heart of the global financial system. Banks will become more similar to nonfinancial corporations. They will experience higher funding costs as their business risk
and their leverage increases. The reform of resolution presages an end to “too big to fail.” Investors, not taxpayers, will have to bear the cost of bank resolution. This means that creditors will be exposed to more risk (see box on page 82 for explanation) and the risk premium that banks will pay to borrow on an unsecured basis will, ceteris paribus, rise as well.

Supervisors, however, have no intention of leaving all other things equal. Their strategy is to reduce the risk that banks can take. In finance terms, supervisors are seeking first of all to eliminate the ratings pickup that banks derive from implicit government support so that there is no difference between a bank’s stand-alone rating and its overall rating (Figure 2). This element of the strategy is already showing some signs of success. In response to reform in resolution regimes, rating agencies have begun to reduce the credit that they give to banks for the support that governments might supply to banks if they were to become distressed. In some cases this revision has resulted in a downgrade in a bank’s overall rating even though its stand-alone rating remained unchanged.

Second, supervisors are aiming to ensure that banks maintain a stand-alone rating that is comfortably investment grade at all times (even at the trough of the cycle) so that the bank will remain a considerable distance from resolution (Figure 2). The principal means to this end is higher capital requirements. Under Basel III the minimum capital ratio will increase substantially. If a bank wishes to have the flexibility to pay dividends or make distributions to shareholders, it must maintain capital greater than or equal to 7% of its risk-weighted assets. For systemically important banks, the ratio must be higher still. Such banks will incur a surcharge of up to 2.5% of risk-weighted assets. And, banks must meet these higher requirements with higher-quality capital, namely tangible common equity.

Supervisors are also reinforcing this tougher capital regime in three ways. Stress tests are being employed to ensure that banks meet the capital requirements, not only on a current basis, but also prospectively even under very adverse economic conditions. In a number of jurisdictions, notably the U.K. and Switzerland, banks are being required to maintain substantially higher total capital ratios (primary loss absorbing capacity) than the minimum required under Basel III. Leverage ratios are being introduced as a backstop to risk-weighted capital requirements. Finally, regulators are introducing a global liquidity standard. This will force banks to measure and manage their liquidity risk as well as keep a liquid asset buffer. As noted above, the overall objective is to reduce the risk that the bank will fail. In finance terms, supervisors are seeking to eliminate banks’ ability to adopt a high-risk strategy.

**Shareholders seek to assure banks earn their cost of equity capital**

Over time, the bank has to earn its cost of equity capital. Otherwise, equity investors will divert their funds to companies that do so, and the bank will find it costly or difficult to raise new equity capital.

But the cost of equity capital is not constant. It depends on the risk of that capital (Figure 1). This depends on the business risk that the firm runs and the leverage that the firm employs. Supervision aims to limit the former factor and regulatory reform will lower the latter. This should reduce the risk of the bank’s equity and the cost of (or required rate of return on) that equity.

However, this new lower risk or lower return environment need not reduce the return to shareholders (Figure 3). Return to equity investors depends primarily on whether or not the bank earns a return in excess of its cost of equity, not on the level of the cost of equity. Low return-low risk equities can perform just as well as, or
butter than, high return-high risk equities. In fact, total return to shareholders from holding utility stocks (a proxy for low return-low risk equities) has compared favorably with the total return realized by shareholders in financial institutions [Huertas (2009)].

**Does the bank have the right business model?**

Today, few banks have an investment-grade stand-alone credit rating and even fewer are covering their cost of equity capital. The question is how they can expect to achieve both, simultaneously and consistently, over time. The answer lies in the bank’s business model. This has to spell out how the bank earns its money. What risks is it taking and what is it being paid to take those risks? What capital does it require in order to support those risks, both in normal times and under stress?

For example, take a simple borrow-short/lend-long business model. The bank should be able to build up to its target rate of return on equity by identifying the return that it expects to get from taking credit risk, interest-rate risk, and liquidity risk as well as the contribution that leverage will make (Figure 4). It should consider whether the credit risk premium is adequate to cover the losses that may arise as a result of operational risk, conduct risk, or documentation risk. The bank should also be able to calculate the risk premium that it will have to pay on its own funding. Together, these factors should yield the target rate of return on the bank’s equity.

At a minimum, the board will wish to periodically conduct what might be called a “business-model checkup.” This can be the first step toward improving the bank’s performance. Such a checkup would include a review of whether: the bank is ignoring one or more risks; is mis-pricing one or more risks; is spending too much to acquire or administer the risks that it does take; has sufficient risk capacity to survive under stress; and has a risk appetite in line with its risk capacity.

**Does the business model include all the risks the bank takes?**

The business model must include all the risks that the bank takes. Had such an analysis been conducted prior to the crisis, banks might have seen that they were ignoring or underestimating liquidity risk as well as conduct and operational risks. Particularly high returns may in fact signal that the bank is assuming a risk without knowing it is doing so or without appropriate controls. Take, for example, the very high returns that banks were recording on their holdings of mortgage-backed securities in the years immediately prior to the crisis. After the crisis, it became obvious that a good deal of the pre-crisis return stemmed from the assumption of market liquidity risk — a risk that banks’ models by and large failed to incorporate. Another example is the very high returns that banks in the U.K. earned from selling payment protection insurance. Subsequently, it
became apparent that a good deal of these returns stemmed from mis-selling. A third example is unauthorized trading. In a number of instances the trader in question reported significant profits on what should have been low-risk positions. This should have triggered alarm bells. Indeed, examining the reasons for superior returns may be equally or more important as diagnosing why a bank is losing money. Rather than simply ascribing superior returns to management proficiency, boards should assure themselves that such returns are not based on hidden risks. That can help the bank avoid future losses.

**Does the business model price risk correctly?**

The business model should ensure that the bank prices risk correctly. It should induce the bank's line businesses to demand externally a full price from the customer for the risks that the bank assumes and to pay internally a full price for the capital and liquidity that the line businesses require.

Undercharging customers for the risk of credit is particularly dangerous. Banks that follow such a policy will gain a disproportionate share of such credit and fail to earn enough income over time to provision adequately for the impairments that will arise. Should the bank have to sell such underpriced credit to third parties, it will be able to do so (even if the loan is performing) only at a discount.

Poor funds transfer pricing also undermines good business performance. Prior to the crisis, banks commonly failed to credit deposit-generating businesses with the full market value of the funds that they supplied to the bank's central treasury. Many banks further compounded this error by allowing business divisions that required funding to borrow from the central treasury pool at a rate that did not fully reflect the risk of the division that borrowed (and in some cases, at a rate that was below the bank's own overall marginal cost of funds). Such arrangements artificially inflated returns in the divisions that were permitted to borrow on preferential terms, and may have induced banks to allocate assets and capital to such divisions.

Correcting such mis-pricing is important. Indeed, in many respects the proposals to separate trading businesses from retail and commercial banking and to place a ring fence around either the trading business (Liikanen Group (2012)) or the retail and commercial bank (HMT (2012)) can be seen as efforts to correct the funds transfer mis-pricing that had occurred in many banks prior to the crisis. But such regulation is no substitute for an accurate funds transfer pricing mechanism, and banks should have systems in place to ensure that internal transfer prices are in accord with market rates.

**Is the business model efficient?**

The business model also has to ensure that the bank operates efficiently. Although banks have made vast strides in this area in recent years, two areas for possible improvement stand out: asset/capital efficiency and compensation.

- **Asset/capital efficiency** - with respect to assets and capital, banks have to ask the same types of questions as nonfinancial firms: are they operating in lines of business in which they have a competitive advantage, and are they maximizing the turnover (revenue) that they get from holding each asset?

  Restricting activity to lines of business in which the bank has a competitive advantage should raise margins and, therefore, increase the bank's rate of return. As the bank exits businesses in which it has no competitive advantage, it frees up capital, allowing such capital to be reallocated to more efficient uses.

  But allocation is only the first step to capital efficiency. Even for businesses in which the bank has a competitive advantage, it is important for the bank to employ assets and capital efficiently. Just as retailers keep a keen eye on their turnover to inventory ratio, so should banks review exactly which assets they need to hold for how long in order to generate revenue and profit.

  What distinguishes banks is their ability to respond quickly to customers' demands for new credit — be that intra-day credit to facilitate payments, short-term credit to finance receivables or consumer purchases, or longer-term credit or bonds to finance investment in factories or houses. In order to be able to continue to offer consumers credit card limits or overdrafts, to offer businesses revolving credit facilities, and to respond quickly to requests for long-term credit (be that in the form of loans or bonds), banks must have sufficient capacity available, both vis-à-vis the individual borrower (so that the new loan will not break counterparty limits) and the market (so that the new credit in aggregate will not cause the bank to breach capital or liquidity limits).
Once a loan has been made, or a bond underwritten, banks are not necessarily the most efficient holders of assets. Investment funds exempt from corporate tax may be much better placed to hold assets.

This suggests that banks may wish to revisit the “originate to distribute” strategy. In concept, the strategy makes sense – for it focused on what should be banks’ comparative advantage (origination) and minimized the amount of assets that the bank needed to hold relative to turnover (revenue) that the bank could achieve. But to make such a strategy work in the future, banks will have to develop the distribution leg3 and put in place processes to better manage the conflicts of interest that such a strategy may imply.4

- **Compensation** – compensation is the largest component of a bank’s cost after interest expense and any business model must ensure that the bank’s compensation system sets the right incentives, i.e., that it is consistent with effective risk management and with shareholders receiving an adequate return on capital employed. As is well known, prior to the crisis, compensation structures, particularly at major trading banks, had departed from these principles. Many banks accrued bonuses as a percentage of revenues, allowed management a significant say in policies with respect to revenue recognition, and paid the total bonus in cash annually on the basis of reported earnings for that year. This practice contributed to excessive risk-taking (IIF (2009)).

Regulation has forced banks to align compensation structures with effective risk management. Bonuses must be adjusted for the risk incurred to generate the profit that justified the bonus and bonuses can no longer be paid fully in cash. At least one-half must be deferred, and the deferred amount must be subject to claw back and paid in a form (i.e., equity) that stands to lose value over the deferral period if risks to the firm crystallize.

But regulation does not ensure that compensation systems are consistent with ensuring that capital receives an adequate rate of return. Bonus is still accrued “above the line” rather than (as is the case in private equity or hedge fund contracts, with investors) as a percentage of the profit realized after equity capital has been paid a minimum rate of return. Shifting some portion of bonus to “below the line” would further align compensation to effective risk management (in that it would force management to recognize the cost of equity capital) and it would help ensure that equity receives a minimum rate of return (Huertas (2011)).

**Will the business model survive under stress: what is the bank’s risk capacity?**

A bank’s business model must also ensure that the bank can survive under stress. The first step toward an answer is to estimate the bank’s risk capacity – roughly the amount of money (in absolute terms) that the bank could afford to lose over the next year without reaching the point of non-viability.

Regulation is seeking to ensure that banks will have adequate risk capacity. Increasingly, regulation is requiring banks to hold up front, sufficient capital to withstand a fairly adverse macroeconomic scenario, so that the bank will still meet minimum capital requirements after taking into account the losses that it might incur in such an economic environment. Banks that do not meet such criteria are forced to file plans that will enable them to do so within a relatively short time frame. Such plans may include raising capital or changing their business model so that they can reduce risk, reduce capital usage, and generate significant retained earnings.

Risk capacity can be quite sensitive to changes in the business model, and banks will want to employ tools, such as strategic balance sheet forecasting, that allow them to estimate quickly how capital and liquidity ratios would evolve under various strategies and economic scenarios. This will allow the bank to maximize its risk capacity as well as to take precautions that will enable it to survive should stress materialize.

**Will the business model make the bank prone to stress: what is the bank’s risk appetite?**

Risk capacity sets out how much risk the bank could take. Risk appetite indicates how much of that capacity the bank wishes to utilize. Boards need to ensure that the bank’s risk appetite is in
line with the bank’s risk capacity while recognizing that the bank has to be able to take a minimum level of risk if it is to compete successfully in various lines of business.

In the long run, the bank’s risk appetite cannot exceed its risk capacity. Either the bank has to raise its risk capacity or cut its risk appetite. Conversely, as risk capacity contracts, so should risk appetite, unless the bank can take countervailing measures quickly to restore capacity.

The prudent bank will, therefore, leave a cushion between risk capacity and risk appetite. It will wish to hold some risk capacity in reserve, particularly as it is difficult to forecast how well funding will hold up once the bank suffers a loss to capital. This reserve should be higher, the fewer or weaker the bank’s recovery options are. Banks should anticipate that supervisors will pay increasing attention to this space. They have already demanded recovery plans from banks and they are stepping up their review of risk appetite statements.

Setting the overall risk appetite is, however, only the first step. The aggregate appetite (and capacity) has to be allocated to the bank’s businesses and this needs to be reconciled with the bank’s business model. For example, if a bank is to compete in offering credit cards to consumers, it will wish to hold some risk capacity in reserve, particularly as it is difficult to forecast how well funding will hold up once the bank suffers a loss to capital. This reserve should be higher, the fewer or weaker the bank’s recovery options are. Banks should anticipate that supervisors will pay increasing attention to this space. They have already demanded recovery plans from banks and they are stepping up their review of risk appetite statements.

The allocation mechanism should also have an effective brake. Boards should ask whether the bank can continue to control its risk appetite if others in the industry lose their risk discipline. At what point does pricing become so unsatisfactory that the bank will stop writing new business, and does the bank’s business model give it the option to take a temporary pause until market pricing again reflects the risk to the bank?

**Is the bank taking risk correctly?**

Finally, boards will want to know how management assures itself that the bank is taking risk correctly. In effect, boards must ask how well the bank’s “three lines of defense” (management, risk management and compliance, and internal audit) actually work. To some extent, boards may rely on the opinion of the bank’s external auditor, but boards will also want to satisfy themselves more directly that the three lines of defense are working as they should. The first line of defense is the management itself. How does it regard risk? Does management view risk as the core of what a bank has to manage or as an appendix to the “real” business of selling products to clients? In financial services, taking risk is a necessary input to generating returns, and the board will want to know that management places the proper emphasis on managing risk.

Evaluating the second and third lines of defense will predominantly fall to the risk and audit committees of the board. The risk committee of the board will naturally focus on the second line of defense — risk management and compliance. To facilitate this review, the bank’s chief risk officer should have a reporting line to the chair of the risk committee (similar to the relationship that the bank’s chief auditor has with the chair of the audit committee).

With respect to compliance, the board will want to know that the bank is in compliance with all the relevant rules and regulations in each of the jurisdictions in which the bank operates. How does compliance convince management that this is the case?

The board will also want to think through issues that could become compliance concerns in the future. Complying with the letter of a regulation may not provide an adequate defense against supervisory or judicial action if the results flaunt what society perceives to be fair. Nor will reliance on market practice necessarily be a good defense against a violation of the letter of a regulation, if a government or court reviews that practice in the cold light of day at some point in the future.

The audit committee of the board will focus on the third line of defense. Central to the effort of the board in this area is a dialogue with, and assessment of, internal audit (IA). Is IA covering all the major risks the bank takes? Is IA discovering the right issues? Is management recommending the right remedies and implementing them promptly? Does the director of internal audit have a direct line to the audit committee of the board, and how do they use it?

---

5 The Risk and Audit Committees should work in a coordinated fashion, possibly via cross-membership on the two committees. At some institutions, for example, the Chair of the Risk Committee is also a member of the Audit Committee, and the Chair of the Audit Committee is a member of the Risk Committee. This helps ensure that each committee is aware of the agenda of the other.
The board should be asking the big questions

In summary, risk governance is central to the business of banking, for banking is the business of taking — and being adequately compensated for taking — risk in various forms, including, without limitation, credit, liquidity, interest rate, market, and operational risks. Both shareholders and supervisors are looking to boards of directors to ensure that:

- The bank targets a return above its cost of capital while maintaining a risk profile that ensures that it can retain an investment grade stand-alone rating.
- The bank constructs a business model that will generate the target combination of risk and return.
- The bank assumes risk in a controlled fashion so that its three lines of defense — line management, risk management and compliance, and internal audit — each function correctly and reinforce one another.

Bringing these elements together would create a strong risk culture at the bank and help lay the foundations for its lasting success.

Resolution determines risk to investors

The resolution regime has a significant impact on the loss that investors expect when they extend unsecured credit to a bank and, therefore, on the risk premium that a bank will have to pay when it borrows.

<table>
<thead>
<tr>
<th></th>
<th>Too big to fail</th>
<th>Safe to fail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability that bank will fail to meet threshold conditions</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>Probability of rescue</td>
<td>95%</td>
<td>5%</td>
</tr>
<tr>
<td>Probability of resolution</td>
<td>5%</td>
<td>95%</td>
</tr>
<tr>
<td>Loss given resolution</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>Expected loss</td>
<td>25bp</td>
<td>475bp</td>
</tr>
</tbody>
</table>

A simple example illustrates why. Take an institution that is too big to fail. The probability that it will fail to meet threshold conditions and that intervention will be required is 20%. If the market judges that the bank is too big to fail, the market is effectively stating there is a very high probability (say 95%) that the government will rescue the bank (so that there is no loss to investors). Correspondingly, the market is stating that there is a low probability (say 5%) that the authorities will place the bank into resolution. If the bank does go into resolution, the expected loss to investors is 25% of their exposure at the point of resolution. Accordingly, the expected loss to investors in the too big to fail bank is 25 basis points (bp).

Now make the same bank safe to fail. Assume there is no change in the probability that the bank will fail to meet threshold conditions (this remains at 20%). Assume as well that the loss the investor suffers if the bank does go into resolution remains at 25%. The only change refers to the probability that the bank will be rescued. If the bank is safe to fail this falls to 5% and the probability that the bank will be put into resolution rises to 95%. As a result, the expected loss to investors in the safe to fail bank is 475bp.
References


Part 2: Insurance and Asset Management

Implied cost of equity capital in the U.S. insurance industry

The Dodd-Frank Act, Solvency II, and U.S. insurance regulation

What drives the development of the insurance sector? An empirical analysis based on a panel of developed and developing countries

The customers have spoken, are the insurance companies listening? A survey of the global insurance market

Do firm characteristics influence mutual fund performance? An empirical study for European mutual funds

Fund transfer pricing for deposits and loans, foundation and advanced

The cultural revolution in risk management
Implied cost of equity capital in the U.S. insurance industry

Doron Nissim
Ernst & Young Professor of Accounting & Finance, Columbia Business School

Abstract
This study derives and evaluates estimates of the equity risk premium inferred from the stock prices and analysts’ earnings forecasts of U.S. insurance companies. During most of the sample period, April 1983 through September 2012, the quarterly median implied equity risk premium (IERP) of U.S. insurers was relatively stable, fluctuating mildly around an average value of 5.5%. However, during the financial crisis of 2007-2009, the median IERP reached unprecedented levels, exceeding 15% in the first quarter of 2009. Following the financial crisis, the IERP declined substantially but it remained at historically high levels, exceeding 9% on average. In spite of significant differences in operations and financial profile, the median IERP of Life and Health insurers was similar to that of Property and Casualty insurers during most of the sample period. However, during the financial crisis the median IERP of Life and Health insurers was substantially larger than that of Property and Casualty insurers, consistent with the higher sensitivity of Life and Health insurers to fluctuations in financial markets. The differences in the IERP across the insurance sub-industries remained substantial after the crisis, indicating a structural change in the pricing of Life and Health insurers. Consistent with investors demanding relatively high rates of return in periods of poor economic performance or high uncertainty, the IERP is positively related to the credit spread, term spread, and inflation, and negatively related to the 10-year Treasury yield. The relations with firm-specific risk factors are similarly consistent with expectations: the IERP is positively related to market beta, and negatively related to size and the equity-to-assets ratio. These risk-factor sensitivities are generally higher for Life and Health insurers as well as during the financial crisis. Finally, consistent with the strong correlations between the IERP and the macro and firm-specific risk factors, the IERP performs well in predicting subsequent excess stock returns. One implication of the results is that the current trend in accounting regulation to eliminate accounting differences across insurance operations may not be desirable.

1 The author gratefully acknowledges the helpful comments and suggestions made by Trevor Harris, Urooj Khan, Stephen Penman, Eddie Riedl, Stephen Ryan, Harold Schroeder, Leslie Seidman, and seminar participants at the CARE-CEASA Conference (Accounting for Uncertainty and Risk), the Financial Accounting Standards Board, Columbia Business School (accounting and finance workshops), Harvard University, and the University of Washington. This research was supported by the Chazen Institute of International Business at Columbia Business School and the Center for Excellence in Accounting and Security Analysis (CEASA).
Introduction

When pricing equity securities, investors discount expected flows (e.g., dividend, free cash flow, residual income) using required rates of return commensurate with the riskiness of those flows. Consequently, given price and estimates of expected flows to equity holders, one can invert an equity valuation model to obtain an estimate of the average required rate of return used by investors in pricing the stock. From the company’s perspective, this estimate reflects the cost of equity capital and is accordingly referred to as the implied cost of equity capital (ICEC). From investors’ perspective, the ICEC is an estimate of the expected return on their investment. In this study I derive and evaluate estimates of the implied cost of equity capital for U.S. insurance companies.2

The ICEC is useful in various settings. Analysts, investors, and other stakeholders may use it to estimate expected returns, to price risk factors, or to calculate intrinsic equity value. For example, if an analyst perceives his/her expectations of future earnings to be different from the “consensus” earnings forecasts, he/she may estimate the ICEC using the stock price and the consensus forecasts by inverting an equity valuation model that discounts earnings forecasts, and then estimate intrinsic equity value by discounting his/her forecasts of future earnings using the ICEC. Creditors and regulators may use the ICEC as a market-based proxy for the riskiness of the company. Academics and practitioners may correlate the ICEC with firm characteristics to quantify the risk pricing of those attributes. The ICEC may also be used to estimate the expected rate of return on an equity investment, or to derive ex-ante estimates of the market risk premium (Claus and Thomas (2001), Fama and French (2002)).

While ICEC estimates are relevant in many settings and for essentially all companies, studying their properties for insurance companies is particularly important, given the fundamental role of risk in that industry. The primary purpose of insurance is the spreading of risks. Insurance is valuable because the risks associated with different policies are not perfectly correlated, and so the total risk of a portfolio of policies is smaller than the sum of the policies’ risks. Thus, insurance functions as a mechanism to diversify insurable risks, similar to the role of mutual funds in diversifying investment risks. In fact, because insurers accumulate substantial funds in conducting their business, they also diversify investment risks for their stakeholders by investing in diversified portfolios. Yet, insurers’ ability to reduce portfolio risks through diversification is limited. Some risks are not fully diversifiable (e.g., catastrophes, longevity/mortality, market risks) and, for various reasons (e.g., size, cost, management skills, line or geographic concentration, speculation), insurers may retain some diversifiable risks.3 To the extent that such residual risks are priced by investors, they should be reflected in insurers’ ICEC. Thus, unlike other industries where operating risks are primarily industry-specific, insurers’ ICEC may reflect the pricing of all insurable risks.4

Focusing on insurers’ ICEC estimates is relevant for two additional reasons. First, the insurance industry — primarily its life segment — experienced unprecedented volatility during the financial crisis of 2007-2009, and thus studying changes in insurers’ ICEC during that period may yield interesting insights. Second, for reasons discussed later, insurers’ ICEC estimates are likely to be more precise than those of most other companies and may therefore facilitate a better understanding of the time-series and cross-sectional determinants of equity risk premiums in general.

The precision of an ICEC estimate depends on the accuracy of the valuation model that implies that estimate. While equity valuation models that discount dividends, free cash flows, or residual earnings are well-grounded in theory, their empirical counterparts are often very “noisy.” In discounted dividends or free cashflows valuations, a large portion of equity value is captured by a “terminal value” calculation, which measures the

---

2 Over the last decade, the implied cost of capital has been the focus of, or a primary variable in, many studies, including Botosan (1997), Botosan and Plumlee (2002, 2005), Claus and Thomas (2001), Dholakia et al. (2005), Easton and Monahan (2005), Easton and Sommers (2007), Easton et al. (2002), Fama and French (2002), Francis et al. (2005), Gebhardt et al. (2001), Gode and Mohanram (2003), Hall and Leuz (2006), and Hribar and Jenkins (2004). As discussed later, an important difference between the current study and prior work is the focus on an industry where the model used to reverse-engineer the cost of equity capital is likely to perform relatively well.

3 Beside diversification, insurers may mitigate some risks by engaging in asset-liability management, reinsuring some exposures, using capital market solutions (for example, catastrophe bonds, contingent capital), or employing other techniques. However, these activities are also limited due to the same factors that constrain diversification, including size, cost, and management skills.

4 Insurance companies are unique in another way that affects risk. In most industries, the cost of the product or service provided is known at the time of sale or soon after the sale. In contrast, for insurance companies — primarily property and casualty insurers that specialize in long-tail lines such as medical malpractice — the ultimate cost of insurance is unknown until long after the sale. This characteristic is a main source of uncertainty and risk.
present value of all cashflows subsequent to the explicit forecast period. The terminal value is estimated using either a constant growth formula, an empirical multiple applied to a forecasted fundamental, or some other highly stylized calculation that is likely to contain considerable measurement error. Under the residual income model, the equivalent of the terminal value calculation captures the value of abnormal (residual) steady-state profitability and is therefore relatively small. Still, for companies whose book value fails to capture important assets, primarily internally developed intangible assets, steady-state abnormal profitability may be substantial.

For insurance companies, the residual income model and the corresponding ICEC estimate are likely to perform relatively well, because book value and explicit forecasts of earnings capture most of the modeled value, and “steady state” residual income (and hence the “terminal value”) are relatively small. The insurance industry is highly competitive and thus any abnormal earnings are likely to fade away within a relatively short horizon. In addition, although the reporting of insurance contracts involves some distortions, insurers’ financial statements are, overall, less conservatively biased than other industries. This is due to the relatively small magnitude of economic intangibles and to the financial nature of most assets and liabilities, which are generally reported at amounts close to fair values (for example, available for sale securities, some insurance reserves). Indeed, the average price-to-book ratio of insurance companies is close to one, while for most other industries it is substantially greater than one. With relatively small steady-state residual income, reported book value and explicit earnings forecasts capture most of the modeled value [Nissim (2013a)], and the imprecise terminal value has a less significant role.

The residual income model is likely to perform well in valuing insurance companies also, because the book value of equity, which anchors residual income valuation, is particularly important in this industry. Due to regulation, insurers’ ability to write premiums and generate income is directly related to their surplus, which is a regulatory proxy for equity capital. Relatedly, insurers are required by regulators to maintain minimum equity capital at levels commensurate with the levels and riskiness of their assets, liabilities, and activities; this requirement makes book equity a relatively useful measure of the scale of insurers’ operations.

Consequently, to obtain ICEC estimates for insurance companies, this study uses the residual income model with individual analysts’ earnings forecasts used as proxies for market expectations of future earnings as of the earnings forecasts’ announcement date. The valuation model utilizes the term structure of risk-free interest rates and thus focuses on the implied equity risk premium (IERP), the risk premium component of the implied cost of equity capital. An analysis of the IERP estimates yields the following findings.

During most of the sample period, which spans April 1983 through September 2012, the quarterly median implied equity risk premium was relatively stable, fluctuating mildly around an average value of about 5.5%. However, during the financial crisis of 2007-2009, the median IERP reached unprecedented levels, exceeding 15% in the first quarter of 2009. Following the financial crisis, the IERP declined substantially but it remained at historically high levels, exceeding 9% on average.

In spite of significant differences in operations and financial profile, the median IERP of Life and Health (LH) insurers was similar to that of Property and Casualty (PC) insurers during most of the sample period. However, during the financial crisis the median IERP of LH insurers was substantially larger than that of PC insurers, reaching 25% in the first quarter of 2009 compared to 11% for PC insurers. The large increase in the IERP of LH insurers is consistent with their high sensitivity to fluctuations in financial markets. The differences in the IERP across the industry groups remained substantial post the crisis, indicating a structural change in the pricing of LH insurers. Apparently, investors are now better aware of the high sensitivity of LH insurers to financial markets and price this sensitivity.

The IERP is strongly related to macro risk factors in a way suggesting that investors demand relatively high returns in periods of poor economic performance or high uncertainty. In particular, the IERP is positively related to the credit spread, term spread, and inflation, and negatively related to the 10-year Treasury yield. The cross-sectional correlations between the IERP and firm-specific risk factors are similarly consistent with expectations: the IERP is positively related to market beta, and negatively related to size and the equity-to-assets ratio. These

5 The primary bias is the reporting of most property and casualty loss reserves undiscounted. This bias is particularly large for insurers who specialize in long-tail liability lines.
sensitivities are generally larger for LH insurers and during the financial crisis. Finally, consistent with its strong correlations with macro- and firm-specific risk factors, the IERP performs well in predicting stock returns.

The paper proceeds as follow. The next section describes the residual income model used in deriving the implied equity risk premium. The following section contains the empirical analysis, and the final section summarizes and concludes. Appendix A discusses implementation issues.

The residual income model

The value of any financial claim is the present value of expected net flows to the owners of that claim. Accordingly, the value of common equity (Equity Value or EV) is the present value of expected net flows to common equity holders (Net Equity Flow or NEF):

$$EV_0 = \frac{E[NEF_0]}{1+r_0^*} + \frac{E[NEF_1]}{(1+r_0^*)^{1}} + \ldots = \sum_{t=0}^{\infty} E[NEF_t] \times (1 + r_t^*)^t$$  \(1\)

Where \(r_t^*\) is the risk-adjusted discount rate for cash flows to be received in year \(t\), and NEF include dividends, share repurchases and noncash distributions, net of share issuance.

Valuation model (1) can be restated in terms of comprehensive income available to common equity holders (comprehensive income or CI) and the book value of common equity (common equity or CE) by substituting the following relation for NEF:

$$NEF_t = CI_t - CE_t + CE_{t+1}$$  \(2\)

This relation postulates that changes in common equity are due to either comprehensive income or net equity flows. Given the definitions of NEF (discussed above) and CI (net income plus other comprehensive income), equation (2) accounts for essentially all changes in shareholders’ equity, and it therefore provides a reasonable approximation for the actual relationship between net equity flows, earnings, and book value.

The resulting valuation model (called residual income model) expresses intrinsic equity value as the sum of current book value and the present value of expected residual income in all future years, where residual income \((CI_t - f_t \times CE_{t+1})\) is earnings (CI) in excess of the return required by common equity investors given the amount (CE) and cost \((f_t)\) of common equity capital:

$$EV_0 = CE_0 + \sum_{t=1}^{\infty} \frac{E[CI_t - f_t \times CE_{t+1}]}{(1 + f_t^*)^t}$$  \(3\)

Where \(f_t^*\) is the time zero one-year risk-adjusted forward discount rate for year \(t\), i.e.,

$$f_t^* = \frac{(1 + r_t^*)^t - 1}{(1 + r_t^*)^{t+1}}$$

To demonstrate the residual income model, consider the following example. At time \(t = 0\) (current time), a firm with a book value of $10 is expected to exist for two years (i.e. until \(t = 2\)), and pay dividends of $1 in year 1 and $14 in year 2 (liquidating dividend). Expected earnings for year 1 are $2, and so, expected earnings for year 2 are $3 (= liquidating dividend minus year 1 book value, or $14-(10+2-1)). The price of a zero coupon one-year risk-free bond is $0.90909 per dollar of par value, and the price of a zero coupon two-years risk-free bond is $0.75757. These bond prices imply that the spot rate for one year is 10% (= 1/0.90909-1), the spot rate for two years is 14.89% (= [1/0.75757]^{1/2}-1), and the one year forward rate for year 2 is 20% (= 1.1489^{1/2}.1-1). Assuming that investors are risk-neutral and that dividends are paid at the end of each year, stock value can be calculated using the dividend discount model (equation (1)) as $11.515. Equivalently, value can be calculated using the residual income model. To do so, note that book value is expected to be $11 at time 1 (= 10+2-1), and residual earnings are expected to be $1 in year 1 (= 2-0.1 \times 10) and $0.8 in year 2 (= 3-0.2 \times 11). Using equation (3), therefore, price should equal $11.515 (= 10 + 1/1.1 + 0.8/[1.1489^{1/2}]).

Unlike the above example, investors are not risk-neutral. I, therefore, model the risk-adjusted forward discount rates \((f_t^*)\) as the sum of the corresponding risk-free forward rate \((f_t)\) and a risk premium \((prem)\). While there is some evidence that the term structure of equity risk premiums is not flat, for tractability I assume a constant risk premium, i.e., \(f_t^* = f_t + prem\). Therefore,

$$EV_0 = CE_0 + \sum_{t=1}^{\infty} \frac{E[CI_t - (f_t + prem) \times CE_{t+1}]}{(1 + f_t + prem)^t}$$  \(4\)

6 For a derivation of the model, see Ohlson (1995) and Nissim (2013b).

7 See, for example, Campbell et al. 1997, Chapter 8.
The Journal of Financial Perspectives

The equation (4) provides the theoretical foundation for the derivation of the IERP. However, its empirical implementation necessarily involves several assumptions and approximations. I use the same approach as in Nissim (2013b), and describe the assumptions and calculations in Appendix A. The following is a “big picture” depiction of the methodology. For each date on which an analyst provides a set of earnings forecasts for a given insurance company, I estimate the IERP associated with those forecasts by substituting the earnings forecasts and the closing stock price on that day for comprehensive income (CI) and intrinsic equity value (EV₀), respectively. To measure future book values (CEₜ), I adjust current book value (CE₀) by adding forecasted earnings and subtracting expected dividends, where future dividends are estimated assuming that current dividends will grow at the forecasted long-term earnings growth rate. For values post the explicit forecast period (i.e., for t > 5), I assume that residual income (i.e., CIₜ - (fₜ + prem) × CEₜ₋₁) will grow at a rate equal to the expected long-term economy-wide growth rate. I measure the forward risk-free rates (fₜ) on the forecasts announcement date. As a robustness check, I re-estimate the IERP using alternative assumptions regarding the evolution of residual income post the explicit forecast period and report the results below.

Empirical analysis
Sample and data
The sample used in this study includes all insurance companies with data available in the intersection of three databases: IBES, CRSP, and COMPUSTAT. Insurance companies are identified using the Global Industry Classification (GIC) system (industry GIC 403010), which is obtained from COMPUSTAT. COMPUSTAT is also the source of reported accounting data, including book value, total assets, and other variables. Market-related data (price, stock returns, dividends, shares, and adjustment factors) are extracted from CRSP. Economy-wide variables (interest rates, inflation, VIX, etc.) are obtained from the Federal Reserve Bank of St. Louis (http://research.stlouisfed.org/) and Yahoo! Finance.

Most insurance companies specialize in either property and casualty (PC) or life and health (LH), but some have significant operations in both segments. In addition, while many insurers underwrite reinsurance policies (insurance sold to insurers), some focus on reinsurance as their core activity. Insurers increasingly offer products and services that involve little or no insurance protection, such as investment products and fee-based services. The industry also includes companies that provide insurance brokerage services (sourcing of insurance contracts on behalf of customers). Reflecting this variation in activities, the GIC system classifies insurance companies as either Life and Health Insurers (LH, 40301020, for example, MetLife, Prudential, AFLAC), Property and Casualty Insurers (PC, 40301040, for example, Berkshire Hathaway, Allstate, Progressive), Multi-line Insurers (ML, 40301030, companies with diversified interests in life, health, and property and casualty insurance such as AIG, Hartford, and Lowes), reinsurers (Re, 40301050, for example, Reinsurance Group of America, Everest Re Group, PartnerRe), or insurance brokers (IB, 40301010, for example, AON, Marsh & McLennan, Willis). Because each of the last three sub-industries includes a relatively small number of firms, I treat these sub-industries as one group, and thus classify insurance companies into three groups: LH, PC, and “Other”.

Each sample observation corresponds to a set of EPS forecasts (typically EPS for the current and subsequent year and long-term EPS growth), provided by an analyst for a given insurer on a certain (announcement) date. These forecasts are matched with COMPUSTAT and CRSP to derive the estimated IERP as described in Appendix A. The resulting sample includes 13,916 observations (330 different firms) during the period April 1983 through September 2012. This sample is augmented with economy-wide and firm-specific variables (described below) that are likely to affect or be correlated with the IERP. Finally, to mitigate the effects of outliers, extreme values of the firm-specific variables are trimmed.8

IERP over time
Table 1 presents summary statistics for the IERP, the 10-year Treasury yield, the sum of these two variables (a proxy for the ICEC), and the other analysis variables. The mean IERP across all observations is 6.9%, while the mean 10-year Treasury yield

8 For each variable, I calculate the 1st and 99th percentiles of the empirical distribution (P1 and P99 respectively) and trim observations outside the following range: P1 - (P99 - P1) to P99 + (P99 - P1). For normally distributed variables, this range covers approximately 7 standard deviations from the mean in each direction (= 2.325 + (2.325 - (-2.325)), which is more than 99.99% of the observations. The variables used in this study are generally well-behaved, so few observations were deleted.
is 5.8%. Accordingly, the mean value of the ICEC is 12.6%. The median values of these variables are somewhat smaller – 6.2% for IERP, 4.9% for the 10-year rate, and 12.2% for the ICEC.

To examine the time-series variation in the IERP, the 10-year Treasury yield and their total (ICEC), Figure 1 plots the median values of these variables for each calendar quarter during the period Q2: 1983 through Q3: 2012. For most of the sample period (1986-2007), the median IERP was quite stable, fluctuating between 4% and 8% and averaging about 5.5%. However, in the early eighties the median IERP was less than 4%, and, since 2008, it exceeded 8%, surpassing 15% in the first quarter of 2009. These extreme levels of the IERP were associated with less abnormal values for the ICEC due to a negative correlation between the IERP and the 10-year Treasury yield. During the early eighties – when median IERP was very low – interest rates were very high, and during the financial crises of 2007-2009 – when median IERP reached unprecedented levels – the 10-year Treasury yield was at record low levels.

While the median IERP was relatively stable during most of the sample period, the median ICEC exhibited a negative trend due to a monotonic decline in interest rates. The negative trend of the ICEC was abruptly broken during the financial crisis of 2007-2009, which drastically increased the median IERP and hence median ICEC. In the two quarters following its peak in early 2009, the median IERP declined significantly to approximately 9.5% and it fluctuated around this level through September 2012, the end of the sample

### Table 1: Summary statistics

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>StdDev</th>
<th>5%</th>
<th>25%</th>
<th>Median</th>
<th>75%</th>
<th>95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implied equity risk premium</td>
<td>13,916</td>
<td>0.069</td>
<td>0.038</td>
<td>0.020</td>
<td>0.048</td>
<td>0.054</td>
<td>0.062</td>
<td>0.082</td>
</tr>
<tr>
<td>10-year Treasury yield</td>
<td>13,916</td>
<td>0.058</td>
<td>0.025</td>
<td>0.026</td>
<td>0.041</td>
<td>0.049</td>
<td>0.073</td>
<td>0.113</td>
</tr>
<tr>
<td>Implied cost of equity capital</td>
<td>13,916</td>
<td>0.126</td>
<td>0.034</td>
<td>0.086</td>
<td>0.106</td>
<td>0.122</td>
<td>0.141</td>
<td>0.177</td>
</tr>
<tr>
<td>Term spread</td>
<td>13,916</td>
<td>0.021</td>
<td>0.011</td>
<td>0.001</td>
<td>0.012</td>
<td>0.023</td>
<td>0.030</td>
<td>0.036</td>
</tr>
<tr>
<td>Credit spread</td>
<td>13,916</td>
<td>0.011</td>
<td>0.005</td>
<td>0.006</td>
<td>0.008</td>
<td>0.010</td>
<td>0.012</td>
<td>0.020</td>
</tr>
<tr>
<td>VIX / 100</td>
<td>11,006</td>
<td>0.214</td>
<td>0.095</td>
<td>0.116</td>
<td>0.150</td>
<td>0.192</td>
<td>0.248</td>
<td>0.376</td>
</tr>
<tr>
<td>Production growth</td>
<td>13,916</td>
<td>0.028</td>
<td>0.014</td>
<td>-0.002</td>
<td>0.020</td>
<td>0.028</td>
<td>0.037</td>
<td>0.049</td>
</tr>
<tr>
<td>Total assets (million)</td>
<td>13,916</td>
<td>59.167</td>
<td>131.412</td>
<td>369</td>
<td>3,212</td>
<td>12,378</td>
<td>48,957</td>
<td>314,150</td>
</tr>
<tr>
<td>Market value of equity (million)</td>
<td>13,916</td>
<td>8.095</td>
<td>18.069</td>
<td>188</td>
<td>2,852</td>
<td>8,211</td>
<td>27,851</td>
<td></td>
</tr>
<tr>
<td>Equity-to-asset ratio</td>
<td>13,863</td>
<td>0.219</td>
<td>0.152</td>
<td>0.051</td>
<td>0.109</td>
<td>0.187</td>
<td>0.285</td>
<td>0.511</td>
</tr>
<tr>
<td>Book-to-market ratio</td>
<td>13,811</td>
<td>0.827</td>
<td>0.497</td>
<td>0.270</td>
<td>0.527</td>
<td>0.746</td>
<td>0.995</td>
<td>1.599</td>
</tr>
<tr>
<td>Recurring ROE</td>
<td>13,496</td>
<td>0.137</td>
<td>0.117</td>
<td>-0.029</td>
<td>0.088</td>
<td>0.134</td>
<td>0.185</td>
<td>0.305</td>
</tr>
<tr>
<td>Market beta</td>
<td>12,647</td>
<td>0.958</td>
<td>0.610</td>
<td>0.245</td>
<td>0.625</td>
<td>0.867</td>
<td>1.131</td>
<td>2.290</td>
</tr>
<tr>
<td>Idiosyncratic volatility</td>
<td>12,647</td>
<td>0.080</td>
<td>0.037</td>
<td>0.044</td>
<td>0.058</td>
<td>0.070</td>
<td>0.092</td>
<td>0.142</td>
</tr>
<tr>
<td>Subsequent year excess return</td>
<td>13,145</td>
<td>0.086</td>
<td>0.334</td>
<td>-0.411</td>
<td>-0.102</td>
<td>0.086</td>
<td>0.252</td>
<td>0.584</td>
</tr>
</tbody>
</table>

9 As discussed earlier, the study uses the term “structure of interest rates” to measure the risk-free component of the cost of equity capital, and it thus derives a term structure of ICEC rather than a point estimate. Still, because most analysts use the 10-year Treasury yield as a proxy for the risk-free component of the cost of equity capital, the total of that rate and the estimated IERP can be used as a rough proxy for the ICEC.
period. While the median IERP has remained high compared to historical levels, the median ICEC post the financial crisis is comparable to historical levels due to unusually low interest rates.

Table 2 presents correlation coefficients (Pearson below the diagonal, Spearman above) for the analysis variables. As expected (given the patterns in Figure 1), the IERP is strongly negatively related to the 10-year Treasury yield. Consistent with investors demanding a relatively high-risk premium at times of uncertainty or poor economic conditions, the IERP is also positively correlated with the term spread (the difference between the 10-year and three-month Treasury rates), the credit spread (the difference between Moody’s Seasoned Baa and Aaa Corporate Bond Yields) and the VIX (implied volatility of the S&P 500 index), and negatively correlated with production growth (the percentage change in the Industrial Production Index). As is common in the literature, I use industrial production as a proxy for overall economic activity rather than real GDP. Industrial production has the advantage of being reported at a monthly frequency while real GDP is reported on a quarterly basis. Moreover, Hobijn and Steindel (2009) compare real GDP and industrial production in terms of their ability to indicate movements in aggregate economic activity and conclude that “our results suggest that movements in real GDP are not necessarily better at identifying such developments [in aggregate economic activity] than are movements in industrial production measures.”

The only macro variable that has an unexpected correlation with the IERP is inflation (the percentage change in the Consumer Price Index). However, the negative correlation between inflation and the IERP is probably due to the high correlation between inflation and the 10-year Treasury yield, which in turn is strongly negatively correlated with the IERP. In the regression analysis below I control for indirect effects by examining all relations simultaneously.

**IERP and firm characteristics**

In addition to the economy-wide variables, Tables 1 and 2 present summary statistics and correlation coefficient, respectively, for select firm characteristics: size, leverage, value, profitability, stock return, and systematic and idiosyncratic volatility. I next describe these variables and discuss the related statistics. The exact definitions of the variables are provided in the notes to Table 1.

The most common approach for estimating the cost of equity capital is the Capital Asset Pricing Model (CAPM). The fundamental premise of the CAPM is that the risk of a stock can be decomposed into two components – systematic risk, which is related to the overall market, and non-systematic (idiosyncratic) risk, which is specific to the individual stock. According to the CAPM, idiosyncratic risk is not priced (i.e., does not increase the discount rate) because it can be eliminated by holding a diversified portfolio. Systematic risk, in contrast, cannot be diversified away and therefore commands a risk premium. Under some stringent assumptions, systematic risk can be measured using the slope coefficient from a time series regression of the stock’s return on a proxy for the market return such as the S&P 500 Total Return. This regression is called the “market model” and the slope coefficient is called “market beta.” Thus, if the CAPM holds, a stock’s risk premium should increase with its beta and be unrelated to its idiosyncratic volatility. However, contrary to the CAPM premise, beta is at best weakly related to subsequent stock returns [Fama and French (1992)], and idiosyncratic volatility is correlated with future stock returns, although the sign of the correlation is controversial [for example, Ang et al. (2006) versus Fu (2009)].

Academic research and practice suggest that size (market value of equity) is at least as important as systematic volatility when estimating the cost of equity capital [Banz (1981), Fama and French (1992)]. Compared to small firms, large firms are better diversified, more likely to use financial hedging techniques, and more profitable. They also have greater financial flexibility, lower information risk
and lower variability in profitability and growth rates, and they may be considered “too big to fail” (for example, the government support to a number of the major global financial institutions during the financial crisis). Size is also correlated with stock liquidity, with small firms having high liquidity risk. More fundamentally, these relationships are due to factors such as economies of scale and scope, bargaining power in input and output markets, mature products, access to capital markets, market attention (analysts, institutional investors), and active trading.

Because creditors generally receive a constant return, the variability of the return generated on borrowed funds is absorbed by equity holders. Thus, financial leverage magnifies the variability of equity returns and increases both systematic and idiosyncratic risk. Debt also reduces financial flexibility. Because debt capacity is restricted, high-debt firms have limited ability to borrow additional funds when the need for such borrowing arises. Relatedly, high-debt firms are dependent on debt markets for continued refinancing and so are more sensitive to changes in interest rates, credit spreads, and funds availability, as was evident during the financial crisis of 2007-2009. Financial leverage also affects operating risks. When firms’ fortunes deteriorate, customers and other stakeholders often require additional consideration for transacting with the firm, exacerbating the negative shock that caused the initial decline in fortune. This is especially true for insurance companies, where financial stability is a crucial element of the product provided by the insurer. Additionally, due to extensive regulation – primarily restrictions on the ability to write premiums – insurers’ ability to generate business may deteriorate when losses due to financial leverage mount. Thus, while the empirical performance of financial leverage in predicting stock returns has generally been weak (Fama and French (1992)), its impact on risk is likely to be greater in the insurance industry.

Another financial ratio that is commonly used as a proxy for expected returns is the book-to-market ratio. Unlike financial leverage, this ratio has performed well in predicting stock returns, including in the financial sector (Barber and Lyon (1997)). One explanation for the return predictability of the book-to-market ratio is that book value is a proxy for expected flows (dividends, earnings, cash flow), while market value is a proxy for the present value of those flows. Thus, a high book-to-market ratio implies that investors use a relatively high discount rate in calculating the present value of expected flows, which in turn implies high expected returns and, therefore, high risk. In fact, the IERP is derived using a similar rationale but with a more direct measure of expected flows – earnings forecasts.11

11 In addition to financial leverage and the book-to-market ratio, other fundamentals such as the cashflow-to-price ratio, earnings momentum, accruals, and asset growth have also been shown to predict stock returns (Subrahmanyam (2009)). However, whether these factors are proxying for risk or market inefficiency is subject to debate. In addition, at least some of these variables are less relevant for financial services companies (for example, cashflow versus accruals). Therefore, do not include these variables in the analysis.
The book-to-market ratio is determined primarily by the relationship between profitability and the cost of equity capital. When profitability is higher than the cost of equity capital, the book-to-market ratio, which reflects investors’ pricing of excess profitability, should be less than one. As shown in Table 1, for insurance companies the mean and median values of recurring return on equity (ROE) — a book measure of shareholders’ profitability — are close to the mean and median values of the implied cost of equity capital, respectively, and accordingly, the mean and median values of the book-to-market ratio are close to one. These statistics support my argument above, that for insurance companies the book value of equity and expected residual earnings during the explicit forecast period capture much of the intrinsic value of equity, leaving a relatively small portion to be captured by the terminal value calculation.

The correlation coefficients between the IERP and the firm-specific variables, reported in Table 2, are almost all consistent with expectations and are highly significant (greater than 0.03 in absolute value). As predicted by the CAPM, the correlation between the IERP and market beta is positive and highly significant (0.39 Pearson, 0.16 Spearman). Also highly significant are the correlations with idiosyncratic volatility (0.28 Pearson, 0.22 Spearman), log of book-to-market (0.36 Pearson, 0.20 Spearman), and the equity-to-asset ratio (-0.12 Pearson, -0.13 Spearman). Most importantly, the correlation with the subsequent year excess stock return is positive and highly significant (0.20 Pearson, 0.07 Spearman). These bivariate correlations, however, reflect both direct and indirect effects as well as both time-series and cross-sectional covariation. Above I use fixed effect regressions to estimate the direct effects.

Sub-industry differences
Before turning to the regression analysis, I next examine differences in the risk profiles of insurers across the three sub-industry groups (Life and Health or LH, Property and Casualty or PC, and all others). Table 3 presents summary statistics for each of the sub-industries. As shown, LH insurers have substantially lower equity-to-assets ratios compared to PC and other insurers (median 9.8% versus 23.3% and 21.9% for PC and other insurers, respectively), suggesting that LH insurers are more risky and should therefore have higher risk premium. However, such inference is premature for at least two reasons. First, the difference in leverage is partially due to “separate accounts” which inflate the balance sheet of LH insurers. Separate accounts are similar to assets under management — insurers generally do not bear the risk or receive the return associated with these investments. Second, PC insurers are typically exposed to higher operating risks than LH insurers because both the frequency and magnitude of PC claims are more volatile. PC losses are highly sensitive to catastrophic events such as hurricanes, earthquakes, and terrorism acts, events which typically have limited effect.

<table>
<thead>
<tr>
<th></th>
<th>Life and health (Min N = 3,957)</th>
<th>Property and casualty (Min N = 5,198)</th>
<th>Other insurers (Min N = 3,492)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity risk premium</td>
<td>Median 0.061 Mean 0.068 StdDev 0.041</td>
<td>Median 0.060 Mean 0.063 StdDev 0.031</td>
<td>Median 0.068 Mean 0.077 StdDev 0.041</td>
</tr>
<tr>
<td>Total assets (million)</td>
<td>27,528 91,397 146,953</td>
<td>8,188 23,394 37,194</td>
<td>10,427 74,354 179,545</td>
</tr>
<tr>
<td>Market value of equity (million)</td>
<td>3,177 7,535 10,130</td>
<td>2,414 6,410 11,703</td>
<td>3,405 11,206 29,119</td>
</tr>
<tr>
<td>Equity-to-asset ratio</td>
<td>0.098 0.135 0.120</td>
<td>0.233 0.263 0.128</td>
<td>0.219 0.253 0.174</td>
</tr>
<tr>
<td>Book-to-market ratio</td>
<td>0.817 0.908 0.542</td>
<td>0.748 0.804 0.392</td>
<td>0.649 0.767 0.565</td>
</tr>
<tr>
<td>Recurring ROE</td>
<td>0.119 0.122 0.094</td>
<td>0.138 0.136 0.105</td>
<td>0.153 0.154 0.149</td>
</tr>
<tr>
<td>Market beta</td>
<td>1.009 1.116 0.649</td>
<td>0.777 0.817 0.408</td>
<td>0.888 0.990 0.751</td>
</tr>
<tr>
<td>Idiosyncratic volatility</td>
<td>0.075 0.082 0.035</td>
<td>0.070 0.076 0.026</td>
<td>0.067 0.083 0.051</td>
</tr>
<tr>
<td>Subsequent year excess return</td>
<td>0.116 0.113 0.361</td>
<td>0.077 0.080 0.302</td>
<td>0.064 0.063 0.346</td>
</tr>
</tbody>
</table>

Table 3: Summary statistics by sub-industry
See Table 1 for variable definitions. “Other” includes multi-line insurers (companies with diversified interests in life, health, and property and casualty insurance), reinsurers, and insurance brokers.
on LH claims. In addition, payments for PC insurance claims depend on policyholders’ incurred losses, while for LH insurance it is generally the face value of the policy. Furthermore, the distribution of PC insurance claims at the firm level can be highly skewed and heavy-tailed, implying that stock returns for PC insurers are likely to be particularly non-normal and may therefore command an incremental risk premium [Cummins et al. (1990)].

On the other hand, LH insurers have significant exposures to market factors that have relatively little impact on PC insurers. Similar to banks, a significant portion of LH insurers’ profits is derived from spread income – the difference between the yield on investment assets and the interest cost of insurance reserves. This source of income is sensitive to changes in interest rates, credit spreads and credit losses. In addition, LH insurers increasingly derive income from management and administrative fees on accounts whose balances are sensitive to market returns. Finally, many LH insurers have significant non-linear exposures to market returns due to various minimum benefit guarantees, primarily related to variable annuity products.

Consistent with the high leverage ratios of LH insurers and their exposures to financial markets, the market-related risk proxies in Table 3 indicate that LH insurers are more risky than PC and Other insurers. In particular, the mean beta of LH insurers is significantly larger than that of PC and Other insurers (p-value for the difference is less than 1% in both cases). In addition, compared to PC and other insurers, LH insurers have significantly higher book-to-market ratios and idiosyncratic volatility. Finally, the average excess stock returns of LH insurers are significantly larger than those of PC and other insurers.

The significant differences in operations and financial profile across the insurance sub-industries suggest that the IERP is likely to vary significantly across the three groups and be higher for LH insurers. Yet, the estimates in Table 3 indicate that the mean and median values of LH insurers’ IERP are significantly smaller than those of other insurers and are only marginally larger than those of PC insurers. Moreover, Figure 2, which reports quarterly median values of the IERP for each of the three sub-industries, indicates that, during most of the sample period, the three median IERP series were quite similar. However, starting in 2008 the median IERP of LH insurers increased dramatically and by early 2009 it reached unprecedented levels. The median IERP of PC and other insurers also increased significantly during that period, but by far less than the increase for LH insurers. In the first quarter of 2009, the median IERP was 25.5% for LH insurers, 11.2% for PC insurers, and 16.7% for other insurers. The large increase in the IERP of LH insurers is consistent with their high sensitivity to fluctuations in financial markets. The differences in the IERP

---

12 The interest cost of LH insurance reserves includes both explicit and implicit components. LH insurance reserves consist primarily of the liability for future policy benefits and policyholder account balances. The liability for future policy benefits represents the present value of future benefits to be paid to, or on behalf of, policyholders, including related expenses, less the present value of expected future net premiums to be received (essentially gross premiums minus embedded profits). This liability relates to traditional life insurance products such as term and whole life, and general account annuities with life contingencies. Because the liability is reported at present value, the related expense – the policyholder benefits expense, which measures the total of benefit payments during the period and the change in the liability – includes an implicit interest cost component. The liability for policyholder accounts relates to universal life policies and investment products; similar to bank deposits, these accounts earn interest and so generate an explicit interest cost for the insurance company. See Nissim (2010) for a discussion of accounting by insurance companies.

13 Unlike banks, which generate spread income by taking on both interest rate risk and credit risk, insurance companies generate spread income primarily by taking on credit risk. Banks borrow short-term and invest long-term and so benefit from the (usually) positive slope of the term structure, while, for most life insurance companies, asset duration is shorter than liability duration.

14 These include assets under management, separate accounts, and some policyholder accounts, with related portfolios included in general account assets. The balances of these accounts are affected by capital market performance both directly (the returns) and indirectly (net flows).
across the industry groups remained substantial after the crisis, indicating a structural change in the pricing of LH insurers. Apparently, investors are now better aware of the high sensitivity of LH insurers to financial markets and price this sensitivity.

Regression analysis
The analysis thus far indicates that the IERP is correlated with relevant macro factors and firm characteristics. However, as noted earlier, bivariate correlations reflect both direct and indirect effects. To estimate direct effects, I next turn to a regression analysis. Table 4 reports estimates from seven panel data regressions of the IERP on economy-wide and firm-specific risk factors as well as sub-industry dummies. Model 1 includes all the explanatory variables. The other six regressions omit the VIX variable, which is unavailable prior to 1990 and is insignificant in Model 1. Model 2 is estimated for three samples: using all observations, excluding the financial crisis (30 June 2007 through 31 December 2009 observations), and for the financial crisis. Model 3 includes firm fixed effect, Model 4 includes time fixed effect, and Model 5 includes firm and time fixed effects. The time effect is measured at a daily frequency and so completely spans the economy-wide variables, which are accordingly omitted from Models 4 and 5. The firm effect does not completely span the sub-industry classification variables due to changes in classification over time, but these are uncommon. The sub-industry dummies are, therefore, not included in Models 3 and 5. With fixed firm effect, each coefficient measures the impact on the dependent variable of a unit deviation of the related explanatory variable from its firm-specific average. Thus, the coefficients from firm fixed effect regressions (Models 3 and 5) are akin to coefficients from a change specification and so are more likely to capture direct effects of the related variables.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 2</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>All</td>
<td>Non FC</td>
<td>FC</td>
<td>All</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Intercept/fixed effect</td>
<td>0.0938</td>
<td>0.1022</td>
<td>0.0949</td>
<td>0.0946</td>
<td>Firm</td>
<td>Time</td>
<td>Firm&amp;Time</td>
</tr>
<tr>
<td>10-year Treasury yield</td>
<td>-0.634</td>
<td>-0.7632</td>
<td>-0.7413</td>
<td>-0.3470</td>
<td>-1.1142</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Term spread</td>
<td>0.0636</td>
<td>0.1210</td>
<td>0.1972</td>
<td>0.0612</td>
<td>0.1287</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Credit spread</td>
<td>2.8794</td>
<td>2.4054</td>
<td>1.3945</td>
<td>3.5872</td>
<td>1.9844</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VIX/100</td>
<td>18.8</td>
<td>22.9</td>
<td>13.9</td>
<td>10.7</td>
<td>22.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflation</td>
<td>0.1538</td>
<td>0.2165</td>
<td>0.3321</td>
<td>-0.0466</td>
<td>0.1929</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Production growth</td>
<td>0.0183</td>
<td>-0.0017</td>
<td>0.0703</td>
<td>0.0751</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Market beta</td>
<td>0.0152</td>
<td>0.0153</td>
<td>0.0074</td>
<td>0.0356</td>
<td>0.0118</td>
<td>0.0153</td>
<td>0.0158</td>
</tr>
<tr>
<td>Idiosyncratic volatility</td>
<td>0.0348</td>
<td>0.0551</td>
<td>0.1075</td>
<td>-0.1840</td>
<td>-0.0502</td>
<td>0.0765</td>
<td>-0.0745</td>
</tr>
<tr>
<td>Equity-to-asset ratio</td>
<td>-0.0365</td>
<td>-0.0326</td>
<td>-0.0218</td>
<td>-0.0694</td>
<td>-0.0608</td>
<td>-0.0258</td>
<td>-0.0531</td>
</tr>
<tr>
<td>Log of market value of equity/100</td>
<td>-0.3617</td>
<td>-0.3846</td>
<td>-0.2881</td>
<td>-0.5040</td>
<td>-1.8216</td>
<td>-0.3314</td>
<td>-1.8101</td>
</tr>
<tr>
<td>Dummy for LH insurers</td>
<td>-0.0089</td>
<td>-0.0067</td>
<td>-0.0065</td>
<td>-0.0214</td>
<td>-0.0084</td>
<td>-0.0067</td>
<td>-0.0067</td>
</tr>
<tr>
<td>Number of observations</td>
<td>9,835</td>
<td>12,596</td>
<td>10,884</td>
<td>1,712</td>
<td>12,596</td>
<td>12,596</td>
<td>12,596</td>
</tr>
</tbody>
</table>

Table 4: Regressions examining the determinants of insurers’ implied equity risk premium

The dependent variable in each regression is a firm/analyst/announcement date-specific estimate of the IERP. “FC” refers to observations during the financial crisis, from 1 July 2007 through 31 December 2009. See Table 1 for variable definitions. T-statistics are calculated using White’s (1980) heteroskedasticity-consistent standard errors.
In each of the five regressions of Table 4 that includes the macro variables, the credit spread is strongly positively related to the IERP – investors in insurance companies demand large risk premiums when economy-wide credit spreads are high. This was especially true during the financial crisis, as the coefficient on the credit spread is significantly larger in the financial crisis (FC) regression compared to the non-FC regression. (The t-statistics for the FC regression are relatively small due to the small number of observations and short time interval.) The coefficient on VIX is insignificant, but unreported results indicate that this is due to the credit spread subsuming the information in VIX. The 10-year Treasury yield is strongly negatively related to the IERP in each of the three “all” sample regressions as well as for the non-FC observations. Inflation and the term premium are positively related to the risk premium. These results are generally consistent with expectations. Investors demand a relatively high risk premium when the economy is performing poorly or when there is high uncertainty. The only coefficient which is inconsistent with this interpretation is that of production growth, which is positive rather than negative. This result may be due to production growth affecting the demand for equity capital, which in turn increases the equity premium.

Turning next to the firm-specific characteristics, the results are also consistent with expectations. In each of the seven regressions, the IERP is positively and strongly related to market beta and negatively and strongly related to the equity-to-assets ratio and size. Moreover, each of the three relations is substantially stronger during the financial crisis, and the differences in the coefficients between the FC and non-FC samples are statistically significant. Idiosyncratic volatility is positively related to the premium, but this relation reverses sign when firm fixed effects are included as well during the financial crisis.¹⁵ The coefficients of the LH and PC dummy variables are both negative and significant, suggesting that, after controlling for the risk factors, LH and PC insurers are on average less risky than other insurance companies (multi-line insurance companies, reinsurers, and insurance brokers).

¹⁵ I do not include the book-to-market ratio because, as discussed above, the IERP is derived based on the relation between the market and book values, so any measurement error in IERP is likely to be correlated with the book-to-market ratio and therefore bias the estimates.

Table 5 reports estimation results for the three industry subsamples. Most of the relations hold for each of the three sub-industry groups. In each case, the IERP is negatively and strongly related to the 10-year Treasury yield, the credit spread, the equity-to-assets ratio, and size. In addition, the IERP is positively related to inflation and market beta. The coefficients on the other variables – the term spread, production growth, and idiosyncratic volatility – are less consistent, but they are also less significant for the whole sample (Table 4). The primary differences across the industries are in the magnitudes rather than signs of the coefficients. In particular, the credit spread coefficient is significantly larger for LH insurers than for PC and other insurers. This result is consistent with the large magnitude (relative to equity) of LH insurers’ investment portfolios, which consist primarily of credit risky bonds. In addition, as noted earlier, LH insurers derive much of their income from various fees that are strongly related to capital market returns, and they provide minimum benefit guarantees that are tied to capital market performance. Indeed the coefficient on market beta is also significantly larger for LH insurers than for PC insurers. Compared to PC insurers, the IERP of LH insurers is also more sensitive to cross-sectional and time-series variation in size and the equity-to-assets ratio.

**Stock return predictability**

The estimates in Tables 4 and 5 show that the IERP is correlated with risk proxies, suggesting that it should predict excess stock returns. Table 6 presents results from regressing excess stock return (over the risk free rate) during the year following the IERP calculation on the IERP and control variables. Three sets of regressions are reported: OLS (Panel A), fixed effects (Panel B), and fixed effects with controls (Panel C). Each set of regressions is run for the full sample as well as for each of the three sub-industry groups. In all cases, the IERP coefficient is positive and highly significant. In contrast, none of the coefficients on the control variables (in Panel C) is consistently significant. The high significance of the IERP coefficient is notable given the low predictability of stock returns and the weak return predictability of IERP estimates in prior studies [Easton (2007)]. As explained earlier, the strong performance of the IERP in this study is likely due to the key role of financial instruments in the insurance industry. Indeed, for PC insurers, whose activities are less financing in nature, the IERP is less strongly correlated with the risk proxies (Table 5) and future excess stock returns (Table 6) compared to LH and other insurers.
If the IERP is measured with no error, and stock prices are efficient with respect to the information contained in the IERP, the IERP coefficient should equal one. If the IERP captures pricing inefficiencies and is measured with no error, the IERP coefficient should be greater than one as prices gravitate to intrinsic values.

Any non-systematic error contained in the IERP will bias the IERP coefficient downward. As shown in Table 6, the IERP coefficients are significantly greater than one in all regressions except for PC insurers. Thus, it appears that the IERP captures market mispricing in addition to risk. In other words, when price is low compared to analysts’ earnings forecasts and book value (so the IERP is high), subsequent stock returns are relatively high as price adjusts upward. The converse is true when price is relatively high.

Robustness
As described above and in Appendix A, the derivation of the IERP requires some strong assumptions, particularly regarding the evolution of residual income after the explicit forecast period. To evaluate the robustness of the results, I re-run all analyses using alternative IERP estimates calculated assuming that after year 5 residual income will (1) remain constant, (2) grow at a rate equal to expected inflation, measured using the University of Michigan Inflation Expectation Survey of Consumers, or (3) grow at a rate equal to expected inflation, measured based on the relationship between nominal and inflation-protected 10-year Treasury yields. The rationale for these choices is explained in Appendix A. The quarterly median IERP calculated using each of the three alternative

### Table 5: Regressions examining the determinants of insurers’ implied equity risk premium by sub-industry

<table>
<thead>
<tr>
<th></th>
<th>Life and health (LH)</th>
<th>Property and casualty (PC)</th>
<th>Other insurers (Other)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept/fixed effect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 2</td>
<td>0.1000 13.7</td>
<td>0.0764 19.8</td>
<td>0.1196 24.4</td>
</tr>
<tr>
<td>Model 5</td>
<td>-0.8032 -23.7</td>
<td>-0.6598 -32.2</td>
<td>-0.8427 -30.4</td>
</tr>
<tr>
<td>10-year Treasury yield</td>
<td>-0.1167 -7.7</td>
<td>0.2002 5.5</td>
<td>0.1892 4.1</td>
</tr>
<tr>
<td>Term spread</td>
<td>3.4236 17.5</td>
<td>1.2385 8.5</td>
<td>2.8637 14.2</td>
</tr>
<tr>
<td>Credit spread</td>
<td>0.2584 5.1</td>
<td>0.1533 4.0</td>
<td>0.2780 4.9</td>
</tr>
<tr>
<td>Inflation</td>
<td>0.0076 0.6</td>
<td>0.0234 2.2</td>
<td>-0.0277 -2.2</td>
</tr>
<tr>
<td>Production growth</td>
<td>0.0153 15.2</td>
<td>0.0168 4.8</td>
<td>0.0122 9.6</td>
</tr>
<tr>
<td>Market beta</td>
<td>0.1397 4.5</td>
<td>-0.0335 -6.6</td>
<td>0.091 5.0</td>
</tr>
<tr>
<td>Idiosyncratic volatility</td>
<td>-0.0698 -10.8</td>
<td>-0.0842 -5.8</td>
<td>-0.0277 -8.0</td>
</tr>
<tr>
<td>Equity-to-asset ratio</td>
<td>-0.5640 -7.9</td>
<td>-0.0852 -2.6</td>
<td>-1.1027 -8.1</td>
</tr>
<tr>
<td>Log of market value of equity/100</td>
<td>-10.0</td>
<td>-0.569 0.970</td>
<td>0.3175 0.8437</td>
</tr>
<tr>
<td>R-square</td>
<td>3.906 3.906</td>
<td>5.198 5.198</td>
<td>3.492 3.492</td>
</tr>
<tr>
<td>Number of observations</td>
<td>3,906 3,906</td>
<td>5,198 5,198</td>
<td>3,492 3,492</td>
</tr>
</tbody>
</table>

Note that the IERP measures the average premium in all future years, if the premium reflects mispricing, it is unlikely that that mispricing will be corrected at a constant rate over all future years. Mispricing tends to be corrected within a reasonably short period of time. The stock return during the correction period will be substantially larger than the mispricing component of the IERP. For example, if the risk free rate is 4%, the stock price is $100, and analysts correctly expect the company to pay $10 perpetual dividend, then the IERP is 6% (solve $100 = $10 / (0.04 + IERP)). If price reflects these expectations, the expected excess stock return next year is 6%. However, if price incorrectly reflects expected perpetual dividends of only $9 (i.e., investors use a discount rate of 9% (solve $100 = $9/r)), and price adjusts to reflect the correct dividend expectations in the following year, the new price will be $111.11 ($10/0.09) and the excess stock return will be 17.11% ($10 + ($111.11 - $100)) / $100 - 4%.

The rationale for these choices is explained in Appendix A. The quarterly median IERP calculated using each of the three alternative
Implied cost of equity capital in the U.S. insurance industry

Panel A: OLS regressions

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Life and health</th>
<th>Property and casualty</th>
<th>Other insurers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.0393</td>
<td>-0.0798</td>
<td>0.0554</td>
<td>0.0993</td>
</tr>
<tr>
<td>Equity risk premium</td>
<td>1.8496</td>
<td>2.8877</td>
<td>0.3964</td>
<td>2.1378</td>
</tr>
<tr>
<td>R-square</td>
<td>0.0405</td>
<td>0.1002</td>
<td>0.0016</td>
<td>0.0597</td>
</tr>
<tr>
<td>Number of observations</td>
<td>13,145</td>
<td>4,136</td>
<td>5,361</td>
<td>3,648</td>
</tr>
</tbody>
</table>

Panel B: Fixed firm and time regressions

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Life and health</th>
<th>Property and casualty</th>
<th>Other insurers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity risk premium</td>
<td>2.617</td>
<td>3.2356</td>
<td>1.2878</td>
<td>3.0184</td>
</tr>
<tr>
<td>R-square</td>
<td>0.7191</td>
<td>0.8776</td>
<td>0.8121</td>
<td>0.8017</td>
</tr>
<tr>
<td>Number of observations</td>
<td>13,145</td>
<td>4,136</td>
<td>5,361</td>
<td>3,648</td>
</tr>
</tbody>
</table>

Panel C: Fixed firm and time regressions, controlling for firm characteristics

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Life and health</th>
<th>Property and casualty</th>
<th>Other insurers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity risk premium</td>
<td>2.1104</td>
<td>2.0703</td>
<td>1.2449</td>
<td>2.3959</td>
</tr>
<tr>
<td>Market beta</td>
<td>-0.0169</td>
<td>0.0813</td>
<td>-0.046</td>
<td>-0.0727</td>
</tr>
<tr>
<td>Idiosyncratic volatility</td>
<td>-0.7472</td>
<td>-3.0592</td>
<td>2.4525</td>
<td>-0.4369</td>
</tr>
<tr>
<td>Equity-to-asset ratio</td>
<td>0.2773</td>
<td>-0.5136</td>
<td>0.4167</td>
<td>0.3916</td>
</tr>
<tr>
<td>Log of market value of equity/100</td>
<td>-11.9146</td>
<td>-10.775</td>
<td>3.9799</td>
<td>-22.3344</td>
</tr>
<tr>
<td>Log of the book-to-market ratio</td>
<td>0.0890</td>
<td>0.2629</td>
<td>0.2869</td>
<td>-0.0461</td>
</tr>
<tr>
<td>R-square</td>
<td>0.7502</td>
<td>0.9025</td>
<td>0.8394</td>
<td>0.8218</td>
</tr>
<tr>
<td>Number of observations</td>
<td>11,899</td>
<td>3.643</td>
<td>4.956</td>
<td>3.300</td>
</tr>
</tbody>
</table>

Table 6: Regressions examining the relationship between the implied equity risk premium and future excess stock return

The dependent variable in each regression is the excess stock return over the year following the earnings forecasts’ announcement date, measured relative to the risk free rate. See Table 1 for variable definitions. t-statistics are calculated using White’s (1980) heteroskedasticity-consistent standard errors. “Other” includes multi-line insurers (companies with diversified interests in life, health, and property and casualty insurance), reinsurers, and insurance brokers.

terminal value assumptions as well as under the initial assumption are presented in Figure 3. As shown, for the 80s and 90s the alternative assumptions reduce the IERP by about one-to-two percentage points, but they generally do not change its pattern over time. From the early 2000s, the differences across the estimates decline substantially, and from the beginning of the financial crisis the four estimates converge. I also replicate all tables with each of the three alternative IERP estimates and, in all cases, find similar results to those reported, confirming the robustness of the findings with respect to the most critical assumption made in the IERP calculation.

Conclusion

This paper derives and evaluates estimates of the implied equity risk premium (IERP) of U.S. insurance companies by inverting the residual income model, utilizing analysts’ earnings forecasts as proxies for market expectations of future earnings. During most of the sample period, April 1983 through September 2012, the quarterly median IERP was relatively stable, fluctuating mildly around an average value of about 5.5%. However, during the financial crisis of 2007-2009, the median IERP reached unprecedented levels, exceeding 15% in the first quarter of 2009. Following the financial crisis, the IERP declined substantially but it remained at historically high levels, exceeding 9% on average.

In spite of significant differences in operations and financial profile, the median IERP of Life and Health (LH) insurers was similar to that of Property and Casualty (PC) insurers during most of the sample period. However, during the financial crisis the median IERP...
of LH insurers was substantially larger than that of PC insurers, consistent with the higher sensitivity of LH insurers to fluctuations in financial markets. The differences in the IERP across the sub-industry groups remained substantial post the crisis, indicating a structural change in the pricing of LH insurers.

Consistent with investors demanding relatively high expected returns in periods of poor economic performance or high uncertainty, the IERP is positively related to the credit spread, term spread, and inflation, and negatively related to the 10-year Treasury yield. The relations with firm-specific risk factors are similarly consistent with expectations: the IERP is positively related to market beta, and negatively related to size and the equity-to-assets ratio. These sensitivities are generally higher for LH insurers and during the financial crisis. Finally, consistent with the strong correlations between the IERP and the macro- and firm-specific risk factors, the IERP performs well in predicting subsequent excess stock returns. The stock return predictability apparently reflects mispricing in addition to compensation for risk.

These findings improve our understanding of the determinants of, and proxies for, equity risk in the insurance industry. The IERP estimates derived in the study perform better than standard risk measures and also indicate potential mispricing. The results demonstrate a structural change in the risk pricing of insurance companies: following the financial crisis, investors demand a substantially higher risk premium, primarily from LH insurance companies. In addition, the risk sensitivities of LH insurers are substantially higher risk premium, primarily from LH insurance companies: following the financial crisis, investors demand a substantially higher risk premium, primarily from LH insurance companies. In addition, the risk sensitivities of LH insurers are significantly different from, and generally larger than, those of other insurers. This evidence suggests that the current trend in accounting regulation to eliminate accounting differences across insurance operations may not be desirable [see, for example, Nissim (2010)].

Appendix A: Derivation of the implied equity risk premium

Given a set of EPS forecasts announced by an analyst (typically EPS for the current and subsequent year and long term EPS growth), I estimate the IERP for that insurer/analyst/announcement date observation by solving the following equation for \( \text{prem} \):\(^{17}\)

\[
P = \left[ \frac{\text{BVPS}_0 + \sum_{t=1}^{5} \left( \text{EPS}_t (1 + \text{prem}) \times \text{BVPS}_0 \right)}{\prod_{t=1}^{5} \left( 1 + f_t + \text{prem} \right)} \right] \times (1 + f_t + \text{prem})^0
\]

Time 0 is the end of the most recent fiscal year (year 0) for which 10-K has been filed as of the earnings forecast announcement date. \( PC \) (cum-dividend price) is the closing share price on the earnings forecast announcement date, adjusted for the re-investment of dividends between the end of fiscal year 0 and the forecast announcement date. The rationale for the dividend adjustment is that the residual income model values the stock relative to book value at the beginning of the first residual income period (time 0, beginning of fiscal year 1). Consequently, any dividend paid between time 0 and the forecast announcement date reduces price but has no direct effect on the valuation.

The other variables are measured as follows. \( \text{BVPS}_0 \) or book value per share at time 0, is calculated by dividing book value by shares outstanding, and is adjusted for stock splits and stock dividends between time 0 and the forecast announcement date. \( \text{EPS}_t \) is the analyst’s EPS forecast for fiscal year \( t \), \( t = 1, \ldots, 5 \), where \( t = 1 \) denotes the fiscal year that starts at time 0. Forecasts for \( t = 3, 4, \ldots \)

\(^{17}\) In some cases analysts announce subsets of the EPS forecast series (EPS1-EPS5 and EPS growth) on different dates. To complete forecast series, I look back up to 31 days for annual EPS forecasts and 93 days for EPS long-term growth forecasts, provided that there is no change in the fiscal year to which the forecasts correspond. These adjustments substantially increase the sample size but do not change any of the inferences.

Figure 3: U.S. insurers’ median implied equity risk premium under alternative terminal value assumptions

The plots present median values for each calendar quarter of four alternative estimates of the equity risk premium of insurance companies. The estimates differ with respect to the terminal value assumption used in the valuation model from which the implied premium estimate is derived.
or 5 are often unavailable. In such cases, I estimate missing forecasts by assuming a constant growth rate equal to the long-term EPS growth forecast, provided that the base forecast is positive.\footnote{18} BVPS, for $t = 1, \ldots , 4$, is forecasted using the following relation:

$$BVPS_t = \frac{DPS_t}{1 + f_{LT}}$$

where $DPS_t$ is forecasted dividend per share for year $t$, $t = 1, \ldots , 4$. DPS is measured as the total of dividends from time 0 through the forecast announcement date and expected dividends for the remainder of the year. Expected dividends are measured based on the most recently declared dividend.\footnote{19} DPS is measured by assuming that the annualized most recently declared dividend will grow at the long-term earnings growth rate in year 2. DPS, and DVPS are estimated by assuming that DPS will continue to grow annually at the same rate as the long-term earnings growth forecast.

$f_{LT}$ is the forward risk-free rate for future year $t$, derived from the term-structure of U.S. Treasuries as of the earnings forecast announcement date. $f_{LT}$ is a proxy for expected long-term interest rates at time $t=5$, measured using the annualized risk-free forward rate from $t=5$ to $t=10$, as derived from the term-structure of U.S. Treasuries as of the earnings forecast announcement date. This rate tends to be less volatile than the 10-year rate and lies between the 10 and 30 years rates. prem is the insurer/analyst/reporting firm’s estimate of a 10-year par yield.

The “max” term measures the present value of expected residual income in all years after year 5. It is the equivalent of the terminal-value calculation in the DCF model. Although this calculation might appear somewhat arbitrary, it is based on both economic and statistical considerations as explained below. In addition, I conduct extensive robustness checks using alternative terminal-value calculations and report the results above.

The first term inside the max function assumes that after year 5 residual earnings will grow at a constant rate equal to $f_{LT}$, a proxy for the expected long-term risk-free rate. The rationale for this assumption is that steady state firms are generally expected to grow at a rate consistent with the nominal long-term growth in overall economic activity, which in turn should be close to the long-term risk-free rate. The nominal long-term growth rate is approximately equal to the total of expected inflation and real growth, while the risk-free rate is approximately equal to the total of expected inflation and the real rate of interest. To the extent that real interest rates predict real returns, which in turn determine real growth, the long-term risk-free rate can serve as a proxy for nominal long-term growth. Given this assumption, the first term in the max function should be familiar — it is essentially the Gordon model applied to residual income.\footnote{20} An important statistical property of this calculation is that it generates few if any outliers because changes in the risk-free and long-term growth rates offset each other (without this offset, the denominator in a Gordon-like calculation may be close to zero or even negative, generating outliers or meaningless estimates, respectively).

Because residual income measures return relative to the cost of equity capital, negative residual income means that the firm is generating a negative net return to shareholders after considering the cost of equity capital. In such cases, growth destroys rather than creates value, and firms have no incentives to increase invested capital. Accordingly, when residual income is negative, the second term in the max function effectively sets expected residual income after year 5 equal to its level in year 5.\footnote{21}

The final term in the equation adjusts the valuation for the expected increase in value from time 0 (the end of the most recent fiscal year for which financial results have been reported) through the forecast announcement date, where $F$ measures the length of that period.

\footnote{20} According to the Gordon model, price is equal to $d \times (1 + g)/(r - g)$, where $d$ is dividend per share, $g$ is the constant long-term growth rate, and $r$ is the cost of equity capital. If the long term growth rate is equal to the risk free rate, $r - g$ is equal to the risk premium.

\footnote{21} An alternative assumption, which is used in the robustness section, is to set residual income after year 5 equal to zero when residual income in year 5 is negative. However, companies are often unable to exit investments without incurring the full loss. Consider, for example, life insurers’ investments in long-term fixed income securities. If interest rates rise, residual income will be negative over many years. And if the firm sells the investment, it will still incur the full loss.
References


Nissim, D., 2013a, “Relative valuation of U.S. insurance companies,” Review of Accounting Studies, 18(2)


Subrahmanyam, A., 2009, “The cross-section of expected stock returns: what have we learnt from the past twenty-five years of research?” Working Paper, UCLA

The Dodd-Frank Act, Solvency II, and U.S. insurance regulation

Scott E. Harrington
Alan B. Miller Professor, Director, Wharton/Penn Risk and Insurance Program, The Wharton School, University of Pennsylvania

Abstract
This article discusses key issues that policymakers should consider when evaluating potential changes to U.S. insurance regulation in the aftermath of the financial crisis, including implications of both the Dodd-Frank Act and the Solvency II initiative in the European Union. Fundamental differences in the U.S. between banking and insurance are emphasized, including much lower systemic risk potential and greater market discipline in insurance, and why those differences favor capital regulation and policyholder guaranty systems that reflect the distinctive features of each sector.

1 This paper updates and expands papers presented at the Networks Financial Institute 7th Annual Insurance Reform Summit, Washington, D.C., 16 March 2011 (distributed as Networks Financial Institute Policy Brief 2011-PB-01) and at the Temple University conference on Convergence, Interconnectedness, and Crises: Insurance and Banking, 8-10 December, Philadelphia, PA. Some of the discussion draws from my September 2009 NAIC Issue Brief, The Financial Crisis, Systemic Risk, and the Future of Insurance Regulation (edited version published as Harrington, 2009). The views expressed and any errors are mine and unrelated to organizations with which I am affiliated.
Introduction
The Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010 could have far-reaching implications for U.S. insurance regulation. Paralleling in many respects global initiatives of the G20 and the Financial Stability Board, the Dodd-Frank Act created the Financial Stability Oversight Council (FSOC) to oversee systemic risk, with the authority to identify systemically significant “nonbank financial companies,” including insurance companies, which will be subject to heightened supervision by the Federal Reserve. It also created the Federal Insurance Office (FIO) within the U.S. Department of Treasury to monitor all aspects of the U.S. insurance industry and negotiate, and enter into, international agreements concerning prudential matters for insurance and reinsurance. The FIO is required to study and report to the U.S. Congress on numerous aspects of insurance regulation, including systemic risk and capital requirements, regulation of insurance companies and affiliates on a consolidated basis, international coordination of insurance regulation, the degree of national uniformity of state regulation, the costs and benefits of federal regulation for different lines of insurance, and the potential consequences of subjecting insurance companies to federal resolution authority, including the effects on state insurance guaranty funds.

The FSOC has thus far promulgated rules for the identification of systemically important nonbank institutions, and the Federal Reserve is developing enhanced supervision procedures and additional capital requirements for systemically important nonbanks. The Federal Reserve has also proposed applying the Basel III bank regulatory regime to nonbank financial institutions that own savings and loans entities, including organizations that predominantly engage in insurance activities, and several insurance organizations have chosen to divest savings and loan subsidiaries.

As the implementation of the Dodd-Frank Act evolves, the National Association of Insurance Commissioners (NAIC) is exploring key issues in solvency regulation and supervision through its Solvency Modernization Initiative, including core principles, group solvency, capital requirements, corporate governance, international reinsurance transactions, and international accounting and regulatory standards. Debate over U.S. insurance regulation encompasses those efforts and the adequacy and effectiveness of the NAIC’s risk-based capital (RBC) system. Possible alternatives to the NAIC RBC system in principle include the adoption of capital standards along the lines of the E.U.’s Solvency II initiative. Solvency II requires that non-E.U. countries’ insurance regulatory systems meet certain criteria in order for their capital, group supervision, and reinsurance standards to be deemed equivalent to those of the E.U. If deemed equivalent, E.U. supervisors will essentially rely on the home country’s supervision and treat insurers and reinsurers from that country the same as E.U. insurers and reinsurers.

This article discusses key issues that policymakers should consider when evaluating potential changes to U.S. insurance regulation in the aftermath of the financial crisis, including the implications of the Dodd-Frank Act and the Solvency II initiative. Despite some convergence, most U.S. financial institutions can be expected to specialize predominantly in either banking (and/or securities) or insurance for the foreseeable future. The article emphasizes fundamental differences between insurance and banking, including low systemic risk potential and stronger market discipline in U.S. insurance, and how those differences favor regulatory and policyholder guaranty systems that reflect the distinctive features of each sector.

The paper first reviews the insurance sector’s role in the financial crisis and the extent to which insurance poses systemic risk. It turns next to possible FSOC designation of one or more insurance entities as posing a threat to the financial stability of the U.S., and the potential consequences of such designation. I then discuss key issues that should be considered carefully in the policy debate over additional changes in U.S. insurance regulation, including whether the U.S. should move towards a banking model of insurance solvency regulation in general and a Solvency II approach in particular.

The financial crisis and systemic risk
Prior to the financial crisis, many U.S. banks, investment banks, thrifts, hedge funds, mortgage originators, and mortgage borrowers assumed substantial risk in anticipation of continued housing price appreciation. If prices continued to climb, they would have achieved large returns. If prices fell, much of the resulting loss would be borne by others. Explicit or implicit guarantees of debt issued by a number of GSEs (Government Sponsored Enterprises) lowered financing costs and contributed to the housing bubble in general and rapid expansion of subprime

106  The Journal of Financial Perspectives
lending in particular. Deposit insurance and implicit guarantees of banks’ obligations likewise encouraged risky lending, especially as the U.S. Congress pressed banks and the GSEs to expand lending to low-income borrowers.

Some banks located significant amounts of leveraged investment in off-balance sheet vehicles. Investment banks helped spread the risk of housing price declines and mortgage defaults through securitization. The conversion from partnerships to corporate ownership very likely encouraged major investment banks to assume more risk in relation to capital, especially given expanded competition with investment bank affiliates of bank holding companies. The Securities and Exchange Commission’s adoption in 2004 of consolidated supervision of the major investment banks allowed them to substantially increase leverage and take on more subprime exposure.

Many subprime mortgage originators were relatively new entrants with little reputational capital at risk and only modest participation in the risk of underlying mortgages. In tandem with growing speculation on housing prices, the Federal Reserve kept short-term interest rates at historically low levels, fueling demand for credit and housing, and it encouraged relaxation in historical underwriting standards. Regulators in the U.S. and abroad uniformly failed to anticipate the coming crisis.

Government assistance
The financial meltdown led to U.S. government assistance to several major insurance companies, including total federal commitments to provide assistance to just one major insurer growing to U.S.$182 billion and assistance provided reaching over U.S.$130 billion. The government’s funding support was used predominantly to meet demands of bank and investment bank counterparties to credit default swap and securities lending transactions.2

Although securities lending activities contributed to significant problems and collateral calls, the original government assistance was precipitated by a liquidity crisis resulting largely from non-insurance activities, not core insurance operations, especially credit default swap activities that were not conducted by regulated insurance subsidiaries. Otherwise, with the exception of a few specialty mortgage/bond insurers, the insurance sector withstood the crisis reasonably well. General property/casualty insurers and most life insurers avoided severe adverse consequences. It remains uncertain whether any insurance operating companies would have become insolvent if the government had not intervened.

The financial crisis reflected fundamental failures in U.S. and foreign banking regulation, including in the U.S., by the Office of Thrift Supervision (OTS), the Office of the Comptroller of the Currency (the regulator of federally chartered commercial banks), the FDIC, the U.S. Securities and Exchange Commission (SEC), and the Federal Reserve.3 Banking regulation and supervision permitted high leverage, aggressive investment, inadequate capital for risky loans and securitizations, and complex and highly leveraged off-balance sheet vehicles, often financed by commercial paper.

Many credit default swap and securities lending counterparties were regulated by U.S. and foreign banking regulators. Because broad regulatory authority encompasses responsibility for monitoring an institution’s relationships with counterparties, banking regulators also bear significant responsibility for not recognizing the risks of allowing regulated banking entities to (1) buy extensive amounts of credit protection from insurance companies, and (2) provide large amounts of securities lending collateral to them.

Systemic risk
In addition to motivating the adoption in the U.S. of regulation requiring the identification and enhanced supervision of systemically important nonbank financial institutions, the events of 2008 have stimulated substantial research and analysis of the extent to which insurance involves systemic risk. Although there is no uniform definition, the term “systemic risk” generally is used broadly to encompass the risk of any large, macroeconomic shock that affects financial stability, and the risk arising from extensive interconnectedness among firms, with an attendant risk

2 Congressional Oversight Panel (2010) provides a comprehensive review of the problems and government assistance. Also see Harrington (2009).

3 Until the crisis, U.S. insurance organizations that owned savings and loan subsidiaries were subject to consolidated regulation and oversight by the OTS. The OTS, which was eliminated by the Dodd-Frank Act, was also responsible for regulating several large mortgage finance organizations that eventually failed and were merged with, or acquired by, other entities with Federal Deposit Insurance Corporation (FDIC) assistance.
of significant economic spillovers on the real economy.\(^4\) There is a distinction between the risk of common shocks to the economy, such as widespread reductions in housing prices or large changes in interest rates, which have the potential to directly harm large numbers of people and firms, and financial risk that arises from interconnectedness and contagion. It often is difficult empirically, however, to distinguish any contagion effects from the effects of common shocks.

While there were elements of counterparty and information-based contagion, the driving factor underlying the general financial crisis was a common shock: the bursting of the housing bubble and declines in values of mortgage-related securities. Regarding the insurance sector, whether credit default swaps and securities lending presented significant risk of contagion remains uncertain, including the extent to which a major bankruptcy would have had significant adverse effects beyond counterparties, or the extent to which counterparties had hedged their exposure or otherwise reduced their risk. A failure by the U.S. government to channel hundreds of billions of dollars of assistance to those counterparties would have weakened their financial condition, forcing some to sell more assets and reducing their ability to invest, trade, and/or make loans. Some E.U. banking counterparties would have needed to raise more capital or reduce their risk exposure. Without federal government intervention, many more insurance policyholders might have terminated or declined to renew their policies. That by itself, however, would not imply contagion or that those customers would be significantly harmed.

Analyses generally conclude that the core activities of insurers pose little systemic risk, especially compared with banking, in part because many insurers hold relatively large amounts of capital in relation to their liabilities and have relatively little exposure to short-term liabilities, reducing their vulnerability to shocks (Swiss Re (2003); Harrington (2004a); Geneva Association (2010, 2012); IAIS (2011); Cummins and Weiss (2013),\(^5\) Shocks to insurers do not threaten the economy’s payment system and short-term lending. Banking crises have much greater potential to produce rapid and widespread harm to economic activity and employment.

Differences in systemic risk between insurers and banks help explain historical differences in regulation across the sectors. Other things being equal, greater systemic risk favors stronger government guarantees of financial institutions’ obligations to protect consumers and deter economically inefficient runs. Greater systemic risk also favors more stringent capital requirements. The issues are related. Stronger guarantees increase moral hazard: they reduce market discipline for financial institutions to be financially strong, increasing the need for more stringent capital rules. Thus, systemic risk strengthens the case for relatively broad government guarantees of bank obligations and relatively strict financial regulation, including capital requirements.

Because insurance activities pose much less systemic risk, there is less need for strict capital requirements and relatively broad guarantees to prevent inefficient runs that would destabilize the economy. Insurance guarantees in the U.S. have been appropriately narrower in scope than in banking, and, as is elaborated below, market discipline is reasonably strong. Insurers commonly have held much more capital than required by regulation and have not faced strong incentives for regulatory arbitrage and other evasion.

\(^4\) In their detailed review of the literature on systemic risk in insurance and reinsurance, Cummins and Weiss (2013) define systemic risk as “the risk that an event will trigger a loss of economic value or confidence in a substantial segment of the financial system that is serious enough to have significant adverse effects on the real economy with a high probability.” They suggest that primary indicators of systemic risk at the firm level include size (volume of exposures), interconnectedness, and a lack of substitutability for a firm’s services.

\(^5\) Based on their detailed literature and analysis, for example, Cummins and Weiss (2013) conclude that “the core activities of U.S. insurers do not pose systemic risk.” In contrast to IAIS (2011), Cummins and Weiss conclude that “life insurers are vulnerable to intra-sector crises” and that “both property-casualty and life insurers are vulnerable to reinsurance crises arising from counterparty credit risk.” Research on financial institutions’ stock prices provides evidence of interconnectedness among insurers and other financial firms and develops new metrics for measuring systemic risk with stock price data. Billio et al. (2011), for example, use principal components analysis and Granger causality tests to analyze stock returns for insurers, banks, securities brokers, and hedge funds during 1994-2008. They find evidence of causal relationships between the sectors during 2001-2008 but not 1994-2000, and they identify several insurers as systematically important. Acharya et al. (2010) develop a measure of systemic risk (systemic or marginal expected shortfall) to reflect a firm’s tendency to lose value when the overall market suffers large losses. Their analysis of stock returns for insurers and other financial firms during 2006-2008 suggests that insurance firms were the least systemically risky. Insurers with the largest systemic risk measures had significant activity in credit derivatives and financial guarantees.
The FSOC and systemic risk regulation

The FSOC is charged with (1) identifying risks to financial stability from “the material financial distress of large, interconnected bank holding companies or nonbank financial companies, or that could arise outside the financial services marketplace”; (2) promoting market discipline “by eliminating expectations on the part of shareholders, creditors, and counterparties that the government will shield them from losses in the event of failure”; and (3) responding to “emerging threats to the stability of the U.S. financial system.” The FSOC has 10 voting members from member agencies (including, among others, the Federal Reserve Chair, Treasury Secretary, SEC Chair, FDIC Chair, Comptroller of the Currency Director, and a presidential appointee with expertise in insurance) and five non-voting members, including the Office of Financial Research Director, the FIO Director, a state insurance commissioner, a state banking commissioner, and a state securities commissioner.6

Section 113 of the Dodd-Frank Act provides the FSOC with the authority, by a two-thirds vote, to designate a nonbank financial company, including an insurance company, as systemically important (by imposing a threat to the financial stability of the U.S.) and subject to enhanced regulation and supervision by the Federal Reserve. The Federal Reserve is required to establish, with input from the FSOC, enhanced risk-based capital requirements, leverage rules, resolution standards, and other requirements for systemically important nonbank financial companies. Section 113 specifies factors the FSOC must consider in determining whether a company will be subject to enhanced regulation and supervision by the Federal Reserve. The Federal Reserve is required to establish, with input from the FSOC, enhanced risk-based capital requirements, leverage rules, resolution standards, and other requirements for systemically important nonbank financial companies. Section 113 specifies factors the FSOC must consider in determining whether a company will be subject to enhanced regulation and supervision by the Federal Reserve. The Federal Reserve is required to establish, with input from the FSOC, enhanced risk-based capital requirements, leverage rules, resolution standards, and other requirements for systemically important nonbank financial companies. Section 113 specifies factors the FSOC must consider in determining whether a company will be subject to enhanced regulation and supervision by the Federal Reserve.

The rules establish a three-stage process for determination of whether a nonbank financial company poses a threat to the financial stability of the U.S. Stage 1 employs publicly available information and information from member regulatory agencies to identify nonbank financial companies for more detailed evaluation in Stage 2. A company is evaluated further in Stage 2 if its global consolidated assets are U.S.$50 billion or greater and it meets at least one of five additional quantitative thresholds. Stage 2 entails a review and prioritization of Stage 2 entities based on analysis of each company, using information available to the FSOC through existing public and regulatory agencies and information obtained from the company voluntarily. Based on this analysis, the FSOC notifies companies it believes merit further evaluation in Stage 3, including analysis of additional information collected directly from the company. The evaluation considers the company’s resolvability in the event of financial distress and includes consultation with the company’s primary regulator. Following Stage 3 analysis, a “Proposed Determination” of a company as systemically significant requires a two-thirds vote of the FSOC, followed by a hearing if the company requests, and, if so, a final vote. As of December, 2012, the FSOC had notified several insurance organizations that they are subject to Stage 3 evaluation.

I view Section 113’s provisions for identifying systemically significant nonbanks for enhanced supervision by the Federal Reserve as misguided. The large U.S. investment banks that survived the financial crisis became bank holding companies and are, therefore, regulated by the Federal Reserve, with enhanced supervision required by the Dodd-Frank Act (based on their size, as are other large banks). Arguments against designating specific nonbanks as systemically significant and subject to enhanced regulation include:7

---

6 The Dodd-Frank Act established the Office of Financial Research within Treasury to provide support to the FSOC through data collection, data standardization, long-term research, and development of risk-measurement and monitoring metrics.

7 I made these arguments in Harrington (2009).
• Greater capital requirements and tighter regulation for entities designated as systemically important raises the risk that they could face excessive burdens and costs that would disrupt competition and harm customers.

• The possibility of being designated as systemically important and subject to more stringent regulation creates uncertainty for large, nonbank financial companies that could likewise distort their financial and operating decisions in undesirable ways.

• Given the failure of regulators (and other professionals and academics) to anticipate the financial crisis, there can be little confidence that enhanced supervision of companies designated as systemically important will significantly reduce the likelihood of any future crisis.

• Most important, there is little reason to hope that “systemically important” will not translate as “too big to fail.” Regardless of legislative and agency assertions that creditors and shareholders of companies designated as systemically important will not be bailed out in the event of financial distress, there is a real risk that government assistance would be likely for companies identified as systemically important, which could ultimately worsen the moral hazard problem.8

These objections notwithstanding and as noted above, strong arguments can be made that the core activities of insurers do not pose systemic risk. With very few exceptions, reasonable application by the FSOC of the statutory criteria and/or six categories of factors in its rules and guidance should not result in designation of any U.S. companies that predominantly write property/casualty insurance as systemically important. Although some large life insurers may pose greater systemic risk than property/casualty insurers, reasonable application of the criteria should also exclude most if not all life insurers from designation as systemically important. If any insurers are so designated, the competitive dynamics of the insurance market will likely be affected in unpredictable ways.

U.S. regulatory modernization
Key and related issues regarding the potential for additional changes in U.S. insurance regulation include: (1) whether federal regulation should be further expanded, (2) how to encourage robust market discipline, (3) the appropriate design of policyholder guarantees, (4) the appropriate design of capital requirements, and (5) group supervision. These issues are all linked to the possible expansion of bank regulatory models to insurance. The last two are closely related to the E.U.’s Solvency II initiative.

Federal regulation
Pre-crisis pressure in the U.S. for optional federal chartering or other fundamental changes in state insurance regulation focused on the costs and delays of regulatory approval of policy forms in over 50 different jurisdictions; the costs, delays, and possible short-run suppression of rates below insurers’ projected costs associated with prior regulatory approval of insurers’ rate changes; and restrictions on insurers’ underwriting and risk classification.9 The pricing and underwriting issues were primarily relevant for property/casualty insurers. Regulation of policy forms was the overriding issue for life insurers. Optional federal chartering was often viewed as a potential mechanism for achieving desirable regulatory change with suitable deregulation, including improved ability to deal with multi-jurisdictional issues within the U.S. and internationally.

The financial crisis fundamentally changed the debate as to whether expanded federal regulation is appropriate for dealing with systemic risk and solvency issues. While the scope is uncertain, the Dodd-Frank Act has the potential to significantly expand federal regulation of U.S. insurance through possible designation of systemically significant insurance organizations for enhanced supervision by the Federal Reserve. The application of the Basel III framework to insurance organizations that own savings and loans entities would likewise expand the reach of federal regulation.

One traditional argument against optional federal chartering of insurers is that the states have performed reasonably well in many respects, including regulating solvency and analyzing and revising regulation to deal with perceived problems (including, for example, the current NAIC Solvency Modernization Initiative).

8 To quote one example of this view, Hubbard et al. (2009) write: “Identifying an institution as systemically important creates a moral hazard, since the market will view this designation as the equivalent of a bailout guarantee. A perceived bailout guarantee will decrease these institutions’ costs of raising capital.”

9 I provided details in Harrington (2006). In contrast to the early 1970s and early 1990s, when temporary increases in the frequency and severity of insurance company insolvencies motivated optional federal chartering proposals, pressure for optional federal chartering from the late 1990s until the crisis was not motivated by solvency concerns.
Another key argument is that the potential risks and costs of optional federal chartering are large compared with the uncertain benefits. In particular, the history of federal deposit insurance and “too big to fail” policy creates some risk that optional federal chartering could expand government guarantees of U.S. insurers’ obligations, undermine market discipline and incentives for safety and soundness, and increase the likelihood of future federal bailouts of insurance organizations.

In addition, even if optional federal chartering were to reduce the scope of insurance price regulation initially, it could ultimately produce broad restrictions on pricing and underwriting at the federal level, to achieve political or social goals (as has been illustrated by the health insurance rate review and minimum medical loss ratio requirements in the U.S. health-care reform law enacted in 2010). Such restrictions would increase cross-subsidies among policyholders, inefficiently distort some policyholders’ incentives to reduce the risk of loss, and increase risk to federal taxpayers if political pressure led to inadequate rates. Alternatives to optional federal chartering that might have the potential for improving insurance regulation with less risk include: (1) selective federal preemption of inefficient state regulations, such as prior approval rate regulation in competitive markets and inefficient impediments to nationwide approval of certain products, and (2) allowing insurers to choose a state for primary regulation with authorization to operate nationwide, primarily under the rules of that state.

The recent crisis and government bailouts did not establish the case for some form of federal chartering of U.S. insurers. In addition to the arguments against federal regulation of individual insurers designated as systemically significant, discussed above, it is not clear how the financial crisis fundamentally altered the potential benefits and costs of optional federal chartering. The problems that some firms experienced cannot be primarily attributed to any insurance regulatory failure. While disputed and subject to some uncertainty regarding the fungibility of capital among affiliates, it appears that state regulated insurance operating entities would likely have been able to largely or completely meet their obligations to policyholders without federal intervention, with state insurance guaranty funds serving as a potentially important back-up if they could not.

The alternative scenario — where insolvency of insurance operating entities had required multibillion dollar assessments under the state guaranty system — might have strengthened the argument for federal regulation. Even then, however, a strong case for federal regulation in response to the crisis would need to explain how pre-crisis federal insurance regulation would have mitigated the industry’s problems. That would not be easy in view of what happened at major bank and investment bank holding companies. There can be no presumption that federal regulation of insurance operations would have mitigated risk-taking, or that optional federal chartering, with or without mandatory federal regulation of insurers deemed systemically important, would mitigate any role of insurance in some future financial crisis.

Market discipline
A lack of market discipline represents a key underlying cause of the financial crisis. A simple explanation for much of the aggressive risk-taking by commercial banks, investment banks, savings and loans, mortgage originators, and mortgage borrowers was that the potential gains and losses were asymmetric. If housing prices continued upwards, participants could achieve large profits. If prices stabilized, or even fell, the losses would be borne largely by other parties, including taxpayers. The extended period of historically low interest rates further encouraged high leverage and fueled risky borrowing, lending, and investment.

A primary objective of additional legislative, regulatory, and administrative responses to the financial crisis should be to encourage market discipline, to promote safety and soundness in all types of financial institutions. A key, stated objective of the Dodd-Frank Act is to reduce the likelihood of future bailouts of financial institutions. Whether it is likely to achieve that goal is uncertain, but an overriding goal of its implementation, and of any future changes in financial regulation, should be to constrain “too big to fail” policy. As noted above, a potential downside to FSOC designation of an insurer as systemically important and subject to enhanced supervision by the Federal Reserve is that it could ultimately undermine market discipline if creditors, counterparties, investors, and/or consumers anticipate additional protection against the consequences of an insurer’s financial distress.
It is fundamentally important for policy to recognize that insurance markets generally have been characterized by relatively strong market discipline and correspondingly low insolvency risk. Many, if not most, policyholders prefer to deal with financially strong insurers and are willing to pay the higher costs that greater financial strength require (even with limited state guarantees, see below). A variety of institutions, including the widespread use of insurance intermediaries (agents, brokers, advisors), private ratings of insurers’ financial strength, and, for business coverages, knowledgeable corporate staff who oversee risk-management and insurance programs, help match risk-sensitive policyholders with financially strong insurers. In addition, insurance production and distribution often involve the creation of sizable franchise value, which could diminish or evaporate if an insurer experiences financial distress. Protection of those assets from loss due to financial difficulty, therefore, provides a significant incentive for adequate capitalization and other forms of risk management by insurers. Finally, many insurers in the U.S. issue debt, primarily at the holding company level. Until the events of 2008, that debt was effectively subordinated to policyholder claims, creating an additional category of stakeholders concerned with risk management.

**Policyholder guarantees**

Given the importance and scope of existing market discipline, policy should recognize the fundamental importance of avoiding any significant expansion of government guarantees of insurers’ obligations and carefully consider the potential effects of possible federal regulation on the state guaranty system. In contrast to “too big to fail” policy, and consistent with lower systemic risk in insurance than in banking, protection provided by state guaranty funds is relatively narrow, which reduces moral hazard and helps preserve market discipline.

The policy debate will likely consider arguments that the state system of ex-post funding of guarantees is defective. That system, however, has several advantages. Coupled with potential borrowing by state guaranty associations, pre-funding is not necessary for the system to have substantial capacity, and ex-post funding helps provide incentives for financially strong insurers to press for effective regulatory oversight. Some academics assume that pre-funding of guarantees would be preferable, in large part because they believe it would permit insurers to be charged accurate, risk-based premiums for guaranty protection, thus encouraging incentives for financial soundness. In practice, however, premiums would likely be based only crudely on risk. The adoption of pre-funded guarantees could yield its disadvantages without enough risk-rating to significantly improve incentives.\(^\text{11}\)

Given that federal guarantees would likely be viewed by policyholders as stronger than state guarantees, a federal guaranty system for federally chartered insurers in a system of optional federal chartering might soon supplant state guarantees — so that most, or even all, insurers would end up being backed by a federal guaranty. As an alternative (and as has been previously proposed), federally chartered insurers could be required to participate in the state guaranty system. However, the longrun stability of such a system would be uncertain. If a federally chartered insurer operating in multiple states failed, there would likely be substantial criticism of any state differences in guaranty coverage. Failure of a state chartered insurer — with federally chartered insurers having to contribute — could likewise lead to substantial criticism of state guarantees and solvency oversight.

A possible belief that the federal government would stand behind federally chartered insurers would distort competition between state and federally chartered insurers, especially if any insurer were designated as systemically important.

As a result, substantial pressure for federal guarantees could be inevitable, with optional federal chartering, perhaps by designing federal guarantees along the lines of existing state guarantees. The design of any government guarantees also might be tailored in principle, to help encourage additional market discipline. It should be recognized, however, that a monopoly federal guaranty program might ultimately ensue. That result could undermine market discipline, in principle requiring stricter capital requirements and regulation.

---


\(^{11}\)  The Dodd-Frank Act adopted ex-post funding of the Orderly Liquidation Fund to repay possible borrowings from the Treasury by the FDIC, via “postevent” assessments on bank holding companies with U.S.$50 billion or more in assets and any nonbank financial companies designated as systemically important (with graduated assessments to be proposed by the FSOC based on assets, risk, and other factors).
Capital requirements and Solvency II

The effectiveness of the NAIC risk-based capital (RBC) system and possible alternatives, such as Solvency II, are important issues in the U.S. solvency modernization debate, including whether significant changes will be required if the U.S. is to achieve equivalence under Solvency II. In contrast to the U.S. RBC system and associated solvency regulation, Solvency II is patterned after the three-pillar Basel II (and III) framework for bank regulation. It relies on market (consistent) valuation of assets and liabilities, in contrast to the U.S. reliance on statutory accounting valuation. Solvency II’s quantitative capital requirements (Pillar I) involve calculation of quantitative (stochastic) risk models. The “standard capital model” will employ a 99.5% value-at-risk metric for a one-year horizon in determining required capital. Moreover, and again borrowing from the bank regulatory model of Basel II, the Solvency II approach is viewed as principles-based rather than rules-based, including that it will provide an insurer with the opportunity to convince regulators that it needs less capital, based on its own (internal) capital model.

As part of its Solvency Modernization Initiative, the NAIC is considering changes to its RBC system that currently appear likely to retain its basic structure. The NAIC RBC standards have been criticized on a number of dimensions, including that the types of risk reflected, risk charges, and aggregation methods are ad hoc and unnecessarily crude, especially compared with more sophisticated stochastic risk models. Some academic analyses have concluded that the U.S. should move in the direction of Solvency II’s reliance on stochastic modeling and regulatory authorization/approval of internal models. Cummins and Phillips (2009) (also see Holzmüller (2009)), for example, recommend U.S. adoption of market valuation, stochastic risk modeling, and internal capital models.

Solvency II might well be appropriate for the European Union, and it has been designed with great study and effort. I am not convinced, however, that more sophisticated quantitative models to derive “value-at-risk” (or “tail value-at-risk”) are appropriate for achieving the objectives of insurance capital regulation in the U.S., given the potential benefits and costs involved, and the demonstrable failures of the Basel system leading up to the financial crisis. It almost always can be argued that capital standards, even those based on relatively complex formulas, such as the NAIC standards, are insufficiently rigorous and that additional refinements and analytical sophistication are warranted. But analytical sophistication invariably outpaces reliable, practical application with real-world data. While seemingly precise, mathematically sophisticated risk models need not be accurate when applied in practice. Their potential value in analyzing a firm’s risk, conceptually and in providing guidance to its management, need not imply that they are an appropriate foundation for regulatory capital, especially through deference to a regulated entity’s own model.

Some observers suggest that the relatively low levels of aggregate U.S. RBC compared to total industry capital indicate that the NAIC RBC formulas do not require sufficient capital. It is important for U.S. policymakers to consider, however, that relatively low levels of RBC in relation to actual capital for the bulk of insurers makes sense in an environment of significant market discipline. Current RBC standards in the U.S. have relatively little effect on the operating and financial decisions of a large majority of insurers, which hold much more capital than required by the standards. Even with refinements of risk measures, or changes in the basic modeling framework, to improve accuracy, any significant increase in the overall level of U.S. RBC could lead to undesirable distortions in decisions of many financially sound insurers. The effects could include reduced willingness to offer coverage, less appropriate investment strategies, and higher prices, especially following any large, negative shocks to insurer capital.

In Harrington (2004a), I discussed this issue in detail and developed a simple model of the optimal stringency of capital standards in relation to the degree of market discipline when imperfect information about insurer risk is available to regulators.

12 Eling and Holzmüller (2008) and Cummins and Phillips (2009) compare NAIC RBC requirements and Solvency II. Also see Eling et al. (2007).

13 A variety of academic studies analyze the ability of the NAIC RBC requirements to predict U.S. insurer insolvencies (e.g., Cummins et al. (1995); Grace et al. (1998); Cummins et al. (1999); Pottier and Sommer (2002)). The evidence generally indicates that NAIC RBC ratios lack predictive power compared with models that include other metrics.

14 Dowd et al. (2011), for example, provides a sharp and detailed critique of the Basel regime of bank capital regulation and its contribution to the financial crisis.

15 According to an early economic evaluation of U.S. RBC requirements, Cummins et al. (1993) stated that: “The insolvency record and the potentially severe adverse consequences of excessive risk-based capital charges provide substantial support for using a ‘minimum threshold’ approach to setting the stringency of risk-based capital. Under this approach, risk-based capital would serve as a minimum or floor on capital that would be expected to constrain only the weakest insurers.” The importance of this point has been overlooked in some subsequent evaluations that build on Cummins et al. (e.g., Holzmüller (2009)).
information concerning capital adequacy results in costly Type 1 and 2 errors (i.e., failing to constrain some inadequately capitalized insurers and inefficiently distorting the decisions of some adequately capitalized insurers). The model formalizes the intuitive prediction that capital standards that minimize the total cost of Type 1 and 2 errors will be less stringent, the greater the proportion of insurers that would be adequately capitalized without regulation. The key implication is that optimal capital standards should not bind most insurers in a market characterized by strong market discipline.

The U.S. RBC system, where most insurers hold significantly more capital than the required minimums [see Harrington (2004a); Cummins and Phillips (2009)], is consistent with reasonably strong market discipline in U.S. insurance markets. For 2011, the median ratio of total adjusted capital to Company Action Level RBC was 526% for all life companies and 520% for all property/casualty insurers [NAIC (2012a,b); data reported for individual insurers].

On average, U.S. RBC ratios decline with insurer size (assets). Figure 1 shows median RBC ratios by year during 2007-2011 for U.S. companies with more than $10 billion of assets. The median ratios for life companies range from a low of 389% in 2008 to a high of 462% in 2011. The medians for property/casualty companies range from a low of 237% in 2008 to a high of 278% in 2010. About 97% of all companies in both sectors had total adjusted capital above their company action RBC level each year during 2007-2011.

Evidence suggests that Solvency II could be binding for relatively more E.U. insurers. Regardless of whether that would be appropriate for the E.U., the case has not been made that tighter requirements would be appropriate for the U.S. A move towards a Solvency II model with an emphasis on sophisticated risk modeling in the U.S. could be accompanied by more stringent requirements that unnecessarily distort the decisions of some financially strong insurers and do not pass a cost/benefit test. Yet, without additional stringency, the question would arise as to why the potential for modest improvements in the quantification of standards that are non-binding for most insurers would justify the large costs of substantially changing the U.S. system.

Some specific questions that should be explored carefully by policymakers (and in academic analysis) in evaluating whether it would be appropriate to move U.S. capital regulation toward a Solvency II model include:

- Does the magnitude of insolvency risk in U.S. insurance markets justify the large personnel and systems costs that would be necessitated by fundamental changes in capital requirements?
- To what extent would close regulatory oversight of capital decisions and regulatory discretion to impose additional capital charges under Pillar 2 of a Solvency II-type system risk excessive intrusion into management decision-making and de facto shadow management by regulators?
- Is stochastic modeling likely to enhance the accuracy of risk assessment significantly compared with less sophisticated and more judgmental approaches, given the potential for large

16 Companies with total adjusted capital below the Company Action Level are required to submit plans for increasing capital to regulators.

17 According to the Solvency II QIS5 report (EIOPA (2011, p. 25)), the median E.U. insurer among sample participants had a ratio of funds available to the calculated Solvency Capital Requirement (SCR) of between 150 and 200%. Fifteen percent of participants had insufficient funds to meet the SCR. Once Solvency II is implemented, companies with funds below their SCR will be required to increase capital. A number of studies have considered the extent to which the Solvency II standards will affect insurers’ decisions, including investment decisions. Höring (2012) reviews some of the literature and provides evidence that, while required capital charges for the market-risk component under Solvency II exceed those of Standard & Poor’s for a comparable value-at-risk benchmark, Solvency II’s diversification and loss absorption adjustments cause the Standard & Poor’s model to require 68% more capital for the same market risks. He concludes that the Solvency II requirements will not significantly influence insurers’ investment strategies.

Figure 1: Median RBC ratios for U.S. life and property/casualty insurers with assets above U.S.$10 billion (2007-2011)
Note: Ratios of total adjusted capital to company action level RBC.
Source: NAIC (2012a,b); author’s conversion to company action level basis.
errors in estimating model parameters (such as value-at-risk inputs for particular risks and correlations between risks) and the reliance on some degree of judgment in selecting many parameters?18

Would allowing the use of regulator-approved internal capital models serve a useful purpose when RBC standards are non-binding for most insurers?

To what extent would allowing the use of regulator-approved internal capital models disadvantage smaller insurers, which lack sufficient volume to parameterize internal models and spread fixed costs of model development, implementation, and maintenance?

To what extent would a Solvency II approach lead to greater uniformity of models used to assess capital adequacy by insurers, rating agencies, and other entities, thus reducing diversity and competition in risk assessment and increasing the risk of system-wide consequences from model errors?

To what extent would arguably more-precise capital standards increase consumerist and political pressure for regulatory suppression of rates in some regulated lines of business and jurisdictions for insurers’ with substantial “excess” capital?

As it stands, my overall assessment is that the insolvency record in the U.S., the design of U.S. capital standards compared with Solvency II, and the direct and indirect costs involved do not support fundamental change in U.S. RBC requirements.19

Ongoing refinements in U.S. requirements should, instead, continue to recognize the distinctive nature of U.S. insurance markets.

As explained above, systemic risk in banking has encouraged broad government guarantees and capital requirements that constrain risk-taking by many institutions, in part to reduce moral hazard. Binding capital requirements generate pressure from banks to relax requirements, including by allowing the use of internal models. Insurance involves much less systemic risk and thus need for stringent capital requirements combined with relatively broad guarantees of firms’ obligations. Relatively strong market discipline favors capital requirements that generally are easily met by the bulk of insurers, reducing undesirable distortions of sound companies’ operating decisions and incentives for evading the requirements. Less constraining capital requirements make costly attempts at greater precision less desirable.

Group supervision

State regulation in the U.S. has generally focused on individual insurers (legal entities), in contrast to insurance regulation in some non-U.S. jurisdictions and U.S. banking regulation, which focus more attention on consolidated supervision. A certain degree of coordination in U.S. insurance solvency regulation is achieved through deference to regulators in an insurer’s state of domicile (or lead domiciliary regulator in the case of multiple domiciles) and through coordinated analysis and dissemination of financial metrics for larger, multistate insurers. State regulators and the NAIC have also relied on an elaborate statutory and administrative framework to deter holding companies that experience financial difficulties from draining funds or otherwise undermining the solvency of insurance subsidiaries.

Solvency II requires group-wide supervision of E.U. insurance entities and, as noted earlier, “equivalent” supervision by non-E.U. countries. In 2009, the International Monetary Fund’s Financial Sector Assessment program recommended that U.S. regulators expand assessments of individual insurers to the group level. The NAIC is revising its model holding company acts, dealing with relationships between parents and state-licensed subsidiaries. It is also exploring enhanced methods for information sharing regarding holding company risk and solvency and the possibility of developing group-wide capital standards. Similar to Solvency II, the NAIC is proposing that groups conduct and provide regulators with “Own Risk Solvency Assessments” of their risks, capital needs, and capital adequacy on a consolidated basis. U.S. state regulators have also begun to participate in “supervisory colleges” of regulators from different countries, to share information about internationally active insurance groups, and the NAIC is working with the International Association of Insurance Supervisors on the ComFrame (Common Framework for the Supervision of Internationally Active Insurance Groups) initiative for enhancing oversight with greater consistency and coordination.

18 Mittnik (2011) provides evidence that calibration of the Solvency II equity risk module is seriously flawed, by relying on a rolling window approach to expand the number of annual data points for use in calibration. He concludes (p. 37): “The results of this study strongly suggest that the implementation of the Standard Formula with its currently proposed equity-risk calibrations is imprudent, if not irresponsible.”

19 Nor do I find compelling, arguments that bank-type solvency regulation and capital requirements are needed for U.S. insurers, to discourage regulatory arbitrage.
When assessing possible changes in U.S. regulation with respect to group supervision, policymakers should carefully consider these efforts at greater coordination. A key goal should be to understand and assess the potential benefits of collaborative frameworks for multi-jurisdictional supervision that do not rely on federal insurance regulation in the U.S., in comparison with systems that would provide a greater role for federal regulation.

Conclusion
Despite some convergence, most U.S. financial institutions can be expected to specialize predominantly in either banking or insurance activities for the foreseeable future. Insurance has fundamental differences from banking, including much lower potential for systemic risk and stronger market discipline. These differences favor regulatory and guaranty frameworks in the U.S. that reflect the distinctive features of each sector. They should be carefully considered by U.S. policymakers in the potential identification of systemically important insurance companies and the analysis of potential changes in insurance regulation, including any that would move U.S. capital and solvency regulation towards a Solvency II model.

References
Congressional Oversight Panel, 2010, The AIG rescue, its impact on markets, and the government's exit strategy, June
EIOPA, 2011, EIOPA Report on the Fifth Quantitative Impact Study (QIS5) for Solvency II, EIOPA-TFQIS5-11/001, March
Harrington, S. E., 2004b, “Capital adequacy and insurance risk- based capital systems,” Journal of Insurance Regulation, 28, 26-72
IAIS, 2011, Insurance and financial stability, International Association of Insurance Supervisors, November
NAIC, 2012a, “Health industry RBC results for 2011,” Kansas City, Mo
NAIC, 2012c, “Property & casualty industry RBC results for 2011,” Kansas City, Mo

116 The Journal of Financial Perspectives
What drives the development of the insurance sector? 
An empirical analysis based on a panel of developed and developing countries

Erik Feyen
Senior Financial Specialist, Financial and Private Sector Department, World Bank

Rodney Lester
Senior Consultant, World Bank

Roberto Rocha
Senior Advisor, Financial and Private Sector Department, World Bank

Abstract
The insurance sector can play a critical role in financial and economic development. By reducing uncertainty and the impact of large losses the sector can encourage new investments, innovation, and competition. As financial intermediaries with long investment horizons, insurance companies can contribute to the provision of long-term instruments to finance corporate investment and housing. There is evidence of a causal relationship between insurance sector development and economic growth. However, there have been few studies examining the factors that drive the development of the insurance industry. This paper contributes to the literature by examining the determinants of insurance premiums (both life and non-life premiums) and total assets for a panel of about 90 countries during the period 2000-08. The results show that life sector premiums are driven by per capita income, population size and density, demographic structures, income distribution, the size of the public pension system, state ownership of insurance companies, the availability of private credit, and religion. The non-life sector is affected by these and other variables. While some of these drivers are structural, the results also show that the development of the insurance sector can be influenced by a number of policy variables.

Key words: Life and non-life, insurance development determinants, econometric analysis

1 We would like to thank Alain Ize, Heinz Rudolph, and Rodolfo Wehrhahn for detailed comments and Diego Sourrouille for valuable research assistance.
What drives the development of the insurance sector?
An empirical analysis based on a panel of developed and developing countries

Introduction
The insurance sector plays a critical role in financial and economic development. By introducing risk pooling and reducing the impact of large losses on firms and households, the sector reduces the amount of capital that would be needed to cover these losses individually, encouraging additional output, investment, innovation, and competition. By introducing risk-based pricing for insurance protection, the sector can change the behavior of economic agents, contributing inter alia to the prevention of accidents, improved health outcomes, and efficiency gains. As financial intermediaries with long investment horizons, life insurance companies can contribute to the provision of long-term finance and more effective risk management. Finally, the sector can also improve the efficiency of other segments of the financial sector, such as banking and bond markets (e.g., by enhancing the value of collateral through property insurance, and reducing losses at default through credit guarantees and enhancements).

Empirical research generally finds evidence of a causal relationship between insurance sector development and economic growth, although some of the studies report mixed results. Ward and Zurbruegg (2000) find evidence of causation in some OECD countries. Webb et al. (2002) find that measures of banking and life insurance predict economic growth, although these individual measures lose power in the presence of interaction terms, suggesting complementarities between these two sectors. Kugler and Ofoghi (2006) find evidence of long-run causality from insurance to GDP growth for eight insurance categories in the U.K.. Arena (2008) finds evidence of a causal relationship of insurance development on economic growth in a large panel of 56 countries and 28 years (1976 to 2004). In the case of life insurance, the results are driven by developed countries, while in the case of non-life insurance the results are driven by both developed and developing countries. Haiss and Sumegi (2008) build a panel of 29 E.U. countries from 1992 to 2005 and conclude that the life sector had a positive impact on growth in the E.U.-15 countries, while the non-life sector had a larger impact in Central and Eastern Europe. Finally, Ćurak et al. (2009) examine the relationship between insurance sector development and growth in the 10 new E.U. Member States between 1992 and 2007 and conclude that the sector has promoted economic growth in these countries.

Despite the critical role that the insurance sector may play for financial and economic development and the reasonable evidence that the sector has promoted economic growth, there have been few studies examining the factors that drive the development of the insurance sector. Moreover, the bulk of the existing empirical research focuses on the growth of the life sector, as measured by life insurance premiums. This paper contributes to the body of research in several ways. First, by building a large and recent database comprising 90 developed and developing countries over the 2000-08 period. Second, by examining the factors that drive the development of both the life and non-life sectors. Third, by exploring the impact of a broad set of potential determinants, including new variables that have not yet been tested. Finally, by measuring the development of the life and non-life sectors using both premiums and assets.

The paper is structured as follows. Section 2 provides a review of the empirical research on the determinants of the insurance sector’s development, with focus on the more recent studies. Section 3 discusses the data and the methodology used in the paper. Section 4 presents and discusses the main results, including univariate, two-group comparison tests and regression results. Finally, section 5 summarizes the main findings and identifies the major policy implications for the development of the insurance sector.

A review of the empirical literature on the insurance sector
The empirical research on the determinants of the insurance sector has essentially focused on the life sector. Early contributions to the empirical literature were mostly focused on the U.S., and explored the role of education, income, religion, and cultural factors, as well as prices. By way of illustration, Hammond et al. (1967) stress the importance of the occupation of the main wage earner. Headen and Lee (1974) show that the demand for life insurance depends on savings and interest rates. Burnett and Palmer (1984) show that education, income, and religion are key determinants of the demand for life insurance. Babbel (1985) shows that prices have a negative effect on the demand for life insurance. Beenstock et al. (1986) provided one of the first empirical studies exploring cross-country data. Using a dataset of 10 developed countries, the authors conclude that income, life expectancy, and the dependency ratio have a positive impact on life insurance demand, while social security expenditures have a negative impact.
More recent empirical studies include Browne and Kim (1993), Outreville (1996), Beck and Webb (2003), and Li et al. (2007). These studies rely on larger cross-country datasets, especially the latter two. Browne and Kim (1993) use data from 1987 for a cross-section of 45 developed and developing countries, while Outreville (1996) bases his analysis on a cross-section of 48 developing countries for the year 1986. Beck and Webb (2003) use a large dataset of 68 developed and developing countries over the 1961-2000 period, while Li et al. (2007) rely on a panel of 30 OECD countries over the 1993-2000 period. These studies use life premiums as the dependent variable (usually expressed as ratios to GDP or the population). Browne and Kim (2003) and Li et al. (2007) define the estimated equation as the demand for life insurance. Outreville (1996) specifies briefly the demand and supply for life insurance and defines the estimated equation as a reduced form. Beck and Webb (2003) also indicate that the life premium reflects both demand and supply factors, while stressing the difficulties of distinguishing between the demand and supply sides.

Table 1 summarizes the main variables that should drive the development of the life insurance sector. Most of these variables have been explored in empirical research. The variables are clustered into four major groups: (1) economic variables (income level, income inequality, inflation, and interest rates); (2) demographic variables (population, population density, age dependency, and life expectancy); (3) social/cultural variables (education, religion); and (4) institutional/market structure variables (size/scope of the social security system, the level of financial development, the share of state and foreign insurers, market concentration, and the quality of the legal/regulatory framework). The second column indicates the expected impact of the variable on the life insurance sector, and the four following columns show the signs of the estimated coefficients in recent studies: positive (+), negative (-), or not significant (N/S). Shaded cells indicate that the variable was not included in the particular study.

Income is a central variable in insurance-demand models. Higher income should increase the demand for life insurance for several reasons, including the greater affordability of life insurance products, the stronger need to safeguard the potential income and expected consumption of the dependents (which increase with the level of income) against the premature death of the wage earner, and the reduced unit price of larger life policies. As shown in Table 1, all recent studies conclude that income (measured by per capita income) has a positive and significant effect on life insurance premiums.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Income</td>
<td>Positive</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Income inequality</td>
<td>Ambiguous</td>
<td></td>
<td>N/S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inflation rate</td>
<td>Negative</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Real interest rate</td>
<td>Ambiguous</td>
<td></td>
<td>N/S</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Population</td>
<td>Positive</td>
<td></td>
<td>N/S</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Population density</td>
<td>Positive</td>
<td></td>
<td>N/S</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Age dependency ratio</td>
<td>Ambiguous</td>
<td>+</td>
<td>N/S</td>
<td>Young ratio: N/S</td>
<td>+</td>
</tr>
<tr>
<td>Life expectancy</td>
<td>Ambiguous</td>
<td>+</td>
<td>N/S</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Education</td>
<td>Positive</td>
<td>+</td>
<td>N/S</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Religion (muslim)</td>
<td>Negative</td>
<td>-</td>
<td>N/S</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Social security</td>
<td>Ambiguous</td>
<td>+</td>
<td>N/S</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Financial development</td>
<td>Positive</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Share of state insurers</td>
<td>Negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Share of foreign insurers</td>
<td>Positive</td>
<td>N/S</td>
<td>+</td>
<td>Mixed^1</td>
<td></td>
</tr>
<tr>
<td>Market concentration</td>
<td>Negative</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legal environment</td>
<td>Positive</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

Table 1: Main determinants of the life insurance sector: expected and actual results
Notes: 1) N/S = Not significant; 2) Share of rural population; 3) Negative for first term, positive for quadratic term
The expected impact of income inequality on life insurance demand, on the other hand, is ambiguous. While demand should generally increase with income, very wealthy groups may not need life insurance because they have surplus/excess assets, while the very poor do not have the means to buy life products. As argued by Beck and Webb (2003), life insurance should generally prove more attractive to the middle classes, but in lower-income countries life products may still be unaffordable to this group. In these cases, the presence of a wealthier upper class could drive the demand for insurance. In sum, life insurance demand would depend on the interactions between the level of income and the shape of the income distribution, which could make the effect of income inequality on insurance demand ambiguous. Beck and Webb (2003) find that higher inequality (measured by the Gini coefficient) does not have any significant effect on the demand for life insurance, while the other studies do not include this variable.\(^2\)

Inflation is expected to have a negative effect on the demand for life insurance, as it erodes the value of insurance policies and makes them less attractive. All the studies surveyed find a negative and significant relationship between inflation and the life insurance premiums, as shown in Table 1.\(^3\) On the other hand, the effect of higher real interest rates on life premiums is ambiguous, because the effects on supply and demand could prove complex. Beck and Webb (2003) argue that higher real interest rates would increase the investment return of providers, which would enable them to offer more attractive returns to consumers. However, the impact of the real interest rate on insurance premiums reflects more complex interactions between supply and demand, and the net effect is ultimately determined by the term-structure of interest rates, the composition of business lines, and the composition of insurers’ portfolios. For example, in a detailed study of the Chilean annuities market, Rocha and Thorburn (2007) and Rocha et al. (2008) show that an increase in real interest rates has a positive effect on real annuity rates, but an ambiguous effect on the number of new annuity policies and the annuity premium (a large component of the overall life insurance premium in Chile). Consequently, it is not surprising to find different results across different empirical studies. Browne and Kim (2003) neglect this variable, Outreville (1996) does not find a significant effect, Beck and Webb (2003) find a positive and significant effect, but Li et al. (2007) find exactly the opposite result.\(^4\)

The size of the population should have a positive effect on the demand for life insurance. For given levels of per capita income and other relevant variables, a larger population not only implies a larger clientele for insurance companies, but also larger risk pools, which reduces the risks for insurers and allows them to reduce fees per dollar of coverage. It is, therefore, rather surprising that this variable has not been included in most empirical research. Population density should also have a positive effect on life insurance, by reducing marketing and distribution costs and the price of insurance. This variable has also been generally neglected in empirical research. Outreville (1996) tests the effect of the share of the urban population, which should be correlated with population density, and finds that the effect is not significant. However, urbanization may be defined differently in different countries, and, for this reason, may not be a perfect proxy for population density. In addition, most research shows significant differences in income levels according to size of the city [Ferre (2010)].

The age dependency ratio (defined as the ratio of people under 15 and above 65 years of age over the working age population) is traditionally assumed to have a positive effect on life insurance demand, on the grounds that wage earners buy life insurance primarily to protect their dependents against mortality risk. However, Beck and Webb (2003) argue persuasively that the effect is rather ambiguous, because dependency ratios can have different effects across different business lines (mortality risk, savings, and annuities). Moreover, they also show the importance of breaking the age dependency ratio between the young and the old dependency ratios. As shown in Table 1, Browne and Kim (1993) and Li et al. (2007) find a positive and significant coefficient, and Beck and Webb (2003) also find a positive and significant coefficient, but only for the old dependency ratio. Outreville (1996) does not find significant effects, and Beck and Webb (2003) do not find significant effects for the young ratio. It is possible that these mixed results reflect different compositions of the business lines of insurers across the different samples.

\(^2\) The early study by Beenstock et al. (1986) finds a negative relationship between the Gini coefficient and life insurance penetration.

\(^3\) Babbel (1981) shows that even the introduction of price-indexed life products in Brazil was not able to increase the overall demand for life insurance.

\(^4\) Li et al. (2007) also point out that the result may depend on the particular interest rate selected.
Most researchers expect life expectancy to have a negative impact on the demand for life insurance, on the grounds that a longer life expectancy is associated with a lower probability of premature death and lower need for life insurance. However, Beck and Webb (2003) show that the effect of life expectancy on life insurance demand is ambiguous, considering the other business lines provided by insurers (savings for retirement, annuities). Consequently, it is not surprising to find very mixed results in the empirical literature. Browne and Kim (1993) and Beck and Webb (2003) do not find significant effects, Outreville (1996) finds a positive and significant coefficient, but Li et al. (2007) find exactly the opposite result. Again, this may be due to different compositions of the business lines of insurers across the different samples.

Education should increase the demand for life insurance, not only because it increases the level of awareness of the relevant risks and the degree of risk aversion, but also because it increases the period of dependency. Surprisingly, only Li et al. (2007) find a positive and significant effect. Religious/cultural factors could also influence the demand for life insurance. In this regard, empirical studies have focused primarily on Islamic countries, as life insurance is frequently disapproved of in some countries on the grounds that it is not compliant with Shariah law (insurance is considered by some to be a hedge against the will of God). Outreville (1996) does not find a significant effect, but Browne and Kim (2003) and Beck and Webb (2003) find a negative and significant coefficient for this variable. The low demand for life insurance in many Muslim countries has prompted the emergence of Takaful insurance, which is structured in compliance with Shariah law [Redzuan et al. (2009)].

Most researchers consider that social security schemes provide protection against mortality risk and should therefore affect life insurance demand negatively. However, Browne and Kim (1993) argue that the effect could be ambiguous, considering that social security also provides protection against disability and old age, and the provision of these benefits could produce a positive income effect on life insurance. It is, therefore, not surprising to find very mixed results across different studies, as shown in Table 1. This pattern could reflect the different composition of social security programs across the different countries in different samples. Ideally, the effect of social security should be tested by disaggregating social security expenditures, but this information is not easily available in many countries.

Financial development should have a positive effect on the life insurance sector, and this effect could operate both from the demand and supply sides. For example, commercial banks expanding into mortgages and other personal loans (e.g., cars) could require the purchase of life insurance to approve these loans. Likewise, the greater availability of private fixed-income instruments, allowing higher spreads for insurers, could motivate them to offer life policies at more attractive terms and increase sales of life products. It is not surprising, therefore, that most studies show a positive and significant effect of financial development, even after controlling for causality and endogeneity bias.

The structure of the insurance market could have significant effects on the growth of the market, but there have been few attempts to test these effects. For example, the presence of state insurers could stifle market development, but this factor was not tested in any of the existing studies. The presence of foreign insurers would be expected to contribute to market development through product innovation and marketing techniques, but has produced mixed results, as shown in Table 1. Regarding market structure, only Outreville (1996) tested the impact of oligopolistic markets on market development, finding a negative and significant effect.

Finally, the quality of the legal and regulatory environment could also have a significant effect on market development, inter alia by enhancing the credibility of insurance contracts. It is surprising that few studies have tested this type of factor, but this may have been due to the lack of good indicators when the studies were conducted. Beck and Webb (2003) use the rule of law index and find a positive and significant effect. Since then, other indicators have become available, enabling researchers to test this critical factor.

5 Zelitzer (1979) notes that, until the 19th century, European nations also condemned life insurance on religious grounds.

6 Rocha and Thorburn (2007) and Rocha et al. (2008) show that the increase in the participation of higher-yield corporate bonds in the portfolios of Chilean annuities providers had a positive impact on the real annuity rate.
In contrast to the literature on life insurance, there have been very few attempts to explore the determinants of the non-life sector. Nakata and Sawada (2007) test a semi-parametric model including per capita income, population, the Gini coefficient, financial development, and contract enforceability. The coefficients usually have the expected signs but only the contract enforceability variable is significant.

Data and empirical methodology
In this study we empirically explore the relationship between various measures of insurance activity and a set of potential drivers. We consider three main dependent variables that capture insurance industry development: 1) the ratio of gross life insurance premiums to GDP (LIFE); 2) the ratio of gross non-life insurance premiums to GDP (NON-LIFE), and 3) the ratio of total assets of insurance companies to GDP (ASSETS). The first two are insurance penetration variables that capture the extent of risk management, but life insurance premiums can also reflect a savings motive. The third variable captures the size of both the life and non-life sectors, but especially the former, as the life sector operates with longer-term contracts and needs to accumulate a larger volume of assets to honor future disbursements.

Regarding the potential drivers of the insurance sector, we obtained data for most of the variables listed in Table 1. These variables have been primarily used for empirical studies of the life sector, but several of them are in principle relevant for both the non-life sector. However, we also obtained economic variables that are more closely related to the non-life sector, such as the size of the fleet of cars and the volume of external trade. To examine the drivers of insurance assets, we combined the variables used for both the life and non-life sectors.

We use annual, country-level data for the period 2000-08 which cover about 90 countries. The insurance variables were mostly taken from the insurance data provider Axco, although asset data were also taken from national sources. The independent variables were obtained from a variety of sources, including Axco, the International Financial Statistics (IMF), the World Development Indicators (World Bank), and national sources. Table 2 shows the definitions, sources and descriptive statistics of the variables. Table 3 shows the correlations between the dependent and explanatory variables.

As shown in Table 2, there is substantial variation in the insurance variables. NON-LIFE, the most ubiquitous form of insurance, ranges from virtually 0% (i.e., Angola, Bangladesh, Myanmar) to over 3.5% of GDP (i.e., U.S., U.K.). LIFE, a form of insurance which emerges at more advanced stages of development, ranges from virtually 0% (i.e., Kazakhstan, Syria, Vietnam) to over 10% of GDP (i.e., South Africa, Portugal). ASSETS range from virtually 0 percent (i.e., Tanzania, Kenya, Saudi Arabia) to over 90% of GDP (i.e., Sweden, U.K., France).

In line with the previous section, we group our explanatory variables into four categories. First, we consider economic variables. Our main income level variable is GDP per capita (GDPPC). Our measure of income inequality is the fraction of income held by the richest 20% of the population (INCOMETOP20) and we use annual inflation (INFLATION) as a proxy for inflationary expectations. As mentioned before, we add two additional economic variables to examine the non-life sector. We include the number of passenger vehicles per 1,000 inhabitants (CARS) because most countries require at least third-party liability insurance (comprehensive car insurance is usually voluntary but also common in many countries). We also consider trade activity, the sum of import and export activities as a fraction of GDP (TRADE), since trade often relies on the availability of marine, cargo, and liability insurance.

Second, we explore the impact of demographic variables such as population size (POP), population density (POP_DENSITY), and average life expectancy (LIFE_EXP). The first two variables have been neglected in empirical research, but could prove relevant for both the life and non-life sector, especially the former. We also include the overall age dependency ratio (AGE_DEP), defined as the ratio of the non-working (<=15 and >=65 years) to the working population (>15 and <65 years). To disentangle the effects of a relatively young versus an old population, we also study the young (AGE) and old age dependency ratios, respectively.

---

7 Separately identified health insurance is not included in the premiums. While health insurance can be a significant component of total premium in some countries, research has shown that its penetration tends to be very weakly correlated to GDP per capita and to be largely driven by idiosyncratic factors.

8 As shown in Table 3, ASSET is more strongly correlated to LIFE than NON-LIFE.
### Panel A

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(L)NONLIFE</td>
<td>(Log) Gross volume of non-life premiums to GDP (%)</td>
<td>Axco</td>
</tr>
<tr>
<td>(L)LIFE</td>
<td>(Log) Gross volume of life premiums to GDP (%)</td>
<td>Axco</td>
</tr>
<tr>
<td>(L)INSURANCE_AS.</td>
<td>(Log) Insurance assets to GDP (%)</td>
<td>Axco</td>
</tr>
<tr>
<td><strong>Economic variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(L)GDPPC</td>
<td>(Log) GDP per capita (constant U.S.$)</td>
<td>WDI</td>
</tr>
<tr>
<td>(L)INCOMETOP20</td>
<td>(Log) Fraction of income held by richest quintile (%)</td>
<td>WDI</td>
</tr>
<tr>
<td><strong>Demographic variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(L)POP</td>
<td>(Log) Population size (million)</td>
<td>WDI</td>
</tr>
<tr>
<td>(L)LAGEDEP_OLD</td>
<td>(Log) Age dependency (%). Percentage population &gt;=65 yr of working population (15 yr and &lt; 65 yr).</td>
<td>WDI</td>
</tr>
<tr>
<td>(L)LAGEDEP_YNG</td>
<td>(Log) Age dependency (%). Percentage population &lt;=15 yr of working population (15 yr and &lt; 65 yr).</td>
<td>WDI</td>
</tr>
<tr>
<td>(L)LAGEDEP</td>
<td>(Log) Age dependency (%). Percentage population &lt;=15 yr and &gt;=65 yr of working population (15 yr and &lt; 65 yr).</td>
<td>WDI</td>
</tr>
<tr>
<td><strong>Social/cultural variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(L)SCHOOLING</td>
<td>(Log) Expected number of years of schooling (year)</td>
<td>WDI</td>
</tr>
<tr>
<td>MUSLIM</td>
<td>Dummy whether majority of population is Muslim</td>
<td>CIA Factbook</td>
</tr>
<tr>
<td><strong>Institutional/ market structure variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRIVATE</td>
<td>Dummy whether sector is mainly private</td>
<td>Various World Bank financial sector specialists</td>
</tr>
<tr>
<td>(L)CONTRIB_RATE</td>
<td>(Log) Contribution rate (%)</td>
<td>Axco</td>
</tr>
<tr>
<td>CONCENTR.</td>
<td>Proxy for market concentration. DUMMY that take on a value of 1 if the largest market player covers &gt;=50% of life-insurance industry assets and/or, the top 2 largest players cover &gt;=60% of life-insurance industry assets and/or, the top 3 cover &gt;=70% and/or, the top 5 cover &gt;=80%. The indicator is 0 otherwise.</td>
<td>Axco</td>
</tr>
<tr>
<td>(L)PC</td>
<td>(Log) Private credit to GDP (%)</td>
<td>IMF International Financial Statistics</td>
</tr>
<tr>
<td>(L)PR_BOND</td>
<td>Private domestic bond market capitalization to GDP (%)</td>
<td>Bank for International Settlements</td>
</tr>
<tr>
<td>LEGAL_RIGHTS</td>
<td>Creditor protection rights (1-10; higher is better rights)</td>
<td>Doing Business</td>
</tr>
<tr>
<td>(L)CARS</td>
<td>(Log) Number of passenger cars per 1,000 people</td>
<td>WDI</td>
</tr>
<tr>
<td>(L)TRADE</td>
<td>(Log) Sum imports and exports to GDP (%)</td>
<td>WDI</td>
</tr>
</tbody>
</table>

### Panel B (not in logs)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Country-year observations</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONLIFE</td>
<td>827</td>
<td>1.44</td>
<td>1.13</td>
<td>0.04</td>
<td>18.49</td>
</tr>
<tr>
<td>LIFE</td>
<td>809</td>
<td>1.82</td>
<td>2.70</td>
<td>0.00</td>
<td>30.17</td>
</tr>
<tr>
<td>GDPPC</td>
<td>801</td>
<td>19.69</td>
<td>24.78</td>
<td>0.02</td>
<td>100.84</td>
</tr>
<tr>
<td><strong>Economic variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFLATION</td>
<td>837</td>
<td>1.08</td>
<td>0.17</td>
<td>0.90</td>
<td>4.25</td>
</tr>
<tr>
<td>INCOMETOP20</td>
<td>702</td>
<td>45.56</td>
<td>7.52</td>
<td>33.07</td>
<td>62.94</td>
</tr>
<tr>
<td>CARS</td>
<td>393</td>
<td>244.89</td>
<td>190.02</td>
<td>1.00</td>
<td>632.00</td>
</tr>
<tr>
<td>TRADE</td>
<td>764</td>
<td>81.06</td>
<td>37.89</td>
<td>0.31</td>
<td>220.41</td>
</tr>
<tr>
<td><strong>Demographic variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POP</td>
<td>837</td>
<td>61.69</td>
<td>175.23</td>
<td>0.08</td>
<td>1325.64</td>
</tr>
<tr>
<td>POP_DENSITY</td>
<td>837</td>
<td>124.74</td>
<td>164.14</td>
<td>2.49</td>
<td>1229.16</td>
</tr>
<tr>
<td>LAGEDEP_OLD</td>
<td>810</td>
<td>13.82</td>
<td>7.88</td>
<td>1.26</td>
<td>32.85</td>
</tr>
<tr>
<td>LAGEDEP_YNG</td>
<td>810</td>
<td>42.02</td>
<td>18.91</td>
<td>19.36</td>
<td>97.76</td>
</tr>
<tr>
<td>LIFE_EXP</td>
<td>830</td>
<td>71.84</td>
<td>7.94</td>
<td>43.55</td>
<td>82.59</td>
</tr>
<tr>
<td><strong>Social/cultural variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCHOOLING</td>
<td>530</td>
<td>13.94</td>
<td>2.74</td>
<td>4.37</td>
<td>20.65</td>
</tr>
<tr>
<td>MUSLIM</td>
<td>756</td>
<td>0.26</td>
<td>0.44</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Institutional/ market structure variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRIVATE</td>
<td>783</td>
<td>0.80</td>
<td>0.40</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>CONTRIB_RATE</td>
<td>756</td>
<td>19.43</td>
<td>10.41</td>
<td>3.93</td>
<td>65.40</td>
</tr>
<tr>
<td>CONCENTR.</td>
<td>359</td>
<td>0.36</td>
<td>0.48</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PC</td>
<td>812</td>
<td>57.40</td>
<td>45.64</td>
<td>1.97</td>
<td>319.71</td>
</tr>
<tr>
<td>PR_BOND</td>
<td>360</td>
<td>33.57</td>
<td>43.24</td>
<td>0.00</td>
<td>374.07</td>
</tr>
<tr>
<td>LEGAL_RIGHTS</td>
<td>447</td>
<td>5.39</td>
<td>2.28</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

**Table 2 - Definitions and descriptive statistics**

The sample consists of country-level, annual data for the period 2000-08. (L)SCHOOLING, (L)LIFE_EXP, (L)INCOMETOP20, and (L)CONTRIB_RATE are sample period averages. WDI — World Bank World Development Indicators; Axco — Axco Insurance Information Services
What drives the development of the insurance sector?
An empirical analysis based on a panel of developed and developing countries

Third, we examine the effects of social and cultural variables. They include the number of schooling years (SCHOOLING) and an indicator variable that captures whether the country is predominantly Muslim (MUSLIM). These variables are probably more relevant for the life sector, as they primarily capture the motivations and incentives of individuals to look for insurance coverage. However, we tested the impact of MUSLIM in the non-life sector as well, as the cultural environment could also have an effect on the decision of businesses to look for insurance protection.

Finally, we investigate a set of institutional and market structure variables. We include the mandatory contribution rate for social security as a proxy for the size of the social security system (CONTRIBUTION_RATE). As mentioned in the previous section, a generous social security system could reduce the incentives and the need to buy retirement products from the life sector. We also include an indicator that measures the share of the private sector in total assets (PRIVATE) to test whether insurance flourishes when the private sector plays a dominant role in the industry. To assess the impact of market concentration (CONCENTRATION), we create a variable that assumes a value of 1 if the largest market player covers ≥ 50% of life-insurance industry assets and/or, the top 2 largest players cover ≥ 60% of life-insurance industry assets and/or, the top 3 cover ≥ 70% and/or, the top 5 cover ≥ 80%. The indicator is 0 otherwise. Finally, we study the impact of financial development, proxied both by the ratio of private credit to GDP (PRIVATE) and the ratio of private bonds to GDP (PRIVATE_BOND).

For our baseline results, we employ multivariate regression analyses on pooled country-year data because of some correlation between the variables. We drop very small economies from the sample (countries with GDP less than U.S.$5 billion constant 2000 dollars) because they tend to have idiosyncratic insurance industries due, for example, to their offshore nature. We adopt Ordinary Least Square (OLS) regressions since most of our explanatory variables are arguably exogenous. All OLS regressions are estimated with year-fixed effects to address worldwide trends and omitted variables such as the global cycle. To account for heterogeneity in our sample we always report Huber-White's heteroskedasticity-consistent standard errors. In addition, we report regression results with and without error clustering on the country level. Note that most of our regressions include time-invariant factors and we are not able to include country-fixed effects to address omitted variable bias. However, to mitigate this problem, we include a set of baseline variables in all the regressions – GDPPC, POP, POP_DENSITY, and INFLATION.

A reverse causality problem could be present for at least two variables: CARS and PC. Clearly, a developed non-life insurance sector makes car ownership more affordable and attractive, thus increasing the demand for cars. Similarly, a developed insurance sector can also mitigate loan collateral problems and reduce risks for lenders (mortality risk and credit risk), thus promoting private credit intermediation (Levine et al. (2000)).

To address the issue of reverse causality we adopt the system, Generalized Method of Moments (GMM) dynamic panel methodology à la Arellano and Bover (1995). In short, this methodology allows the insurance variables to depend on their own past realizations, adds country-fixed effects to the model to address omitted variable bias, and uses lagged values of explanatory variables as internal instruments to deal with reverse causality. For this, we need to assume that the explanatory variables can be affected by past and current realizations of insurance development, but are uncorrelated with the future error term (weak exogeneity). Appropriateness of the instruments can be formally tested with a Sargan or Hansen test of over-identifying restrictions with the null that instruments are valid. Since the methodology uses a system of equations – the first in levels, the second in first differences – to fully exploit the instruments and boost estimate precision, we also need to assume that the differences of explanatory variables are uncorrelated with the error term. Moreover, the methodology requires the assumption that the error term is not serially correlated, which can be tested by a correlation test with the null that the errors in the difference equation are uncorrelated. In addition to the lags of explanatory variables, we also use religious and legal origin indicator variables as instruments.

Lastly, it is important to recognize that our regressions results reflect average effects in our sample. For instance, although we find a negative coefficient for the contribution rate in our regressions, it does not imply that the association is positive in all countries. Rather, depending on its design, individual social security systems can act as a complement or a substitute for life insurance. Similarly, the impact of our drivers can be different between countries because they differ in the development of their business lines such as annuities and mortality-risk products.
Empirical results

This section discusses the empirical results on the drivers of insurance market development. The results consist of 1) parametric and non-parametric group comparison tests and 2) pooled OLS and dynamic panel regression results. The OLS results for each insurance variable are presented in individual sections.

Group comparison test results

Table 4 shows the results of univariate, two-group comparison tests to identify strong correlates with our three main insurance variables. The tests are constructed as follows. All country-year observations in the period 2000-08 for each of the three insurance variables are divided into two groups based on the median values of the explanatory variables described in Section 3.

For nominal variables in Table 4, we simply construct “above-the-period median” and “below-the-period median” groups. For binary variables the groups are determined according to whether the variable takes a value of one or zero. We then conduct simple T-tests (with unequal variance) and Wilcoxon-Mann-Whitney tests to assess whether the insurance variables differ significantly between the groups. The T-test tests for a difference between means of the two groups. The non-parametric Wilcoxon-Mann-Whitney test tests whether the two groups are drawn from the same population or distribution and is more robust in the presence of outliers.

As expected, the three insurance variables usually display high values in favorable income, demographic, social, and institutional/policy environments. In contrast, insurance activity is significantly lower in less benevolent environments. These initial results suggest that insurance market development is sensitive to the

<table>
<thead>
<tr>
<th></th>
<th>NONLIFE</th>
<th>LIFE</th>
<th>INSURANCE AS.</th>
<th>GDP/PC</th>
<th>INFLATION</th>
<th>INCOME TOP20</th>
<th>POP</th>
<th>POP DENSITY</th>
<th>LAGEDEP OLD</th>
<th>LAGEDEP</th>
<th>LAGEDEP_QLD</th>
<th>LAGEDEP_YNG</th>
<th>LAGEDEP</th>
<th>LIFE_EXP</th>
<th>SCHOOLING</th>
<th>MUSLIM</th>
<th>PRIVATE</th>
<th>CONTRIB RATE</th>
<th>CONCENTRATION</th>
<th>PC</th>
<th>PR_BOND</th>
<th>CARS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONLIFE</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIFE</td>
<td>0.63</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INSURANCE AS.</td>
<td>0.64</td>
<td>0.88</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDP/PC</td>
<td>0.67</td>
<td>0.56</td>
<td>0.74</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INFLATION</td>
<td>-0.17</td>
<td>-0.36</td>
<td>-0.24</td>
<td>-0.30</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCOME TOP20</td>
<td>-0.22</td>
<td>-0.19</td>
<td>-0.40</td>
<td>-0.39</td>
<td>0.21</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POP</td>
<td>-0.32</td>
<td>0.07</td>
<td>0.10</td>
<td>-0.37</td>
<td>0.03</td>
<td>0.16</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POP DENSITY</td>
<td>0.03</td>
<td>0.33</td>
<td>0.16</td>
<td>-0.06</td>
<td>-0.08</td>
<td>0.22</td>
<td>0.18</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAGEDEP OLD</td>
<td>0.59</td>
<td>0.49</td>
<td>0.53</td>
<td>0.52</td>
<td>-0.14</td>
<td>-0.63</td>
<td>0.06</td>
<td>0.08</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAGEDEP_QLD</td>
<td>-0.56</td>
<td>-0.43</td>
<td>-0.50</td>
<td>0.19</td>
<td>0.66</td>
<td>0.15</td>
<td>-0.09</td>
<td>-0.68</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAGEDEP_YNG</td>
<td>-0.24</td>
<td>-0.19</td>
<td>-0.23</td>
<td>0.14</td>
<td>0.45</td>
<td>0.14</td>
<td>-0.02</td>
<td>-0.13</td>
<td>0.76</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIFE_EXP</td>
<td>0.47</td>
<td>0.42</td>
<td>0.48</td>
<td>0.76</td>
<td>-0.35</td>
<td>-0.40</td>
<td>-0.23</td>
<td>0.08</td>
<td>0.49</td>
<td>-0.68</td>
<td>-0.59</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCHOOLING</td>
<td>0.65</td>
<td>0.42</td>
<td>0.56</td>
<td>0.75</td>
<td>-0.23</td>
<td>-0.39</td>
<td>-0.30</td>
<td>-0.22</td>
<td>0.65</td>
<td>-0.69</td>
<td>-0.37</td>
<td>0.73</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUSLIM</td>
<td>-0.58</td>
<td>-0.51</td>
<td>-0.42</td>
<td>-0.39</td>
<td>0.03</td>
<td>0.00</td>
<td>0.09</td>
<td>0.04</td>
<td>0.63</td>
<td>0.43</td>
<td>-0.11</td>
<td>-0.29</td>
<td>-0.49</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRIVATE</td>
<td>0.39</td>
<td>0.44</td>
<td>0.29</td>
<td>0.29</td>
<td>-0.21</td>
<td>-0.05</td>
<td>-0.11</td>
<td>0.10</td>
<td>0.21</td>
<td>-0.24</td>
<td>-0.11</td>
<td>0.20</td>
<td>0.23</td>
<td>-0.16</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONTRIB RATE</td>
<td>0.23</td>
<td>0.16</td>
<td>0.18</td>
<td>0.17</td>
<td>-0.07</td>
<td>-0.26</td>
<td>0.02</td>
<td>-0.04</td>
<td>0.35</td>
<td>-0.38</td>
<td>-0.20</td>
<td>0.25</td>
<td>0.35</td>
<td>-0.13</td>
<td>-0.01</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONCENTRATION</td>
<td>-0.32</td>
<td>-0.35</td>
<td>-0.27</td>
<td>-0.25</td>
<td>0.18</td>
<td>-0.04</td>
<td>-0.22</td>
<td>0.17</td>
<td>-0.19</td>
<td>0.12</td>
<td>0.00</td>
<td>-0.14</td>
<td>-0.27</td>
<td>0.09</td>
<td>-0.32</td>
<td>-0.03</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC</td>
<td>0.63</td>
<td>0.75</td>
<td>0.75</td>
<td>0.68</td>
<td>-0.39</td>
<td>0.33</td>
<td>0.11</td>
<td>0.21</td>
<td>0.48</td>
<td>-0.59</td>
<td>-0.40</td>
<td>0.63</td>
<td>0.58</td>
<td>-0.30</td>
<td>0.38</td>
<td>0.17</td>
<td>-0.22</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PR_BOND</td>
<td>0.59</td>
<td>0.57</td>
<td>0.70</td>
<td>0.64</td>
<td>-0.57</td>
<td>-0.41</td>
<td>0.25</td>
<td>0.09</td>
<td>0.54</td>
<td>0.51</td>
<td>-0.22</td>
<td>0.46</td>
<td>0.67</td>
<td>-0.34</td>
<td>0.29</td>
<td>0.15</td>
<td>-0.10</td>
<td>0.65</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CARS</td>
<td>0.68</td>
<td>0.45</td>
<td>0.55</td>
<td>0.86</td>
<td>-0.22</td>
<td>-0.52</td>
<td>0.43</td>
<td>0.10</td>
<td>0.67</td>
<td>-0.77</td>
<td>-0.48</td>
<td>0.70</td>
<td>0.83</td>
<td>-0.39</td>
<td>0.40</td>
<td>0.35</td>
<td>-0.16</td>
<td>0.58</td>
<td>0.59</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRADE</td>
<td>0.38</td>
<td>0.17</td>
<td>0.05</td>
<td>0.15</td>
<td>-0.03</td>
<td>-0.26</td>
<td>0.42</td>
<td>0.09</td>
<td>0.10</td>
<td>-0.25</td>
<td>-0.21</td>
<td>0.18</td>
<td>0.25</td>
<td>-0.05</td>
<td>0.12</td>
<td>0.14</td>
<td>0.04</td>
<td>0.25</td>
<td>0.20</td>
<td>0.31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 - Correlations

The sample consists of country-level, annual data for the period 2000-08. All variables are expressed as natural logarithms (with the exception of dummy variables). LSCHOOLING, LLIFE_EXP, LINCOME TOP20, and LCONTRIB_RATE are sample period averages.
existence of proper conditions, including those that are within the reach of policy-makers. Below we will discuss a few highlights and relegate a comprehensive discussion of all factors to the regression results.

First, we discuss economic variables. We find that the three insurance variables are all significantly higher in the high income group, as expected. For example, the average non-life premium is 1.9% of GDP in the high income group compared to 0.9% of GDP in the low group. The differences are even more pronounced in the case of the life sector, suggesting that this sector starts growing at a much faster speed in high middle-income countries. Not surprisingly, the same sharp differences are observed in the case of assets – the average ratio of assets to GDP is 32.1% in the high group compared to 4.9% in the low group, suggesting again that life insurance companies only start developing rapidly in high middle-income countries (as the volume of assets derives predominantly from the life sector). Finally, the tests also suggest that the fleet of cars is a relevant driver of the non-life sector. For example, the average non-life premium is 1.1% of GDP in the group with a low ratio of cars to the population, versus 1.6% of GDP in the high group. The non-life sector also seems to be positively and significantly affected by the volume of external trade.
We also find that insurance activities are significantly hampered in high inflation countries, especially in the life sector. This is not surprising, as the value of life policies is significantly eroded by high inflation, triggering a contraction in demand. For example, the average life premium in the low inflation group is 2.7% of GDP compared to 0.9% in the high inflation group. Similarly, the average ratio of assets to GDP is 30.7% in the low inflation group compared to only 7.1% in the high group. The tests for the inequality variable show that insurance activity tends to flourish in more equal societies (i.e., lower INCOMETOP20), which may reflect the presence of a broader middle class driving demand, particularly for life insurance. However, this result should be seen as preliminary, because of the correlation of this variable with other variables such as GDP per capita and schooling.

Next, we turn to the demographic variables. Life premiums are larger in countries with larger populations, as expected. Intriguingly, the opposite is true in the case of the non-life sector. This may be capturing the presence of small and open economies with large non-life businesses due to external trade.9 The ratio of assets to GDP is not significantly different in the two groups,

---

9 The correlation coefficient of POP and TRADE is negative and significant, as shown in Table 3.
What drives the development of the insurance sector?  
An empirical analysis based on a panel of developed and developing countries

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGDPPC</td>
<td>1.06</td>
<td>0.96</td>
<td>0.92</td>
<td>0.61</td>
<td>0.37</td>
<td>0.29</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>(16.9)***</td>
<td>(14.4)***</td>
<td>(12.8)***</td>
<td>(12.5)***</td>
<td>(9.03)***</td>
<td>(4.99)***</td>
<td>(3.63)***</td>
</tr>
<tr>
<td></td>
<td>[6.04]***</td>
<td>[5.32]***</td>
<td>[5.56]***</td>
<td>[6.52]***</td>
<td>[3.31]***</td>
<td>(2.07)**</td>
<td>(2.03)**</td>
</tr>
<tr>
<td>LPOP</td>
<td>0.32</td>
<td>0.32</td>
<td>0.24</td>
<td>0.21</td>
<td>0.31</td>
<td>0.12</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>(11.9)***</td>
<td>(12.6)***</td>
<td>(7.32)***</td>
<td>(6.11)***</td>
<td>(12.0)***</td>
<td>(3.30)***</td>
<td>(5.82)***</td>
</tr>
<tr>
<td></td>
<td>[4.21]***</td>
<td>[4.65]***</td>
<td>[2.94]***</td>
<td>[3.02]***</td>
<td>[4.63]***</td>
<td>[1.15]</td>
<td>[3.23]***</td>
</tr>
<tr>
<td></td>
<td>0.48</td>
<td>0.45</td>
<td>0.39</td>
<td>0.38</td>
<td>0.25</td>
<td>0.27</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>(3.78)***</td>
<td>(3.98)***</td>
<td>(3.24)***</td>
<td>(3.94)***</td>
<td>(6.19)***</td>
<td>(8.81)***</td>
<td>(6.05)***</td>
</tr>
<tr>
<td></td>
<td>[2.90]</td>
<td>[3.20]</td>
<td>[6.10]</td>
<td>[8.89]</td>
<td>[0.79]</td>
<td>[5.81]</td>
<td>[8.15]</td>
</tr>
<tr>
<td>INFLATION</td>
<td>-2.90</td>
<td>-3.20</td>
<td>-6.10</td>
<td>-8.89</td>
<td>-0.79</td>
<td>-5.81</td>
<td>-8.15</td>
</tr>
<tr>
<td></td>
<td>(-5.84)***</td>
<td>(-5.96)***</td>
<td>(-5.23)***</td>
<td>(-5.81)***</td>
<td>(-1.23)</td>
<td>(-3.24)***</td>
<td>(-5.71)***</td>
</tr>
<tr>
<td></td>
<td>[-3.35]***</td>
<td>[-3.51]***</td>
<td>[-5.35]***</td>
<td>[-4.15]***</td>
<td>[-0.66]</td>
<td>[-1.64]</td>
<td>[-5.00]***</td>
</tr>
<tr>
<td>LLIFE_EXP</td>
<td>-3.00</td>
<td>-3.71</td>
<td>-0.14</td>
<td>-1.18</td>
<td>-1.64</td>
<td>-0.11</td>
<td>-0.13</td>
</tr>
<tr>
<td></td>
<td>(-3.33)***</td>
<td>(-4.62)***</td>
<td>[-1.18]</td>
<td>[-1.64]</td>
<td>[-0.11]</td>
<td>[-0.13]</td>
<td></td>
</tr>
<tr>
<td>LAGEDEP</td>
<td>0.17</td>
<td>0.55</td>
<td>0.94</td>
<td>0.53</td>
<td>0.37</td>
<td>0.072</td>
<td>0.084</td>
</tr>
<tr>
<td></td>
<td>(-1.40)***</td>
<td>(-3.76)***</td>
<td>(3.11)***</td>
<td>(1.20)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAGEDEP_OLD</td>
<td>0.58</td>
<td>0.68</td>
<td>1.19</td>
<td>0.63</td>
<td>0.66</td>
<td>0.56</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>(-1.48)</td>
<td>(0.68)</td>
<td>1.79</td>
<td>(3.61)***</td>
<td>(1.56)</td>
<td>(2.12)**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.87]</td>
<td>(1.01)</td>
<td>0.36</td>
<td>(3.69)***</td>
<td>(1.56)</td>
<td>(2.12)**</td>
<td></td>
</tr>
<tr>
<td>MUSLIM</td>
<td>-0.66</td>
<td>-0.66</td>
<td>-0.66</td>
<td>-0.66</td>
<td>-0.66</td>
<td>-0.66</td>
<td>-0.66</td>
</tr>
<tr>
<td></td>
<td>(-3.66)***</td>
<td>(-3.66)***</td>
<td>[-1.64]</td>
<td>[-1.64]</td>
<td>[-1.64]</td>
<td>[-1.64]</td>
<td></td>
</tr>
<tr>
<td>LCONTRIB_RATE</td>
<td>-0.15</td>
<td>-0.15</td>
<td>-0.15</td>
<td>-0.15</td>
<td>-0.15</td>
<td>-0.15</td>
<td>-0.15</td>
</tr>
<tr>
<td></td>
<td>(-1.71)*</td>
<td>[-1.71]</td>
<td>[-1.71]</td>
<td>[-1.71]</td>
<td>[-1.71]</td>
<td>[-1.71]</td>
<td></td>
</tr>
<tr>
<td>CONCENTR.</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>(1.88)*</td>
<td>(1.88)*</td>
<td>(1.88)*</td>
<td>(1.88)*</td>
<td>(1.88)*</td>
<td>(1.88)*</td>
<td>(1.88)*</td>
</tr>
<tr>
<td>PRIVATE</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>(4.00)***</td>
<td>(4.00)***</td>
<td>(4.00)***</td>
<td>(4.00)***</td>
<td>(4.00)***</td>
<td>(4.00)***</td>
<td>(4.00)***</td>
</tr>
<tr>
<td>LEGAL_RIGHTS</td>
<td>0.085</td>
<td>0.085</td>
<td>0.085</td>
<td>0.085</td>
<td>0.085</td>
<td>0.085</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td>(2.09)**</td>
<td>(2.09)**</td>
<td>(2.09)**</td>
<td>(2.09)**</td>
<td>(2.09)**</td>
<td>(2.09)**</td>
<td>(2.09)**</td>
</tr>
<tr>
<td>LPC</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>(14.1)***</td>
<td>(14.1)***</td>
<td>(14.1)***</td>
<td>(14.1)***</td>
<td>(14.1)***</td>
<td>(14.1)***</td>
<td>(14.1)***</td>
</tr>
<tr>
<td></td>
<td>[4.52]***</td>
<td>[4.52]***</td>
<td>[4.52]***</td>
<td>[4.52]***</td>
<td>[4.52]***</td>
<td>[4.52]***</td>
<td>[4.52]***</td>
</tr>
<tr>
<td>LPR_BOND</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>(5.50)***</td>
<td>(5.50)***</td>
<td>(5.50)***</td>
<td>(5.50)***</td>
<td>(5.50)***</td>
<td>(5.50)***</td>
<td>(5.50)***</td>
</tr>
</tbody>
</table>

Table 5 - Development drivers of the life insurance industry

The dependent variable is the logarithm of the volume of life insurance premiums as a fraction of GDP. The sample consists of annual data for the period 2000-08. All regressions are estimated using pooled Ordinary Least Squares. Year-fixed effects and a constant were estimated but are not reported. LSCHOOLING, LLIFE_EXP, LINCOMETOP20, and LCONTRIB_RATE are sample period averages. In all regressions, White's heteroskedasticity-consistent (clustered) t-values are given in parentheses (brackets). ***, **, and * represent significance at 1, 5, and 10% level respectively.
reflecting these conflicting results. On the other hand, higher population density tends to facilitate insurance activity in the two sectors, as indicated by larger premiums and assets. The impact of a higher age dependency ratio is positive in the life sector, and this is also reflected in the ratio of assets to GDP, although the Wilcoxon tests are not significant. Higher life expectancies have a positive impact on the two sectors, but this result needs to be interpreted with care due to the strong correlation between this variable and GDP per capita (Table 3).

The tests suggest social and cultural factors also matter. Countries with a higher level of schooling tend to have larger premiums, especially in the case of the life sector. Not surprisingly, this is also reflected in a much higher average volume of assets. However, this result also needs to be interpreted with care, due to the strong correlation between schooling and GDP per capita (Table 3). Countries with predominant Muslim populations have significantly lower life premiums and assets, indicating religion plays a role in life insurance markets. The same result holds for the non-life sector.

Institutional and market factors matter as well. Countries with larger social security schemes (reflected in high mandatory contribution rates) have more insurance development, although this result is somewhat surprising, as social security can reduce the space and incentives for private life insurance and private pensions. On the other hand, the tests suggest that countries with private sector dominance (i.e., PRIVATE=1) in the insurance sector exhibit significantly more insurance activity, both in the non-life and the life sectors. We also find strong results across all insurance indicators for market concentration. For example, in concentrated markets (i.e., CONCENTRATION=1), insurance assets are 8.8% of GDP, versus 25.4% in non-concentrated markets. Although our proxy for legal rights (LEGAL_RIGHTS) is imperfect for insurance, since it is tailored to creditor rights, the differences between the groups are large and statistically significant. For example, life insurance premiums to GDP are almost twice as large in the high creditor-rights group (2.03% versus 1.14%). However, this result could also partly be driven by the fact that financial development is key to insurance sector development. Indeed, we also find that financially more developed countries exhibit significantly more insurance activity. For example, the ratio of insurance assets to GDP is only 3.5% in the group with low ratios of credit to GDP, versus 31.5% in the group with developed credit markets. However, the results are weaker, and even puzzling, for the private bond variable.

These tests produce suggestive evidence and provide insights into how conditions affect insurance activities, but they could be driven by confounding factors. In particular, the test results could be driven by the fact that some of the grouping variables are correlated with general economic development, thus leading to erroneous conclusions about environmental factors that drive insurance market development. To address these concerns we turn to regression analyses next.

Life premium regression results
We consecutively present the regression results for each of the three dependent insurance variables. The regressions aim to explore the robustness of the group comparison tests by assessing the individual and joint significance of the factors identified in section 3: economic, demographic, social/cultural, and institutional/market structure factors. Given their overall structural relevance for the sector, we include a core set of both economic and demographic variables in all regressions as baseline controls: economic development, population size and density, and inflation (LGDP, LPOP, LPOP_DENSITY, INFLATION). We employ pooled OLS regressions on annual country-level data for the period 2000-08. We substitute sample averages for each year in the dataset for factors for which only a few observations per country are available: LLIFE, LINCOME20, and LCONTRIB_RATE. Thus, these variables are time-invariant. We report robust standard errors in two variations: unclustered and clustered at the country level. Year-fixed effects and a constant were estimated but are not reported.

Table 5 presents the results for life insurance. The dependent variable is log of life insurance premiums to GDP (LLIFE). Depending on the explanatory variables included in the models, the sample size varies from 181 (56 countries) to 767 (86 countries) annual country-level observations.

Model (1) explores the demographic variables, life expectancy and age dependency (LLIFE, LAGEDEP) in addition to the baseline variables. LGDP is positively and significantly associated with LLIFE at the 1% confidence level, in line with the findings of previous research. Similarly, we also find that LPOP and LPOP_DENSITY are positively associated with LLIFE at the 1% confidence level. These results persist virtually throughout all specifications, also when standard errors are clustered. As mentioned before, these results are expected, reflecting the positive effects of larger clienteles, deeper risk pools and scale
What drives the development of the insurance sector?  
An empirical analysis based on a panel of developed and developing countries

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGDPPC</td>
<td>0.14</td>
<td>0.26</td>
<td>0.30</td>
<td>0.20</td>
<td>-0.056</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>(3.48)***</td>
<td>(18.4)***</td>
<td>(14.6)***</td>
<td>(10.1)***</td>
<td>(-0.46)</td>
<td>(7.84)***</td>
</tr>
<tr>
<td></td>
<td>[1.81]***</td>
<td>[6.61]***</td>
<td>[7.26]***</td>
<td>[3.59]***</td>
<td>[-0.38]</td>
<td>[3.91]***</td>
</tr>
<tr>
<td>LPOP</td>
<td>-0.0017</td>
<td>-0.035</td>
<td>-0.073</td>
<td>-0.034</td>
<td>-0.020</td>
<td>-0.069</td>
</tr>
<tr>
<td></td>
<td>(-0.081)</td>
<td>(-3.40)***</td>
<td>(-5.38)***</td>
<td>(-2.94)***</td>
<td>(-6.61)</td>
<td>(-5.17)***</td>
</tr>
<tr>
<td></td>
<td>[-0.041]</td>
<td>[-1.22]</td>
<td>[-2.65]***</td>
<td>[-1.05]</td>
<td>[-0.47]</td>
<td>[-2.61]***</td>
</tr>
<tr>
<td>LPOP_DENSITY</td>
<td>0.091</td>
<td>0.0042</td>
<td>-0.022</td>
<td>0.0015</td>
<td>0.031</td>
<td>-0.046</td>
</tr>
<tr>
<td></td>
<td>(2.81)***</td>
<td>(0.32)</td>
<td>(-1.20)</td>
<td>(0.11)</td>
<td>(1.01)</td>
<td>(-2.64)***</td>
</tr>
<tr>
<td></td>
<td>[1.90]***</td>
<td>[0.13]</td>
<td>[-0.60]</td>
<td>[0.039]</td>
<td>[0.78]</td>
<td>[-1.33]</td>
</tr>
<tr>
<td>INFLATION</td>
<td>0.059</td>
<td>0.022</td>
<td>0.98</td>
<td>1.23</td>
<td>1.32</td>
<td>1.23</td>
</tr>
<tr>
<td></td>
<td>(0.14)</td>
<td>(1.41)</td>
<td>(2.19)***</td>
<td>(5.32)***</td>
<td>(2.52)***</td>
<td>(3.11)***</td>
</tr>
<tr>
<td></td>
<td>[0.098]</td>
<td>[0.85]</td>
<td>[1.22]</td>
<td>[4.03]***</td>
<td>[1.91]***</td>
<td>[1.71]***</td>
</tr>
<tr>
<td>MUSLIM</td>
<td>-0.56</td>
<td>-0.57</td>
<td>-0.57</td>
<td>-0.28</td>
<td>-0.11</td>
<td>-0.23</td>
</tr>
<tr>
<td></td>
<td>(-11.3)***</td>
<td>(-4.47)***</td>
<td>(-4.53)***</td>
<td>(-3.57)***</td>
<td>(-2.23)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-4.02]***</td>
<td>[-4.15]</td>
<td>[-2.06]**</td>
<td>[-1.92]***</td>
<td>[-1.59]</td>
<td>[-0.97]</td>
</tr>
<tr>
<td>CONCNR.</td>
<td>-0.060</td>
<td>-0.23</td>
<td>-0.11</td>
<td>-0.11</td>
<td>-0.11</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td>(-1.15)</td>
<td>(-2.06)**</td>
<td>(-1.92)***</td>
<td>(-1.59)</td>
<td>(-0.97)</td>
<td>(-0.97)</td>
</tr>
<tr>
<td>PRIVATE</td>
<td>0.15</td>
<td>-0.091</td>
<td>0.12</td>
<td>-0.97**</td>
<td>-0.97**</td>
<td>-0.97**</td>
</tr>
<tr>
<td></td>
<td>(2.64)***</td>
<td>[-0.75]</td>
<td>(1.97)***</td>
<td>(1.33)</td>
<td>(0.62)</td>
<td>(0.99)</td>
</tr>
<tr>
<td>LEGAL_RIGHTS</td>
<td>0.055</td>
<td>0.048</td>
<td>0.041</td>
<td>0.041</td>
<td>0.041</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>(5.53)***</td>
<td>(2.17)***</td>
<td>(4.20)***</td>
<td>(2.76)***</td>
<td>(1.50)</td>
<td>(2.09)***</td>
</tr>
<tr>
<td>LPC</td>
<td>0.30</td>
<td>0.14</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>(10.2)***</td>
<td>(1.60)</td>
<td>(4.80)***</td>
<td>(4.02)***</td>
<td>(1.09)</td>
<td>(2.35)***</td>
</tr>
<tr>
<td>LCARS</td>
<td>0.19</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>(4.62)***</td>
<td>(2.44)***</td>
<td>(2.44)***</td>
<td>(2.44)***</td>
<td>(2.04)***</td>
<td>(2.04)***</td>
</tr>
<tr>
<td>LTRADE</td>
<td>0.057</td>
<td>0.057</td>
<td>0.057</td>
<td>0.057</td>
<td>0.057</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>(1.02)</td>
<td>(0.54)</td>
<td>(0.54)</td>
<td>(0.54)</td>
<td>(0.54)</td>
<td>(0.54)</td>
</tr>
<tr>
<td>Observations</td>
<td>384</td>
<td>740</td>
<td>321</td>
<td>766</td>
<td>89</td>
<td>308</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.509</td>
<td>0.604</td>
<td>0.668</td>
<td>0.535</td>
<td>0.711</td>
<td>0.708</td>
</tr>
<tr>
<td>Countries</td>
<td>83</td>
<td>83</td>
<td>76</td>
<td>87</td>
<td>49</td>
<td>73</td>
</tr>
</tbody>
</table>

Table 6 - Development drivers of the non-life insurance industry

The dependent variable is the logarithm of the volume of non-life insurance premiums as a fraction of GDP. The sample consists of annual data for the period 2000-08. All regressions are estimated using pooled Ordinary Least Squares. Year-fixed effects and a constant were estimated but are not reported. LSCHOOLING, LLIFE_EXP, LINCOME TOP20, and LCONTRIB_RATE are sample period averages. In all regressions, White's heteroskedasticity-consistent (clustered) t-values are given in parentheses (brackets). ***, **, and * represent significance at 1, 5, and 10% level respectively.

economies, as well as good distribution opportunities. Indeed, it is surprising that these variables were not explored in previous research.

INFLATION has a coefficient of -2.90, with significance at the 1% confidence level, which clearly indicates that inflationary environments are detrimental to life insurance development since they inhibit effective asset management, make product pricing more difficult, and produce uncertainty around the long-term value of the insurance. Again, these results continue to hold across most specifications and confirm previous research findings. LLIFE_EXP has a negative impact on LLIFE, suggesting that a lower probability of premature death is a dominant factor and lowers the demand for life insurance. Finally, AGEDEP, the general age dependency ratio does not enter significantly, a result that is not surprising given the different effects of age dependency on different business lines and the possible different effects of the young and old dependency ratios.

In order to examine further the effects of age dependency, Model 2 splits up AGEDEP into the young and old dependency ratios (AGEDEP_YNG, AGEDEP_OLD), as there is a negative correlation between these ratios (rho=-0.68) that could render LAGEDEP insignificant. In fact, the correlation between LAGEDEP and
large risk pools and scale economies may be critical to the sector’s development. At the same time, Model 4 also shows that private sector dominance of the sector (PRIVATE) is positively associated with LLIFE on the 1-percent confidence level. This result remains when standard errors are clustered at the country level, and suggests that private sector dominance is associated with more product innovation and more effective distribution, which are also critical to the sector’s development. Finally, the legal rights index from Doing Business (LEGAL_RIGHTS) is positive and very significant (although in an unclustered regression only), suggesting that the life sector develops faster under a supportive legal framework.11

Models 5 and 6 show that financial development is strongly associated with LLIFE even when LGDPPC is controlled for. A 1% increase in the log of private credit to GDP (LPC) is associated with a 1.12% increase of LIFE. This finding implies that deeper credit markets spur personal loans, which often require life insurance as collateral. Additionally, financial development in general can help insurance companies conduct better risk and asset and liability management, which can increase the supply of insurance. Interestingly, the inclusion of LPC renders INFLATION insignificant in Model 5. This could be driven by the fact that inflation is detrimental to financial development. Indeed, INFLATION and LPC are negatively correlated (rho=−0.4). Together, these findings suggest that inflation exerts influence over life insurance activity in part via its strong impact on credit market development. In addition, Model 6 suggests that deeper private bond markets help insurance companies manage their assets more efficiently. Note, however, that the number of countries in the sample falls significantly since bond market data are typically only available for more developed countries.

Lastly, Model 7 displays regression results that test for the statistical significance of individual factors in all categories, to address a potential confounding variable problem. Although the sample size decreases significantly, the regression shows that the main results continue to hold with coefficients of the same order of magnitude. Interestingly, LINCOMETOP20 and MUSLIM have now become statistically significant in clustered regressions as well.

10 In unreported results, we included an LGDPPC squared term which is positive and highly significant. This result indeed suggests that life insurance consumption increases exponentially for higher levels of income and in that sense can be considered mostly a luxury good.

11 Admittedly, none of the legal indicators from Doing Business captures with accuracy the quality of the legal framework in the insurance sector. However, it can be argued that countries that have stronger creditor rights also tend to have a more supportive legal and regulatory framework in the insurance sector.
Non-life premium regression results
Next, we turn to a regression analysis of non-life insurance development in Table 6. The dependent variable is the log of non-life premiums to GDP (LNONLIFE). Depending on the explanatory variables included in the models, the sample size varies from 89 (49 countries) to 766 (87 countries) annual country-level observations.

Model 1 shows the result for the baseline controls in addition to two non-life market-specific economic factors: personal cars penetration and trade activities (LCARS, LTRADE). As expected, LGDPCHC is positively and statistically significantly associated with LNONLIFE at the 1% confidence level. The coefficient implies that a 1% increase in GDPCHC is associated with a 0.14% increase of NONLIFE, which is significantly lower than for LIFE. Interestingly, LPOP is not significant in this model, and is negative and significant in most other regressions. As discussed in the previous section, an explanation for this finding could be that smaller countries are more dependent on trade, which increases demand for marine, cargo, and liability insurance. This could also explain why LTRADE does not enter significantly. Indeed, LTRADE is significant and positively associated with LNONLIFE when LPOP is not controlled for (unreported result: the coefficient is 0.145 and P-value is 0.092). This finding further supports that LPOP (negatively) proxies for relatively higher trade activities (the correlation coefficient is -0.42, as shown in Table 3). We also find that LPOP_DENSITY is statistically significant and positive in this regression, but this result is not robust in all specifications.

Interestingly, INFLATION is positive in all models. It is not significant in Model 1, but it is highly significant in most other models. In theory, this result could reflect the portfolio shifts from financial to real assets and the anticipation of consumption in very high inflationary environments, resulting in additional demand for non-life insurance. However, this is not a likely explanation in our sample. Indeed, in unreported regressions we find that the inclusion of a high-inflation dummy (i.e., an indicator variable that assumes the value of 1 if annual inflation is larger than 10% and 0 otherwise) does not render INFLATION insignificant. Another explanation could be that inflationary asset bubbles boost insurable economic activity, particularly in the real estate sector. We also confirm that LCARS is positive and highly statistically significant. The coefficient implies that a 1% increase in CARS is associated with a 0.19% increase of NONLIFE.

Model 2 shows the results for a relevant societal factor: religion (MUSLIM). The regression confirms that MUSLIM is negatively associated with non-life activities as well. Indeed, the coefficient is statistically highly significant and suggests that NONLIFE levels of Muslim countries are at 57% of the expected value of non-Muslim countries, all else equal (exp(-0.56)). Moreover, clustering does not influence the significance of this result.

Model 3 covers institutional and market factors: market concentration, private sector dominance, and contract protection (CONCENTR, PRIVATE, LEGAL_RIGHTS). We find that CONCENTR is not significant for non-life premiums, in contrast to life, although it is weakly significant and negative in Model (6). We also document that private sector dominance (PRIVATE) is positively associated with LNONLIFE on the 1% confidence level, although the result does not withstand clustering. The coefficients suggest non-life insurance levels of countries with private sector dominance are at 116% of the expected value of countries that do not, all else being equal (exp(0.15)). Finally, we confirm that our measure of the quality of the legal framework (LEGAL_RIGHTS) is highly significant for non-life activity as well.

Model 4 shows the result of another relevant market factor: financial development (LPC). We find that credit market development (LPC) is also strongly associated with LNONLIFE. Country level cluster does not affect these results. Indeed, personal loans become available as credit markets develop and are used to acquire goods that need insurance, including cars and houses.

Finally, Models 5 and 6 put all relevant factors together. Model 6 replicates Model 5 but drops LCARS, since it reduces the sample size significantly. The MUSLIM, LEGAL_RIGHTS, and LPC results continue to be very strong, while the somewhat weaker PRIVATE result also remains. Interestingly, market concentration (CONCENTR.) is now also slightly significant and negative, which supports research finding that, after a certain optimal point, a concentrated market structure does impede non-life development [Fenn et al. (2008)]. Overall, these results suggest that the non-life sector is driven positively by income, private sector dominance, a strong legal framework, and a deep credit market. Population size has a negative effect, but this could be capturing the positive effect of trade volumes on the non-life sector. A relatively large fleet of cars also contributes to the sector’s...
development. Surprisingly, inflation seems to contribute positively to the development of the non-life sector, but it is dominated by the negative effect on the life sector, and could reflect a shift from financial to real assets in inflationary environments.

**Insurance assets regression results**

Next, we conduct a regression analysis on the overall size of the insurance sector measured by total assets. The results are presented in Tables 7 and 8. The dependent variable is the log of insurance sector total assets to GDP (LASSETS). Overall, the result for ASSETS resembles those for LIFE, since the life sector accumulates a larger pool of assets compared to the non-life sector. In particular, the baseline variables are significant in most of the regressions and with the same signs as those obtained for LIFE. In other words, the results generally confirm the importance of income levels, population, population density, and inflation.

### Table 7 - Development drivers of the insurance industry's total assets

The dependent variable is the logarithm of total insurance assets as a fraction of GDP. The sample consists of annual data for the period 2000-08. All regressions are estimated using pooled Ordinary Least Squares. Year-fixed effects and a constant were estimated but are not reported. LSCHOOLING, LLIFE_EXP, LINCOMETOP20, and LCONTRIB_RATE are sample period averages. In all regressions, White's heteroskedasticity-consistent (clustered) t-values are given in parentheses (brackets). ***, **, and * represent significance at 1, 5, and 10% level respectively.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGDPPC</td>
<td>1.34</td>
<td>1.10</td>
<td>1.08</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>(21.0)**</td>
<td>(23.6)**</td>
<td>(22.3)**</td>
<td>(13.4)**</td>
</tr>
<tr>
<td></td>
<td>(11.2)**</td>
<td>(10.2)**</td>
<td>(9.72)**</td>
<td>(6.37)**</td>
</tr>
<tr>
<td>LPOP</td>
<td>0.35</td>
<td>0.25</td>
<td>0.26</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>(7.39)**</td>
<td>(10.5)**</td>
<td>(11.1)**</td>
<td>(7.01)**</td>
</tr>
<tr>
<td></td>
<td>(3.76)**</td>
<td>(4.25)**</td>
<td>(4.46)**</td>
<td>(3.11)**</td>
</tr>
<tr>
<td>LPOP_DENSITY</td>
<td>0.050</td>
<td>0.18</td>
<td>0.18</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>(1.16)</td>
<td>(6.53)**</td>
<td>(6.47)**</td>
<td>(3.32)**</td>
</tr>
<tr>
<td></td>
<td>(0.58)</td>
<td>(2.52)**</td>
<td>(2.50)**</td>
<td>(1.40)</td>
</tr>
<tr>
<td>INFLATION</td>
<td>(-4.55)</td>
<td>-0.73</td>
<td>-0.70</td>
<td>-5.28</td>
</tr>
<tr>
<td></td>
<td>(-4.04)**</td>
<td>(-0.90)</td>
<td>(-0.88)</td>
<td>(-4.34)**</td>
</tr>
<tr>
<td></td>
<td>(-3.25)**</td>
<td>[-0.66]</td>
<td>[-0.64]</td>
<td>[-2.79]**</td>
</tr>
<tr>
<td>LCARS</td>
<td>(-7.87)**</td>
<td>(-4.42)**</td>
<td>0.72</td>
<td>(6.05)**</td>
</tr>
<tr>
<td></td>
<td>(-4.52)**</td>
<td>[-1.93]</td>
<td>[-5.66]**</td>
<td>[-2.32]**</td>
</tr>
<tr>
<td>LTRADE</td>
<td>-0.50</td>
<td>0.49</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>(-7.87)**</td>
<td>(1.98)**</td>
<td>(3.78)**</td>
<td>(1.87)**</td>
</tr>
<tr>
<td>LLIFE_EXP</td>
<td>-3.01</td>
<td>-3.46</td>
<td>-3.46</td>
<td>-3.46</td>
</tr>
<tr>
<td></td>
<td>(-4.52)**</td>
<td>(-5.66)**</td>
<td>(-5.66)**</td>
<td>(-5.66)**</td>
</tr>
<tr>
<td>LAGEDEP</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>(1.98)**</td>
<td>(1.98)**</td>
<td>(1.98)**</td>
<td>(1.98)**</td>
</tr>
<tr>
<td>LAGEDEP_OLD</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>(3.78)**</td>
<td>(3.78)**</td>
<td>(3.78)**</td>
<td>(3.78)**</td>
</tr>
<tr>
<td>LAGEDEP_YNG</td>
<td>0.46</td>
<td>0.46</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>(2.68)**</td>
<td>(2.68)**</td>
<td>(2.68)**</td>
<td>(2.68)**</td>
</tr>
<tr>
<td>LSCHOOLING</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.38</td>
<td>-0.38</td>
</tr>
<tr>
<td></td>
<td>(-0.81)</td>
<td>(-0.81)</td>
<td>(-0.81)</td>
<td>(-0.81)</td>
</tr>
<tr>
<td>LINCOMETOP20</td>
<td>-0.39</td>
<td>-0.39</td>
<td>-0.39</td>
<td>-0.39</td>
</tr>
<tr>
<td></td>
<td>(-1.28)</td>
<td>(-1.28)</td>
<td>(-1.28)</td>
<td>(-1.28)</td>
</tr>
<tr>
<td>MUSLIM</td>
<td>-0.49</td>
<td>-0.49</td>
<td>-0.49</td>
<td>-0.49</td>
</tr>
<tr>
<td></td>
<td>(-2.44)**</td>
<td>(-2.44)**</td>
<td>(-2.44)**</td>
<td>(-2.44)**</td>
</tr>
<tr>
<td>Observations</td>
<td>266</td>
<td>538</td>
<td>538</td>
<td>343</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.699</td>
<td>0.677</td>
<td>0.681</td>
<td>0.699</td>
</tr>
<tr>
<td>Countries</td>
<td>65</td>
<td>76</td>
<td>76</td>
<td>60</td>
</tr>
</tbody>
</table>
What drives the development of the insurance sector?  
An empirical analysis based on a panel of developed and developing countries

as important drivers of insurance assets. Depending on the other explanatory variables included in the models, the sample size varies from 538 (76 countries) to 61 (35 countries) annual country-level observations.

We first discuss Table 7. Model 1 documents the results for the baseline variables and economic factors specific to the non-life sector: cars penetration and trade activity (LCARS, LTRADE). As expected, LGDPPC remains positive and highly significant, and so does LPOP. POP_DENSITY is not significant in this regression but is significant in other specifications. Also, INFLATION exhibits a negative and statistically significant association. Interestingly, LCARS is negative, which could point to a substitution effect between life and non-life activities—in the presence of unusually high car penetration, insurance companies might forgo fully developing the non-life industry and accumulate less assets. We also find that LTRADE is very significant, suggesting that more open countries accumulate more insurance assets. Since the correlation between these two factors is low, and LCARS is controlled for as well, this suggests that some forms of trade-related insurance are more asset-intensive than car-related insurance.

Models 2 and 3 investigate other relevant demographic factors: life expectancy and age dependency (LLIFE_EXP, LAGEDEP). The results confirm our previous findings in the life regressions. LLIFE_EXP enters negatively and statistically as very significant, although this result does not hold when additional variables are included, as in the case of LIFE. Interestingly, LAGEDEP is now significant as well. The coefficients of LAGEDEP_OLD and LAGEDEP_YNG are both positive and significant, as in the case of LIFE. This finding confirms that young and old dependency ratios are important drivers of insurance development, operating through different channels. A high LAGEDP_YNG induces the accumulation of assets to ensure protection against mortality risk, while a high LAGEDP_OLD induces the accumulation of assets for retirement (both in the accumulation and payout phases). Model 4 studies additional economic and cultural factors: expected years of schooling, income inequality, and religion (LSCHOOLING, LINCOMETOP20, MUSLIM). We do not find statistical significance of LSCHOOLING and LINCOMETOP20. However, MUSLIM is negative and statistically significant in an unclustered regression, implying that religion not only matters for premiums, but also reduces asset accumulation.

Next we turn to Table 8. Model 1 presents the results for institutional and market factors: social security, market concentration, private sector dominance, and legal rights (LCONTRIB_RATE, CONCENTR., PRIVATE, LEGAL_RIGHTS). The results are very similar to those obtained for LIFE. The baseline variables are all significant and with the same signs as those obtained for LIFE. In addition, we also observe that larger social security systems are associated with lower levels of asset accumulation. Market concentration (CONCENTR.) enters positively and significantly in line with LIFE as well. Finally, private sector dominance (PRIVATE) and legal rights (LEGAL_RIGHTS) exert a positive and significant influence on asset accumulation. Models 2 and 3 show the result for financial development (LPC, LPR_BOND). We find that financial development is key for asset accumulation, and that private credit (LPC) seems to exert a larger impact than private bond markets (LPR_BOND).

Finally, Models 4 and 5 present results for all relevant factors. Model 6 is identical to Model 5 but excludes LCARS to avoid a large reduction in sample size. Model 6 also reconfirms most of previous findings. In the presence of other factors, we find that most baseline variables remain significant and with the expected signs. Additionally, we document the robust positive association with LAGEDEP, CONCENTR., PRIVATE, and LPC. However, in these full models, LEGAL_RIGHTS and LCONTRIB_RATE lose statistical significance.

Dynamical panel regression results
Most of the factors we analyzed are arguably exogenous, thus reducing the problem of determining the direction of causality. However, two factors in particular could be endogenous: PC and CARS. Consequently, we adopt an instrumental variable approach á la Arellano and Bover (1995), using the same sample of annual country-level data for the period 2000-08. This method controls for country-fixed effects which capture time invariant characteristics of the country such as the average level of economic development. It also uses internal and external instruments for PC and CARS to address causality. Note that, since the models include a 1-period lag of the insurance variables, the interpretation of the coefficients is slightly different because they effectively capture the effect on insurance activity growth.
The Journal of Financial Perspectives

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGDPPC</td>
<td>0.72</td>
<td>0.52</td>
<td>0.45</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>(15.8)***</td>
<td>(13.4)***</td>
<td>(6.64)***</td>
<td>(2.64)***</td>
</tr>
<tr>
<td></td>
<td>[7.83]***</td>
<td>[5.33]***</td>
<td>[3.08]***</td>
<td>[2.05]***</td>
</tr>
<tr>
<td>LPOP</td>
<td>0.27</td>
<td>0.27</td>
<td>0.070</td>
<td>0.53</td>
</tr>
<tr>
<td></td>
<td>(8.19)***</td>
<td>(12.8)***</td>
<td>(2.54)***</td>
<td>(5.79)***</td>
</tr>
<tr>
<td></td>
<td>[4.16]***</td>
<td>[5.56]***</td>
<td>[0.95]</td>
<td>[3.96]***</td>
</tr>
<tr>
<td>LPOP_DENSITY</td>
<td>0.24</td>
<td>0.045</td>
<td>0.18</td>
<td>-0.18</td>
</tr>
<tr>
<td></td>
<td>(6.68)***</td>
<td>(1.88)***</td>
<td>(7.91)***</td>
<td>(-1.88)***</td>
</tr>
<tr>
<td></td>
<td>(2.94)***</td>
<td>[0.77]</td>
<td>(2.92)***</td>
<td>(-1.41)***</td>
</tr>
<tr>
<td></td>
<td>(-4.98)***</td>
<td>1.44</td>
<td>-7.76</td>
<td>-0.32</td>
</tr>
<tr>
<td></td>
<td>[3.44]***</td>
<td>[1.79]***</td>
<td>[3.06]***</td>
<td>[0.29]***</td>
</tr>
<tr>
<td></td>
<td>[-2.14]***</td>
<td>[1.25]</td>
<td>[-1.85]***</td>
<td>[0.21]***</td>
</tr>
<tr>
<td></td>
<td>4.39</td>
<td>4.39</td>
<td>0.79</td>
<td>4.39</td>
</tr>
<tr>
<td>LLIFE_EXP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.43)**</td>
<td>(1.96)***</td>
<td>(0.46)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LAGEDEP</td>
<td>2.47</td>
<td>1.33</td>
<td>2.47</td>
</tr>
<tr>
<td></td>
<td>(4.07)**</td>
<td>(2.36)**</td>
<td>(1.55)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[3.00]***</td>
<td>[1.55]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.33</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSCHOOLING</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-1.60)***</td>
<td>(-0.66)***</td>
<td>(-0.48)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-1.17)***</td>
<td>[-1.00]</td>
<td>-0.024</td>
<td></td>
</tr>
<tr>
<td>LINCOMETOP20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-1.58)***</td>
<td>(-0.046)***</td>
<td>(-0.030)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-1.12)***</td>
<td>[-1.00]</td>
<td>-0.024</td>
<td></td>
</tr>
<tr>
<td>MUSLIM</td>
<td>-0.62</td>
<td>-0.61</td>
<td>(-2.95)***</td>
<td>[-1.75]***</td>
</tr>
<tr>
<td></td>
<td>(-2.01)***</td>
<td>[-1.65]***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LCONTRIB_RATE</td>
<td>-0.20</td>
<td>0.074</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>(-2.32)**</td>
<td>(0.53)***</td>
<td>(0.31)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-1.20)***</td>
<td>(0.35)***</td>
<td>(0.20)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.36</td>
<td>-0.38</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.98)***</td>
<td>(-2.00)***</td>
<td>(1.82)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[1.65]***</td>
<td>[-1.58]***</td>
<td>[1.07]***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.46</td>
<td>0.85</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.00)***</td>
<td>(3.64)***</td>
<td>(2.46)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[1.57]***</td>
<td>[2.33]***</td>
<td>[1.54]***</td>
<td></td>
</tr>
<tr>
<td>PRIVATE</td>
<td>0.13</td>
<td>0.13</td>
<td>0.031</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.50)***</td>
<td>(1.01)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[1.77]***</td>
<td>[0.66]***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEGAL_RIGHTS</td>
<td>(4.67)**</td>
<td>0.83</td>
<td>0.72</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>[-1.37]***</td>
<td>[3.96]***</td>
<td>(6.05)***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[5.76]***</td>
<td>[2.66]***</td>
<td>(3.58)***</td>
<td></td>
</tr>
<tr>
<td>LPC</td>
<td>0.28</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.70)***</td>
<td>(2.36)**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LPR_BOND</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.46</td>
<td>-0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-3.36)***</td>
<td>[-2.41]***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LCARS</td>
<td>1.17</td>
<td>0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(4.80)***</td>
<td>(4.05)***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[3.43]***</td>
<td>[2.59]***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LTRADE</td>
<td>Observations 255</td>
<td>534</td>
<td>301</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>R-squared 0.723</td>
<td>0.724</td>
<td>0.671</td>
<td>0.929</td>
</tr>
<tr>
<td></td>
<td>Countries 62</td>
<td>76</td>
<td>39</td>
<td>35</td>
</tr>
</tbody>
</table>

Table 8 - Development drivers of the insurance industry’s total assets

The dependent variable is the logarithm of total insurance assets as a fraction of GDP. The sample consists of annual data for the period 2000-08. All regressions are estimated using pooled Ordinary Least Squares. Year-fixed effects and a constant were estimated but are not reported. LSCHOOLING, LLIFE_EXP, LINCOMETOP20, and LCONTRIB_RATE are sample period averages. In all regressions, White’s heteroskedasticity-consistent (clustered) t-values are given in parentheses (brackets). ***, **, and * represent significance at 1, 5, and 10% level respectively.
Following previous sections, in each of our regressions, we include the baseline variables LPOP, LPOP_DENSITY, and LINFLATION. For consistency with previous regressions, we show results with and without LGDPPC. However, we find that it is difficult to fully disentangle the contemporaneous effects of LGDPPC and our drivers, particularly in the presence of country-fixed effects.

As exogenous (internal) instruments, we use the following set: LPOP, LPOP_DENSITY, (French and British) legal origin, and religion dummies. In our specifications, we consider LGDPPC and INFLATION to be (possibly) endogenous. We include 2- to 4-period lags as internal instruments for the endogenous variables. Depending on the specification, the regressions pertain to 71 up to 82 countries.

Table 9 presents the results. Models 1 and 2 confirm the causal effect of LPC on LLIFE, although the inclusion of LGDPPC produces a p-value for LPC of 0.145. Model 1 also finds consistent results for LPOP and LPOP_DENSITY and shows the causal effect of LGDPPC on LLIFE, consistent with previous research showing that LGDPPC and insurance activities have a causal effect on one another. Similarly, Models 3 and 4 confirm that LPC causes NONLIFE but controlling for LGDPPC produces a p-value of 0.21 for LPC. Not surprisingly, we find that LGDPPC exerts a causal effect on LNON-LIFE. Consistent with earlier findings, the LPC coefficient is smaller for non-life than life, which suggests that financial development is more important for life insurance growth. Models 5 and 6 show that LPC causes ASSETS growth, even in the presence of LGDPPC. We also reconfirm the relevance of population size. Finally, Models 7 and 8 confirm the causal effect of LCARS for NONLIFE. This finding confirms car penetration as a driver of insurance development and not merely a result of it. However, again, LGDPPC absorbs the significance of LCARS (p-value=0.2).
Conclusions and policy implications

In this paper we explored the main factors that drive the development of the insurance sector, including both the life and the non-life sectors. The paper contributes to the body of empirical research by covering both sectors, measuring their development by premiums and assets, and introducing a number of additional explanatory variables. We believe that this more integrated and comprehensive approach provides additional insights over those provided by previous research.

The life insurance regressions confirm some of the findings of previous empirical research on life insurance and add some additional findings. We find that income is an important driver of life insurance, but that population and population density are also important drivers. These two variables were, surprisingly, overlooked in previous multi-country research efforts, but show the importance of larger clienteles, deeper risk pools, and scale economies, as well as easier distribution channels. Inflation hinders the life sector’s development, in line with previous research results.

Life expectancy has a negative and significant influence in some of the specifications, suggesting that higher life expectancy reduces demand for life insurance by reducing the probability of a premature death. Age dependency proves to be an important driver of life insurance when the young and old dependency ratios enter separately. The results confirm that a high young-age dependency ratio drives the demand for insurance against mortality risk, while a high old-age dependency ratio drives the demand for insurance against loss of income at old age. Schooling does not seem to be an important driver of life insurance, in line with previous research, but the share of the richest 20% of the population has a positive effect, suggesting that some life insurance products may be regarded as luxury goods, all else being constant. The predominance of a Muslim population tends to hinder the development of the life sector, in line with previous research results.

The results show that a large social security system (measured by the contribution rate) hinders the development of the life sector, by partly reducing the need for insurance but also by reducing the level of disposable income net of taxes and contributions. The predominance of private ownership in the life industry and a strong legal framework promote the development of the life sector, and so do developed credit and bond markets. Interestingly, more concentrated markets tend to promote faster industry development. This somewhat surprising result may indicate the importance of large risk pools and scale economies.12

The regressions for the non-life sector show that income is an important driver and that the predominance of Muslims in the population slows the sector’s development. The results also show the importance of institutional and market structure variables. As in the case of life insurance, private ownership of the industry, a strong legal framework, and developed credit markets promote the development of the non-life sector. Curiously, market concentration slows the development of the sector, a result that is in principle more intuitive than the one obtained for life insurance, although the result is not very robust across specifications.

The regressions for the non-life sector also show that the fleet of cars and the volume of trade are important drivers of the sector’s development. Interestingly, the impact of trade seems to be captured by the size of population, reflecting the importance of external trade in small and open economies. Surprisingly, inflation seems to promote larger non-life premiums, but this result could be the secondary effect of the shifts from financial to real assets that are observed in high-inflation environments.

The results for insurance assets essentially mirror those obtained for the life sector. The baseline variables are significant in most regressions and the coefficients have the same signs as those obtained for the life regressions. The Muslim variable enters negatively and significantly, in line with the results for both sectors. Reassuringly, we confirm that the combined effect of high inflation on the two sectors is negative, reflecting the much stronger and negative impact on the life sector. We also confirm the positive and significant impact of private ownership and of deep credit and bond markets on the development of the insurance sector. Finally, a strong legal framework also promotes the development of the insurance sector, although this result is not robust across all specifications.

12 However, Rocha et al. (2008) find that the low concentration of the annuities market in Chile contributed to more competition and a higher annuity rate offered to annuitants, probably contributing to a higher demand for annuities and increasing premiums and assets.
What drives the development of the insurance sector?
An empirical analysis based on a panel of developed and developing countries

These results have a number of important policy implications. Some of the drivers of the insurance sector are not within the reach of policy-makers or can only be influenced over long periods of time, but the results indicate that supportive policies can contribute to the acceleration of the sector’s development. They show, for example, the importance of a stable macroeconomic framework and low inflation for the sector’s development. The positive impact of private ownership on the sector’s growth is another important finding, as the state still plays a predominant role in many countries. The insurance sector flourishes under a supportive legal framework, and also benefits significantly from developed credit and bond markets. Religion plays a role, suggesting the need for more progress in introducing insurance institutions and products more harmonized with religious beliefs, such as the case of Takaful arrangements in Muslim countries.

Despite a number of additional findings, our research was not able to capture all the factors that may influence the development of the insurance sector, such as recent severe perturbations (conflict, financial crisis), the regulation of the rapidly growing bancassurance channel, the differential taxation of insurance products, and regulations that enhance consumer protection (and thus support greater levels of trust in the sector). Thus, after controlling for the main drivers identified in this paper it is important to make a more detailed analysis to identify and understand the additional factors that accelerate or hinder the development of the insurance sector in a specific country.

For example, Lester (2010) shows that Middle East and North African countries have generated much smaller revenues from car insurance than would be expected by their fleet of cars, due to several problems in the regulation and enforcement of motor third-party liability insurance (Figure 1). However, our results provide a number of additional insights into the main drivers of the insurance sector and also highlight areas for future research on this important area.

References

Figure 1 - Motor premium penetration versus personal car density
Source: Lester (2010).


The customers have spoken, are the insurance companies listening? A survey of the global insurance market

Paul Clark
Asia Pacific Insurance Leader, Ernst & Young LLP

Christine Delany
Director, Global Insurance Advisory, Ernst & Young LLP

Shona Burns
Senior Manager, Insurance Advisory, Ernst & Young LLP

The challenge of responding effectively to rapidly changing consumer needs and behaviors is recognized as one of the top strategic agenda items. We set out to obtain hard evidence of customer attitudes and behaviors. We undertook a worldwide insurance customer survey – unprecedented in both its scope and scale – to explore what drives consumer behavior across the whole customer life cycle and to test some of the received wisdom held in the industry today; to separate myth from reality, if you will. Our research reveals that while, of course, there are significant variations in customer attitudes and behaviors around the globe, there are some underlying themes that are remarkably consistent. Our findings also clearly show that insurers are far behind other sectors in meeting consumers’ expectations. So how should insurers respond? They undoubtedly need to rethink their approach to customers and seek to implement a tailored, yet lower-cost “customer-centric” operating model to support it better. Insurers will need to define what it means to be customer-centric and develop critical new competencies, aligning their operating models accordingly. Significantly, although the above activities may be considered transformational to many in the industry, we believe they will only provide insurers with the ability to catch up and otherwise “protect the core.” To leapfrog competitors and generate significant growth requires true customer-centric innovation and a significant change in approach to strategic decision-making. Implementing these strategies will future-proof the core business, while allowing insurers to redefine customer relationships and become a different insurer – the insurer of the future.
Introduction
The financial services industry continues to operate in a challenging, volatile, and uncertain economic environment. Global Insurers, in particular, face significant headwinds as competitive forces, increasing regulation, disintermediation, and new technologies impact their business models and put immense pressure on the cost of doing business. Although the demand for insurance, savings, and investment products remains relatively strong across the world, increasingly diverse customer needs, buying behaviors, and expectations, combined with rapidly evolving consumer protection initiatives, make navigation through this difficult time more challenging. The ever-increasing integration of technology and life, through the use of smart devices, mobile connectivity, and social media, are blurring the boundaries of online and offline worlds, giving greater power and choice to consumers and driving a fundamental shift in customer expectations in terms of how products are marketed, priced, sold, and serviced, and, of course, how companies are perceived.

Learning from previous recession experiences, the winners in the insurance market will be those that equip themselves with a clear understanding of customer needs and behaviors in all relevant geographies, sectors, and channels, to help them drive profitable growth strategies and provide the confidence to invest in growth opportunities, at a time when securing budget is extremely difficult.

The discussions we have had with global insurance executives, non-executive directors, regulators, and supervisors corroborate this view and demonstrate, without exception, that the challenge of responding effectively to rapidly changing customer needs and behaviors is one of their top strategic priorities, alongside regulatory change and talent management.

In response to this, we set out to obtain real data and evidence around customer attitudes and behaviors, and, specifically, to answer the following two critical questions: how do consumers view and interact with the insurance industry today, and more importantly, what are their future intentions and expectations? We believe that insurers can no longer rely on long-held assumptions and received wisdom about what consumers think and how they behave. Our objective was to truly listen to “the voice of the customer” and thereby understand how companies need to evolve if they are to capture the growth opportunities now and after the global economic crisis settles. Working with the independent firm Ipsos, we canvassed the opinions of 24,000 consumers across 7 regions1 and 24 countries around the world, covering both life and non-life (personal lines) sectors. In this research, we set out to explore what drives consumer behavior across the whole customer life cycle, from initial research through purchase, servicing, repeat purchases, claims, and termination. We also set out to test some of the received wisdom held in the industry today, to separate myth from reality, and provide hard evidence of what customers really demand. The overwhelmingly positive response and huge level of interest we have received since publication, is yet another indication of how unprecedented this kind of fact-based research is in the insurance industry.

The world is changing: it will be reshaped, driven by changes in the global economy, technology, and customers
The current macroeconomic environment presents a considerable challenge for insurers operating in mature, developed economies. High unemployment rates, downward pressure on household income, continued housing market issues, and other factors, continue to exert pressure on the demand for insurance products. Worldwide, real growth of non-life premiums totaled just 1.9% in 2011, while life premiums declined by 2.7% from their 2010 levels.2 Given this context, insurance companies need to up their game in order to maintain their “share of wallet,” let alone grow it.

Even within emerging markets that have experienced strong growth over the past decade, where insurers have battled hard to win the first product from the customer, the growth curves have now started to flatten. Insurers in southeast Asia are reflecting on the longer-term issues that have plagued the more mature markets and anticipate that similar regulations flowing through to their markets will challenge the traditional distribution models in this region. We are already seeing many multinational companies in this region preemptively placing more focus on customer satisfaction indices and persistency rates as part of senior management performance scorecards, and introducing customer-centric initiatives such as welcome calls. Notably, domestic companies appear to be further behind.

1 For the purpose of this article, we have separated the results for the Americas region between North and South America to reflect the differences between the developed and emerging markets within the region.
2 Swiss Re, sigma No 3/2012.
Industry profitability also continues to be pressured by weak investment returns, driven by the current low interest rate environment and equity market volatility. This, combined with uncertainties surrounding the impact of new regulations and the ongoing Eurozone debt crisis, has raised the bar for business case development for investments in new or expanded capabilities. As a result, the historical cycle of spending on cost cutting during difficult times and spending on growth initiatives during prosperous times has been replaced with the “new normal” of managing investments with the dual purpose of growing the business while maintaining cost competitiveness.

At the same time, a new global economy is emerging that presents opportunities as well as challenges for insurers. A shift from West to East, with developed economies contributing less to the global GDP in future (two-thirds of new middle-class entrants will come from China and India), is introducing new players to the economic scene. Increasing urbanization, as populations shift to economic hubs, is creating megacities. Yet, at the same time, a global village emerges as communities form, based on common interests, unrestricted by geographic boundaries.

Technology is moving faster than anyone ever imagined and adoption is growing at an exponential rate. The integration of technology and life is a fact. We are never offline, we are increasingly connected. The use of smart devices, mobile connectivity, and social media are blurring the boundaries of online and offline worlds. By 2020, 80% of the global population will have access to mobile telephony and more than 60% to smart phones or low-cost tablet computers. It is anticipated there will be over 50 billion connected devices globally, with mobile devices being the primary internet device for most individuals. As a result, real time decision-making will increase and the phone will become the main purchasing and payment tool for many.

Social media and unprecedented access to information, such as peer-to-peer product and service reviews, are giving greater power and choice to consumers, creating more informed and more demanding customers. Insurers are dealing with a very different customer.

What does this mean for insurers? Well, it is a different world. Disintermediation has changed, and continues to change, the customer relationship. By losing touch with customers, insurers risk becoming viewed as only administrators, managing the lowest margin in the value chain. In some non-life markets, this is further exacerbated by the fact that there are two parties involved in the customer relationship, the broker and a comparison website. In mature markets, one of the biggest challenges life insurers with largely intermediated distribution models have encountered is how to introduce and integrate fledging direct operations effectively, and how to manage the ensuing customer ownership issues with existing agents and intermediaries. This will similarly become a significant issue in emerging markets like Asia, as direct channels emerge to complement traditional agent channels, to service more demanding and tech-savvy consumers.

Evolving consumer protection in most mature markets is also posing a challenge, forcing insurers to shift operating models and undertake significant investments in new capabilities. Regulators across the globe have responded to the financial crisis by tightening regulations to improve outcomes and reduce risk for customers, such as the E.U.’s PRIIPs and MiFID regulations, the Dodd-Frank Wall Street Reform and Consumer Protection Act 2010, or the U.K. (abolishing commission and professionalizing advice), and the Australian FOFA reforms, which promise to “go further than RDR.” Even in advanced developing markets we can see signs of a rapidly changing regulatory landscape. Hong Kong and India are introducing remuneration disclosures and moving away from commissions. Singapore is initiating the FAIR review, which will make significant increases in advisor qualifications quite likely and raise questions about future remuneration models. South Africa is considering RDR/TCF-like initiatives to improve the professionalism of advice, and other emerging markets are looking to the U.K.’s Financial Services Authority (FSA) as a.

---

4 Hans Vestberg, President & CEO of Ericsson, GSMA Mobile World Congress 2010.
4 Hans Vestberg, President & CEO of Ericsson, GSMA Mobile World Congress 2010.
model upon which to base their regulatory regimes. All this is placing significant pressure on advice and distribution, to improve customer outcomes while maintaining cost-efficiency.

In response, both insurers and intermediaries need to focus on improving productivity, getting a better understanding of client value by channel and other segments, and reducing the “cost-to-serve,” for example, by passing parts of the process back to the customer in the form of self-service. Distributors who cannot afford to manage their entire client portfolio proactively are looking to product providers to help them reduce costs through advisor technologies, using a B2B2C approach.

The competitor landscape will inevitably look different in the future. Pressure is already building as new entrants, operating models, and technologies become a real threat to the ownership of the customer relationship. Unless insurers adapt quickly and respond to these changes with innovative solutions, they risk becoming irrelevant.

Listening to the voice of the customer
We believe that, faced with these unprecedented challenges in the market, there is a risk that some insurers may not be listening and responding to the most important voice of all – that of their customers. For any insurers hoping to navigate their way through this difficult time, understanding how customer behavior and attitudes are changing is critical. Previous assumptions and received wisdom about customers may no longer be reliable and those insurers who are able to respond best to what customers want now are most likely to succeed. The challenge is to really understand what drives customer behavior today and tomorrow, not yesterday, in all geographies, sectors, and channels where they operate.

Our Ernst & Young 2012 Global Insurance Consumer Survey (published February 2012) found that while, of course, there are significant variations in customer attitudes and behaviors around the globe, there are many underlying themes that are remarkably consistent.

**Theme 1: customers want to be able to buy with confidence**
In both life and non-life sectors, customers want products, and the purchasing process, to be simple and transparent so that they can understand what they are buying. They want to build long-term relationships with insurance providers based on trust, and to have confidence that the products they are buying are right for them and meet their needs.

In the life sector, one of the most striking and positive findings from our research is that the reputation of the industry has not been tarnished by the global financial crisis. Received wisdom is that the financial crisis has created a perception that all financial services companies are untrustworthy. However, our research indicates this is not the case in the insurance sector, where most customers have a surprisingly positive view of the industry. The
majority of customers still have reasonably high satisfaction levels and are confident that the products meet their needs (Table 1). However, as shown in Table 1 and in more detail in Table 2, there is no room for complacency. For a significant minority of customers in Japan, and in some developing markets in Asia Pacific (China, Hong Kong, South Korea) and Europe (Poland and Turkey), insufficient product understanding and information about product suitability is creating a lack of confidence that their product is right for their needs.

Of course, many of the products that exist in today's more developed life markets have historically been driven by the needs of regulators and intermediaries or agents. This is now changing, as providers recognize they need to simplify both products and product information, to meet the customers' need to understand clearly what they are buying. Global insurers' focus should now be on increasing levels of customer confidence in regions such as Asia, Japan, and some of the emerging markets in Europe, where respondents said the reasons they were not satisfied or confident that their policy met their needs were lack of information on how their products were performing, they did not fully understand the terms when they first bought the policy, or the fact that their needs had since changed (and had not been addressed). This implies considerable future challenges for emerging markets with traditional agency models, around agent training and agent attrition (orphan levels are already high and rising in these markets) as regulations tighten and customers become more demanding and sophisticated.

Our survey found a clear correlation between those customers who conduct higher levels of research and their confidence levels in a number of countries, including Singapore, China, Poland, Turkey, the Netherlands, and UAE. Those who conducted a high amount of research are 10% to 17% more confident than the average customer that the product they have purchased is right for their needs. Results from those customers surveyed in Japan reflected a more than 30% increase in confidence as a consequence. Interestingly, Indian and South African customers who appear to conduct significantly more research than any other country also have the highest levels of confidence overall.

Overall, customers are looking for value to be clearly demonstrated, reflecting a balance of price, product features, and service tailored to their needs. They also expect the buying process to be convenient and transparent, allowing them to buy with confidence. Once they have made their choice and have an established relationship with an insurance provider, customers expect the product and service to deliver against their expectations. This may sound obvious, but the survey shows that customers' perceptions are that the industry is failing to deliver this in some key areas, as will be described later.

There is significant room for improvement compared with other industries. Despite the general level of satisfaction, customers believe the industry performs less well compared with others in terms of providing customer service and rewarding their loyalty. Across all regions and in both life and non-life sectors a significant number of consumers agree with the statement that the industry is trailing behind other sectors in both quality of service and how it rewards loyalty. These perceived shortcomings are most prevalent in India, South Africa, and UAE (Table 3).

Customers clearly want the industry to serve them better. Notably, many of the service improvements sought by customers in the life sector are centered around how they interact with insurers, whether it be improving the effectiveness of call centers, simpler and more transparent communications, or enhanced online services and information. There appears to be a shift away from the traditional “service issues” of billing and general administration, with the exception of India, where claims processing is highlighted as being particularly problematic (Figure 1). These findings show that the industry could do a lot more to bring itself into the 21st century.
In non-life insurance in particular, price is often the main measure of value since products are more comparable and the frequency of purchase drives greater customer familiarity and confidence. Price is, of course, important but increasingly, other factors such as brand and reputation are playing a bigger role, perhaps more than insurers either realize or care to acknowledge. This may be a reflection of market structure (such as a recent history of price tariffs) or it may be a more fundamental desire from consumers to find a measure of quality they trust. In highly competitive markets characterized by price transparency, there is a tendency for prices to converge. This leads to non-price factors such as brand becoming more important as customers search for a way to differentiate between providers – financial stability and brand, for example, can become a “kite mark” for quality. In contrast to the life sector, much of the improvements sought in non-life insurance are centered around product fundamentals, including discounts and coverage levels, as well as transparency of terms and conditions (Figure 2).

So what does this mean for insurers? Well, most significantly of all, we found that customers judge insurers against other consumer industries: they expect comparable standards of service and rewards for loyalty, similar to those they already receive from non-financial services companies. In our experience, insurers typically benchmark themselves against their peers in the industry, obviously a massive disconnect. Insurers need to evolve customer propositions continually to meet changing needs and expectations, particularly around improved information and transparency – and they need to look outside the industry to do this effectively.

Of course, the need for clarity and transparency is not driven purely by customer preferences. As mentioned earlier, as a result of the financial crisis, regulators around the globe are focusing on protecting consumer interests more than ever before, with major new regulations being introduced in the E.U. and many other countries around the world. Insurers that align themselves to a customer-centric model of operation will find the transition to the new regulatory environment less painful, and may gain competitive advantage compared to those that continue to rely on outdated models that fail to satisfy customers or regulators.

Figure 1: Which two or three of the following elements, if any, of the service you receive from your product provider would you like to see improved?* – life sector

*Multiple answer question. Customers were asked to select all answers that apply, therefore total may add up to more than 100.

Back office processing

Insurer interaction
Theme 2: the internet and technology are changing the way customers engage with insurers, but traditional channels remain important

There is strong evidence in our findings that the way customers want to interact with insurers is changing. As noted previously, the internet is transforming consumer offerings in other retail industries and setting a new benchmark for how businesses interact with consumers. The customer-centric nature of pure internet businesses, such as Amazon or eBay, is influencing customers’ expectations. These businesses are built on customer data, recognize each customer as a unique individual, and are able to offer product and service suggestions based on customers’ past behavior and known preferences. Many also build in independent data as part of the offering, inviting customers to review what others thought of their purchases and making it easy for consumers to compare pricing and offerings.

Other industries, ranging from consumer goods to airlines and hotels, have to respond to the widespread availability of internet price comparisons and independent quality ratings, often compiled by their own customers. These developments help customers buy with more confidence—and set an expectation that all sectors, including insurance, have to meet. This transparent, information-rich environment has fed an expectation among consumers that they will be able to do more independent research when buying insurance.

While the level of actual insurance purchase over the internet remains low in many countries, its use varies considerably between countries in the survey. In every case we found that customers intend to do more research using the internet, even if they ultimately rely on conventional channels for purchase (Figure 3).

Many insurers are active in developing digital tools across several areas of the value chain. These have focused primarily on the provision of information, modeling personal requirements, and the generation of quotations. Non-life insurance lends itself more to internet purchase than life and pensions, given the higher customer familiarity and comparability of the products. Market innovations, such as mobile phone insurance quotation, policy maintenance, and insurance claims submission apps, signal the beginning of the mobile channel becoming a new medium for competing for customers.

---

11 Price comparison websites allow people to see different prices offered by different insurance companies for an insurance product. Some of these also allow customers to buy insurance on their website, while others do not.
The big question for the industry is what is holding customers back from transacting online? Is it pure preference, the limitations of digital capability currently on offer, data security, or the complexity of products for certain customer segments? Could the currently low levels of online transactions shift exponentially if these root causes were addressed?

In life and pensions in particular, the old adage that “life insurance is sold, not bought” is not the case for a growing minority of consumers. Received wisdom is that, because of a lack of customer knowledge and confidence, life insurance products must be sold to consumers, i.e., the purchasing decision is not customer-driven. Despite this very persistent belief in the industry, we found strong evidence that customers want to take more control when buying life, pensions, and investment products. The survey reveals that they are increasingly seeking information themselves, rather than simply relying on being “sold to.” As stated in Theme 1, research levels are already high in South Africa and India where 67% and 74% of respondents, respectively, say they conducted a great deal or a fair amount of research, in comparison with only 18% in Japan, 24% in South America and 27% across Europe. For future preferences, this is expected to increase by 35%–40%, with the greatest increases anticipated in South America and Japan (Table 4).

The internet is enabling this trend. Use of online information sources is increasing. Customers want to use sources they can trust, without risk of sales bias. In the past, customers often used friends and family as their trusted reference points. Although this is still important, our findings show increasing willingness to use online sources, which they perceive to be objective, for future research. These include comparison sites, blogs, and social media. Advice from an agent is still mentioned as a frequently consulted source, but it is only one of a wide range of sources consulted.

<table>
<thead>
<tr>
<th>Country</th>
<th>Previous</th>
<th>Future</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japan</td>
<td>18%</td>
<td>57%</td>
</tr>
<tr>
<td>Asia Pacific</td>
<td>39%</td>
<td>73%</td>
</tr>
<tr>
<td>Europe</td>
<td>27%</td>
<td>65%</td>
</tr>
<tr>
<td>UAE</td>
<td>49%</td>
<td>83%</td>
</tr>
<tr>
<td>North America</td>
<td>28%</td>
<td>66%</td>
</tr>
<tr>
<td>South America</td>
<td>24%</td>
<td>69%</td>
</tr>
<tr>
<td>South Africa</td>
<td>67%</td>
<td>69%</td>
</tr>
<tr>
<td>India</td>
<td>74%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Table 4: Customers who undertake a fair amount or great deal of research before buying – life sector
We found that customers understand their needs and what products should do. They have a clear understanding of the financial needs they want their products to meet, and this shows a logical variance by life stage. In their research, customers are looking for factual information to confirm they are selecting the right product to meet their needs. The top three factors customers identified across all regions were: financial stability, product features, and product performance track record. Factors such as brand or advisor recommendation are less important, with the exception of India where customers placed the brand of the insurance provider in their top three. This shows that customers are less influenced by providers’ own claims or the recommendation of a salesperson, which could be biased. They want objective information to ensure that they make the right choice.

The implication of all this for insurers is that influencing objective sources of information is critical. Traditional marketing and distribution methods are not well-suited to the emerging customer trend of researching before purchasing, using objective and independent sources. The challenge for insurers is how to communicate their propositions so that objective sources represent them fairly, allowing the customer to make a well-informed choice. Actions speak louder than words. This goes beyond simple advertising. Getting the right message across to customers on internet-based information sources is a new challenge, to which other consumer industries are already responding. In essence, the internet offers huge opportunities for organizations that can harness the digital consumers to their advantage – nothing less than a massive new marketing department that is not even on the payroll. Additionally, billions of dollars are spent annually by insurers on publishing sales literature that consumers are telling us they no longer want – much of this could be saved by investing in brand and product positioning within social media, as this is where consumers today are increasingly building their brand respect. Insurers need to consider how they can simplify and demystify products and ensure they consistently deliver the service customers expect, so that independent commentators (including their own customers) comment positively about the company and its products through the independent sources. Simply telling customers how good you are cannot do this; you need to prove it by what you do and how you do it.

Despite these trends, personal interaction remains important for most customers. The received wisdom is that personal interaction is essential to the purchasing process and our research indicates that this largely remains the case for life sector customers (Table 5).

Most customers still lack the confidence needed to buy without assistance. The two most common reasons cited for the continuing need for personal interaction are that products are too technical and complicated and that customers feel they need expert advice when making important financial decisions. Despite the need for expert assistance, it is important to note, however, that although consumers rely heavily on personal interaction to complete a purchase, our research across Europe reveals that 46% of customers say they would not pay for advice from agents or intermediaries, while 18% are happy to pay commission, 10% an up-front fee, and 11% say they would pay a combination of fee and commission. More worrying, is that a reasonably high percentage of consumers in some countries believe that advice is provided free of charge. This suggests that many customers do not really understand the cost, or appreciate the value, of advice.

<table>
<thead>
<tr>
<th>Respondents who</th>
<th>Respondents who</th>
</tr>
</thead>
<tbody>
<tr>
<td>considered using a comparison site</td>
<td>actually bought through a comparison site</td>
</tr>
<tr>
<td>India</td>
<td>20%</td>
</tr>
<tr>
<td>South Africa</td>
<td>4%</td>
</tr>
<tr>
<td>South America</td>
<td>48%</td>
</tr>
<tr>
<td>North America</td>
<td>28%</td>
</tr>
<tr>
<td>UAE</td>
<td>40%</td>
</tr>
<tr>
<td>Europe</td>
<td>38%</td>
</tr>
<tr>
<td>Asia Pacific</td>
<td>49%</td>
</tr>
<tr>
<td>Japan</td>
<td>17%</td>
</tr>
</tbody>
</table>

Table 6: Importance of the internet in researching and purchasing – non-life sector
This problem may be exacerbated as regulators around the globe move to increase transparency of advisor remuneration and even, in some cases, ban commissions altogether. This is already observed in the U.K., where insurers are no longer able to pay commissions to advisors for the sale of investment products.

Customers would actually prefer products to be simpler and more transparent, making it easier for them to make an informed choice. A not-insignificant minority of the customers, mainly in developed countries (e.g., U.K., Germany, Netherlands, Japan) would prefer to take complete control over the buying process and do not want to be subjected to sales pressure. They also feel it is more convenient and they will get a better deal buying this way.

This has implications for the role of insurance intermediaries or agents – if customers are doing more research themselves and taking more responsibility for the product choice, what is the value-add of the intermediaries and what should they be remunerated for? In an industry where margins are coming under increasing pressure, insurers need to choose where they invest across the value chain, to avoid adding extra costs, while still paying intermediaries for services customers may prefer to access themselves.

We found a growing trend in all countries to using the internet to research non-life products (with nearly half of all respondents in South America and Asia Pacific considering using a comparison website), although levels of actual purchase vary considerably between countries. The highly developed U.K. internet aggregator channel is an outlier in this area (more than two-thirds of new business motor sales now go through this channel). We found that most other countries have a continued preference for more conventional channels to complete the purchase, including agents or brokers and direct channels. But, given the increasing use of the internet as a research channel, purchasing may follow if distributors make it easy for customers to compare prices and buy with confidence (Table 6).

Finally, the degree of use of online by non-life customers also varies by type of transaction: some are happy to use the internet to make a purchase but not to deal with a claim. In all locations, we found an emerging group of customers who had a preference for a wholly online experience, but most respondents still wanted a mix of online and personal contact. This is explored in detail within Figure 4, reflecting the results from customers surveyed in the U.K., a country where online has grown rapidly over the past decade.

As we have seen in other industries, the emergence of new technology does not mean that existing channels become obsolete; just as eBooks have not fully replaced books, magazines, and newspapers and Skype has not replaced face-to-face communication. Digital is a complementary communication tool – and one that needs to be integrated within the business to drive sustained success in the market.

For insurers, this is forcing dramatic change on the operating model. Our research shows that, regardless of purchase channel, customers across both life and non-life sectors show a strong preference for a mixture of online and offline interaction. The challenge for insurers is that undertaking the necessary integration to ensure that customers can contact them by any channel, and get a seamless and consistent response, will require a significant change effort for many. Insurers need to ensure their sales channels focus on where they really add value to the end customer – in particular, providing the right expertise to help people make well-informed financial decisions. This will be helped by greater transparency in the sales process and in product design.

Of course, personal interaction does not necessarily mean face-to-face contact. This can be delivered remotely at lower cost. Telephone contact can be appropriate for some transactions, provided it builds the necessary customer confidence. Increasingly, web-based chat or video interaction is becoming acceptable, leading to much richer opportunities for remotely delivered personal interaction. Insurers may be able to develop new, lower-cost ways of interacting with customers that meet the need for personalized contact without the requirement for expensive face-to-face channels. This could be important, as the cost of sale, including advice fees and commissions, becomes more transparent to customers due to disclosure rules.

Lastly, insurers cannot afford to ignore the new breed of independent consumer, and need to recognize the importance of customers who are confident to buy on their own, and respond to this segment’s needs. These customers are looking for good
value products that they can buy directly, preferably over the internet. New, mainly online, products are emerging to target this segment – particularly in the wealth management space and for buying simple life-protection products – using approaches similar to non-life insurance comparison sites. This sales route may take an increasing share of the market in future, particularly in mature markets, so insurers need to respond to this trend if they want to maintain market share.

**Theme 3: building long-term customer relationships**

Traditionally, this has been left to the agent or intermediary, which is why insurers now lag behind many other industries, including banks, in customer management and engagement. However, it is clear that consumers in the majority of markets perceive their main point of contact for dealing with their policy to be their insurer (potentially due to the lack of interest from their agent beyond the sale). Despite this, almost a third of life sector customers, on average, claim they receive no contact from their insurer each year and a similar proportion are contacted only once.

It is clear from our survey results that improving the level and quality of contact can have a positive impact on customer loyalty and retention. A positive pattern exists when comparing levels of satisfaction with provider contact and future switching level across many countries, including Mexico, U.K., U.S., and India. In the U.K., 3% of satisfied customers said they were certain to, or likely to, switch in the next five years, compared with 17% of dissatisfied customers.

Similarly, there is a strong correlation between comparing quality and levels of contact with cross-sell levels, particularly in Mexico, South Korea, the U.K., and Japan. In Mexico, 38% of satisfied customers said they had bought another product or service, compared with 15% of dissatisfied customers.

Received wisdom in the insurance industry has been that customers are reluctant to buy more products from the same provider – and indeed many insurers with poor cross-sales rates console themselves in the knowledge that their peers have equally dismal results. Our research shows that across both life and non-life sectors, consumers are willing, and indeed prefer, to buy more products from companies that they trust and who can make the purchase experience easy and convenient for them (Table 7).
Customers expect companies to contact them at product renewal or maturity, and as their needs change over their life cycle. By doing so, insurers will demonstrate an interest in retaining their business and can make suggestions about other products that would be appropriate to their needs.

In the life and pensions sector, the inherently more complex nature of the products and the dependence on intermediated distribution makes it harder for providers to move to a customer-centric model. However, our research shows that insurers are failing to capture potential customer lifetime value through repeat sales and improved customer retention. Customers want to build relationships with their provider as their main point of contact for their policy. Insurers must accept this challenge and adjust to live up to this expectation.

A common perception in the life industry is also that it is hard to influence persistency, particularly with largely intermediated sales in which insurers have limited contact with end customers. However, our survey highlights a significantly contrasting view. With the exception of a few outlier countries (e.g., India), customers' perception is that providers do very little to contact them at the point where they are lapsing. Insurance providers across the world make only a “fair or great effort” to retain customers in a minority of cases – 30% in Europe and only 19% in North America, the lowest region (Table 8). These are striking statistics, given that persistency is an increasingly important driver of value for the industry. Other Ernst & Young analysis of European insurers indicates that a 10% improvement in persistency can deliver a 2%-4% improvement in embedded value and as much as a 10% improvement in new business value.

So, in reality, providers can influence persistency. The survey finds that the top two reasons customers cite for switching providers are inability to meet changed needs and poor service – factors that providers can influence directly. Agent recommendation was also cited as a top three reason for customers switching providers. In certain instances, some commentators suspect “agent recommended” switching may be driven by the opportunity for the agent to generate additional commission rather than the need of the customer. However, improved consumer protection regulations in many markets are expected to result in a decline in this behavior.

In the non-life insurance sector, our survey shows that consumers are more likely to renew if insurers make more effort to retain their business, but that in many territories, customers' perception is that insurers fail to make this effort. 82% of respondents in Europe stated that little or no effort was made, compared with 78% in North America, 68% in South America, and 68% in Asia Pacific. This is particularly astonishing, given the relative product simplicity and ease of renewal process in the non-life sector – and the critical need to retain customers to make profits, given the price dynamics.

Conversely, we found that one area where many insurers believe they can differentiate themselves is claims service, but this does not actually make much of a difference in building loyalty. Poor claims service is definitely linked to low renewal rates, but good claims service is largely regarded as a given – “that’s what I paid for” – rather than a reason to renew (Table 9).

Another key area where insurers can encourage longer-term relationships is through rewarding loyalty. As discussed in Theme 1, in most territories consumers perceive the life and non-life insurance industry as lagging behind other consumer industries...
on this measure. There is a strong sense among consumers that insurers could do more to earn their trust and loyalty and to reward them for participating in long-term relationships. Consumers are used to many other industries rewarding their loyalty, such as supermarkets, airlines, or hotels, and they expect the same from insurers. Interestingly, we have observed many leading insurance loyalty propositions within the emerging markets of South Africa and eastern Asia, more so than in the more established European and North American markets.

So, what are the implications of these findings for insurers? Cross-selling is a key lever to boosting profitability. In mature markets where there are high levels of product penetration, so many “new” sales are simply businesses switching from one provider to another. Given the high costs of new customer acquisition, this means profitability is increasingly driven by retaining customers for longer and increasing revenue per existing customer – in other words, increasing customer lifetime value. This makes it essential for insurers to increase their cross-selling efforts and they need to work in partnership with distribution channels to achieve this. In some cases, a perceived lack of ownership of the customer relationship is the underlying issue. Providers must agree protocols for customer access and align economic incentives to obtain the best outcome for all parties – for example, by incentivizing multiple product sales – so that insurers and intermediaries cooperate to act in the best interests of the customer. In recent years, intermediaries have seized control of the customer relationship. Experience from other industries, such as automotive, shows us that a deep understanding of customer need by both manufacturer and distributor is critical to long-term product quality and customer service, for mutual benefit. Providers who make their contact, their products, and their reward structures right are winning repeat business.

Recognizing the value of the customer relationship is vital, both at the point of initial sale and over the customer lifetime. Some insurers are developing effective “sales-through-service” operations, which use routine servicing contacts to understand customers’ changing needs and offer carefully selected, relevant products where a need is identified. These operations are based on customer segmentation models, which identify customers who have the highest propensity to purchase. Such operations can be highly effective and improve customer satisfaction, compared to traditional unfocused “product push” cross-sell efforts.

There is a real opportunity for insurers to influence persistency and providers should consider an effective customer retention function essential. Retention activity should be supported by improved engagement with existing customers. The quality of customer contact across the life cycle has a material influence on retention. This needs to be backed by flexibly designed products that can respond to changing customer needs and financial incentives to reward customer loyalty – over the customer life cycle rather than just at the point of lapse. But the key to improving persistency is removing the reasons why customers consider leaving in the first place. Use of predictive models can be helpful to target customers, based on likelihood of lapse and the value of retaining them, but “test and learn” approaches are essential to work out which interventions are most effective. It is also important to manage the impact intermediaries have on persistency, through better management of intermediary channels, whether tied or independent. Some providers are segmenting distributors, based on a profitability model that takes into account the persistency of business introduced by intermediaries, refocusing their sales management effort on more profitable intermediaries.

### Learning from the voice of the customer – how should insurers respond?

Our research suggests that there is a lot that insurance companies can do to improve their businesses by listening to the voice of the customer. We believe that by getting the customer interaction right, there is an opportunity in every market to do a better job of attracting customers and strengthening relationships in order to drive growth (both revenue and profit).

In essence, in the life sector this means improving customer trust and confidence by putting the customer, rather than the

<table>
<thead>
<tr>
<th>Region</th>
<th>Good claim</th>
<th>Bad claim</th>
<th>No claim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>67%</td>
<td>46%</td>
<td>60%</td>
</tr>
<tr>
<td>Americas</td>
<td>66%</td>
<td>51%</td>
<td>67%</td>
</tr>
<tr>
<td>Asia Pacific</td>
<td>63%</td>
<td>52%</td>
<td>60%</td>
</tr>
</tbody>
</table>

Table 9: Percentage of respondents who said they were unlikely to change insurer, categorized by good claim, bad claim, or no claim – non-life sector
intermediary, at the center of the business model — a “customer-centric operating model” — and using customer data to develop deep insight into their needs, to offer the right product, at the right time, to the right customer, and to follow through with service that responds to their changing needs. This will require a balance between working with intermediary channels to ensure they focus on (and are remunerated for) where they really add value to the customer, and building stronger direct relationships with customers throughout the life cycle. Essential to delivering this will be providing a suite of simple, transparent products tailored to life stage, and making it easy to access relevant products and information, particularly online, but supported by trusted personal interaction where necessary.

In the non-life sector, this means delivering convenience and value by integrating online and offline channels seamlessly to meet changing customers’ needs over the product life cycle; in particular, making renewal simple and convenient for customers across whichever medium they choose, understanding how to personalize service and show customers they are valued, analyzing the cost-to-serve of the many customer micro-segments to understand how service expectations and profitability differ (particularly for managing risk selection and retention effort), and developing and managing insurance brand(s) to ensure they support the key value messages and communicate these effectively in the digital world.

These are clearly not easy actions to implement, particularly in mature economies where insurers have extensive legacy operations. The challenge is how to transition from existing product and traditional distribution models to one that delivers what customers are saying they really want today and in the future, and developing the critical new capabilities to enable this transformation.

**Becoming more customer focused: developing a customer-centric operating model**

So what does it mean to be customer-centric? A customer-centric organization is one with a deep understanding of who their customers are, what they value, and the contribution each customer makes to the profitability of the company — and that builds an operating model around this understanding. This includes designing business processes that recognize different customer segment needs, delivering a positive and seamless customer experience at every touchpoint across the whole customer life cycle, maintaining an active dialogue with customers (including obtaining and acting on feedback), and fostering a culture that places the customer at the heart of all decision-making.

For insurers seeking to improve their level of customer centricity, the first step of any transformation is to understand where you are today and where you want to be in the future, and by when. Assessing your organization against a customer-centric maturity model can help you understand the relative maturity of key capabilities such as segmentation, distributor management, service delivery, people, and product development. Defining and understanding what it really means to be “emerging” versus “advanced” or “leading” can be an enlightening exercise in itself, even before mapping your own company against these.

Indeed, understanding and communicating the scale and nature of the challenge ahead is the first step of any journey. Not all insurers will aspire to be market leaders. It is important to be realistic about what is achievable — leading customer centricity (the far right of the maturity model) is an aspiration, you may never realistically get there. However, the process of considering how to balance customer-centric principles with financial considerations is critical. For some, the trade-offs between the level of investment required and the benefits realized will determine how far they progress — and are often far harder to establish in mature, high market-share incumbents (where the scale of change to legacy operations is larger and perhaps the fear of disruption to existing business more acute).

The end objective is to transition away from traditional product- or channel-organizational structures to a more customer-focused operating model. Every organization’s target operating model (TOM) will be unique. However, they will have some common characteristics and design principles, such as using value-based advisor and customer segmentation, developing and utilizing customer insight, managing the end-to-end customer experience, providing multi-channel access, and integration and maintaining continuous conversations (interaction history) with advisors and customers. These “design principles” need to become the new “mantra” within the organization and rigorous adherence to them across all business units and functions is an essential component to driving a new enterprise-wide customer-focused culture, and will be a critical success factor of the customer transformation.
Typically, most insurers will encounter a number of challenges or barriers inherent in their current operating models. Organizational silos create the inability or unwillingness to share customer data, or cooperate across business units and functions (in effect, internal customer ownership wrangling). In our experience, this is particularly common in businesses that have traditionally been segmented along distribution channels, i.e., intermediary and direct. The traditional insurance customer engagement model presents significant barriers. In many cases, initial customer contact and, indeed, ongoing advice is controlled by an intermediary or agent – often provider engagement at this stage is purely administrative – not an ideal start to any relationship. Legacy systems and processes, not to mention poor quality or insufficient end-customer data, constrain many organizations in managing the end-to-end customer relationship – they are unable to monitor where they lose customers or why dissatisfaction occurs. Typically, a reactive stance to customer interaction (e.g., change of circumstances) is taken, but minimal insight around customer needs and preferences is captured and acted upon, beyond the initial transaction, with limited attempts to intervene proactively in the journey to add value. Even when customers are “orphans” (no longer have an active agent or intermediary), many organizations do not (or rather cannot) seize this opportunity to re-establish a relationship directly, with advice and service offerings. Perhaps surprisingly, many of these challenges also exist within insurers who have large direct businesses, where pre-conceived assumptions around customer behaviors, needs, and actions often prevent much-needed change.

Insurers will need a focused approach to overcome these challenges and barriers. Executive management ownership and commitment to driving customer centrality is imperative, supported by clearly defined customer-centric objectives and metrics across all business units and functions, with performance management and compensation aligned with customer-focused behaviors.

Using data as a source of competitive advantage: advanced segmentation and data analytics
Many insurers have grasped the concept of the need to segment and tailor their service proposition to different segments. In our experience, however, these are often based on socioeconomic factors (e.g., postcode, age, or other demographic information) and product holdings, for example, to determine a customer’s current and potential value.

Customer lifetime value is, of course, an important determinant of how to treat a customer. However, leading insurers are thinking differently about segmentation by grouping customers based on behaviors and values. Making sense of your customers in this way, by creating a micro-segmentation model that tries to understand what customers are really about – what kind of person they are, what they value most, and how they want to interact – and aligning this with a modular distribution model that allows customers to access your organization through multiple channels, when and how they prefer, will optimize customers’ experiences.

Data analytics have become critical to designing an effective customer experience. Insurers need to understand, for every segment what the true drivers of customer satisfaction are. What attributes of their experience (e.g., price, personal attention, response time, service calls, and language) will have the biggest impact on delivering a positive experience? In other words, what is the “unexpected” that delights the customer, versus the expected service that is only a “hygiene” factor (although, notably with claims, appreciating that those hygiene factors must be delivered, otherwise a disproportionately negative impact can occur). Equally, you need to understand what is not driving value – as these are opportunities to reduce cost.

Ultimately, the aim of this exercise is to reshape strategic decision-making by prioritizing investment in service factors that really deliver value for each segment, channel, and product combination. Without such analysis, it is almost impossible to mount a sensible business case.

Enhanced data analytics and predictive modeling capabilities can be applied in a number of other ways within an organization, to enhance customer focus: using behavioral engineering models to unpack the existing risk culture and its implications, as part of a transition to a more customer-centric and innovative culture; customer retention – why, where, and when you are losing customers; customer leads analysis to support development of a fully integrated distribution model; financial well-being modeling (of staff or customer base) to inform customer experience and advisor conversations on how to improve financial well-being of clients; and of course, claims management and leakage.
Harnessing digital as part of an integrated channel strategy to engage better with customers and lower costs

Our survey has clearly shown that, across both life and non-life sectors, most consumers want access to a mix of online and personal contact, throughout the product life cycle. Digital is not the panacea, but it is a critical enabler to delivering the experience customers want. The key challenge for insurers is how to harness the power of digital to engage directly with customers better and, of course, lower costs. Digital technology will simply complicate the process and frustrate customers if it is not joined up seamlessly with other channels, recognizing that existing channels and access points will also need to adapt and, in some cases, play an alternative role within the business.

The overall customer experience is influenced by customers' direct and indirect interactions with an organization – both physical and digital touchpoints. Both need to be managed effectively in order to maximize the experience (Figure 5).

Call centers are a prime example of an existing channel that will continue to play a critical role through the life cycle, in spite of the increasing influence of online channels, but will need to adapt its role within the business to respond to evolving trends and also drive greater value at lower cost. Currently, however, they are largely designed to operate in a telephone-dominated world with little data integration between other channels and are viewed as an expensive mechanism to service basic customer queries. Additionally, call handlers are not always equipped with the right skillset and tools to support a multichannel offering.

Of course, launching digital channels for many insurers with a heavily agency-based or intermediated distribution model is an enormous challenge, requiring knock-on changes to agent or intermediary segmentation, customer management and communication protocols, training, incentives, and rewards. There are massive implications in this new model for agents in particular. They will need to acquire new skills to integrate with the digital world – becoming more technologically savvy and connected themselves, and also playing a completely new role in terms of capturing data about individual customers and feeding this into the insurer’s analytics engine. They will be uniquely positioned on the front line to supply the kind of unstructured, but rich, customer data required.

An agency force with the “listening” capabilities and agility to play this role can become a key strategic asset. Customer satisfaction and engagement, and the value of the in-force book will become just as important as new business targets. Where once, agents were primarily in the business of producing sales volumes, they may soon be more focused on information collection and relationship management. They will both produce leads to the digital channels and in turn receive them from online browsers who decide they would like to speak with an agent.

Figure 5: Overview of direct and indirect interaction flow
Source: Ernst & Young.
Finally, from our experience, investment in digital is often driven by a fear of being left behind rather than a solid business case based on key performance indicators and recognized metrics. Without a solid business case, digital propositions will constantly face challenge and re-evaluation.

Customer-centric innovation and disruptive growth models

Although the above activities can be considered transformational to many in the industry, we believe they will only provide insurers with the ability to “catch up” and otherwise “protect the core.” To leapfrog competitors and generate significant growth, a new framework for innovation within the organization and a significant change in approach to strategic decision-making is required.

The three stages of innovation are depicted in Figure 6, across the dimensions of business models and competencies. Protecting the core is about getting the basics right, becoming more customer-centric and, where appropriate, adopting the leading practices that competitors or other industry sectors have demonstrated. However, it relies on largely existing competencies and business models and one could argue it will only ever be a catch-up strategy. Only genuine customer-centric innovation will produce significant growth and deliver competitive advantage. This will require the adoption of new approaches to embed innovation into the culture of your organization – in essence, moving from “innovate” to “becoming an innovative organization.” Furthermore, working through partnerships and alliances to develop completely new business models can generate “disruptive growth.”

In order to achieve customer-centric innovation, new cultural norms will need to be developed. Embedding a culture of innovation throughout the organization is essential. This will involve tapping into external discovery centers, labs, and industry “observatories” to gain insights from other sectors on a regular basis, as well as introducing internal “test and learn” processes that allow the business to generate ideas, test their feasibility, and pilot initiatives in a rapid and agile manner. Importantly, the new organizational culture needs to embrace the concept of failure as an acceptable by-product of experimentation (i.e., if you haven’t failed, you haven’t tried hard enough) that is often at odds with a typical insurance company approach to risk. Collaboration and empowerment of employees to spot customer improvements and act upon them must be the new norm, as well as physical spaces such as “innovation centers” to help develop and incubate ideas. One of the most significant cultural changes will be a shift from short-term to long-term strategic decision-making. The traditional focus in the industry (and one that is embedded in most performance management and compensation systems) has been around new business and market share rather than deriving value from the in-force book. This often means it is difficult for customer proponents within an organization to secure investment in enabling capabilities such as data analytics and customer MI, or long-term customer relationship-building activities, as these may not present the most compelling revenue-generating business cases in the short term. A change in mindset as to how to approach strategic business case development is required. Rather than the traditional “what will I get?”, organizations need to consider “what is the opportunity cost of not doing this?” in terms of financials, competitive positioning, and customer experience.

Above and beyond customer-centric innovation, some players are thinking even further outside the box to consider “disruptive growth” ideas – developing radical new customer value propositions, beyond the traditional confines of insurance, to meet the demands of customers in other industries. This requires new business models and new competencies, but can generate significant revenue streams by tapping into new profit pools.
Financial services is not the only industry under pressure—trends and megatrends are affecting business as usual across many other industries. This creates unique opportunities to enter other industries, establishing alliances and collaborating with new partners. Innovation will be the driver to identify and create these opportunities—such as “hybrid models.” Hybrid models take customer centricity to the extreme—considering what your customers may need on a more holistic basis, without being limited to traditional definitions of insurance provision. Examples include a Japanese insurer that offers home security audits as part of its home insurance policy, Progressive Insurance in the U.S. with its “concierge claims” offering, and a large Asian insurer focusing on consumer longevity and well-being, developing products that are priced according to loyalty scheme take-up (i.e., use of gyms, regular medicals, health product usage and balanced diet).

Conclusion
There is clearly a lot that insurance companies in all regions of the world can do to improve their businesses by listening to the voice of the customer. The combined impact of our findings is simple, but profound: insurers need to become as customer focused as other consumer businesses and deliver a genuinely customer-centric experience.

In a part online, part offline world, characterized by complex customer segmentation, even getting the basics right has become more complicated—but insurers must rise to this challenge. Customers want to have a trusted relationship and be rewarded for loyalty but insurers are manifestly missing a revenue opportunity here, not only from customer retention but also cross-sell and up-sell. Digital has the potential to make even insurance more interesting and engaging, however, it must be harnessed effectively as part of an integrated distribution strategy.

It could be argued that much of this is not ground-breaking. Customer-centric operating model, customer value management, multi-channel distribution, are all very easy to say but extremely hard to do in practice. Although we depict these as simply keeping pace, in our experience many insurers are far from being able to do these things well today. The inability to execute has been a key issue for the industry and, given the changing dynamics discussed in this paper, this is now a “strategic must,” not a “nice to have.”

Moreover, introducing a culture of innovation and customer-focused strategic decision-making will future-proof the core business while allowing insurers to redefine customer relationships and become a different insurer. In essence, listen to the voice of the customer and become the insurer of the future.
Do firm characteristics influence mutual fund performance? 
An empirical study for European mutual funds

Frank de Jong
Professor of Financial Markets and Risk Management, Tilburg University

Loes Wingens
Kempen Capital Management

Abstract
This study investigates the influence of fund management firm characteristics on mutual fund performance. Using a sample of European-domiciled open-end equity funds for the period 1998-2008, this study finds that the funds of private companies have performed better than the funds of public companies. The degree of focus, the volume of assets under management, and the number of funds offered by a fund management firm also have a positive impact on fund performance. In addition to these four firm characteristics, we find that non-European and U.K. fund managers had better-performing funds.

1 We would like to thank Joost Driessen and Jenke ter Horst for their helpful comments.
Do firm characteristics influence mutual fund performance?
An empirical study for European mutual funds

Introduction
Mutual fund selection is an important task of investment managers, and the performance of the mutual fund is a major factor in the selection process. Several determinants of mutual fund performance have been identified in the literature.

Managers’ skills, such as timing and selection abilities, have been found to positively influence fund performance. Fund characteristics, such as the fund size or the fund's total expense ratio, have also been found to influence performance. Most of the literature has treated mutual funds as if they are a stand-alone entity. This is not always appropriate, since a mutual fund typically belongs to a broader organizational structure, the fund management firm (the fund family). The fund family may impact the decisions of the fund and could potentially have a significant effect on the fund and its performance. Furthermore, there might be performance spill-over effects between funds within a family [see for example Ivkovic (2002) and Gaspar et al. (2006)]. Without an understanding of the role of the fund family we might ignore significant influences on the behavior and performance of individual funds. Literature on this subject however, is not yet very extensive.

In this paper, we contribute to the existing literature and investigate the influence of firm characteristics on fund performance in further detail. We investigate the impact of several factors on the performance of European mutual funds. We argue that the number of funds in the fund family, the ownership (public or private) of the firm, and the size and focus of the firm all influence fund performance.

Most of the existing mutual fund literature is focused on the U.S., which has the largest and perhaps most mature mutual fund industry in the world. Nonetheless, it would be very interesting to see whether the results found for the U.S. are also valid in Europe. The mutual fund industry in Europe is growing and becoming more and more important. Investment fund assets in Europe have doubled in size during the past decade, from €3,042 billion at the end of 1998 to €6,142 at the end of 2008.

At the end of 2008, the U.S. mutual fund industry represented 51% of the global mutual fund industry, but Europe was second, representing 33% of the world’s mutual fund industry. Despite its size and importance, the European mutual fund industry has received little attention in the finance literature. Some argue that this is due to the differences in the institutional setting of the industry in different European countries. Since the creation of the Economic and Monetary Union, however, European financial markets are becoming increasingly integrated and these institutional differences are disappearing.

Several papers have already addressed individual European countries but not Europe as a whole. Otten and Schweitzer (2002), Otten and Bams (2002) and Ramos (2009) find that, within European countries, the mutual fund industry is more concentrated than the U.S. Furthermore, the number of funds offered in Europe is much greater than in the U.S., leading to a much smaller average fund size in Europe. Overall, they find that the European industry is still lagging the U.S. in terms of total asset size, average fund size, and market importance. None of these papers investigates the impact of firm characteristics on fund performance.

The theory
Hypothesis 1: funds managed by a publicly owned company underperform those managed by a private company.

Whether or not the fund management firm is listed on the stock exchange is likely to influence fund performance. In a private firm the management are often also shareholders in the firm. This mitigates the agency conflicts between the manager and the shareholder. Ferris and Yan (2009) argue that the organizational form of a fund management firm influences its focus (short- or long-term). Publicly listed firms typically have dispersed ownership and are subject to mandatory disclosure requirements. Public companies’ stocks are traded on an active market and followed by analysts. These characteristics drive public companies to focus heavily on short-term performance. Privately held companies, however, typically have concentrated and dedicated owners. These firms are not subject to mandatory disclosures and there is no active market for their shares. These characteristics allow private firms to focus on maximizing long-run firm value. Consequently, the agency conflict between fund management and fund investors might be less severe in these companies.

The mutual fund industry is a unique industry for studying the influence of public listing on a firm’s performance. There are both public and private firms in the mutual fund industry but private firms do not have to disclose performance information.
Their open-end funds, however, are all listed and have to disclose the same level of information as the funds of listed companies. A private firm (especially in Europe) is often owned by a large shareholder (Pagano and Röell (1998)), which is generally more willing and able to monitor the management than a minority shareholder. Private firms are, therefore, likely to be better-monitored than public firms with dispersed ownership. This improved monitoring might also affect fund performance. Ferris and Yan (2009) test this hypothesis for the U.S. market and find that funds of privately held fund management firms significantly outperform those that are managed by a publicly held firm. Bogle (2005) identifies a “conglomeratization” of the U.S. fund industry. Large, often public, financial conglomerates have been taking over smaller, often private, players in the market. He argues that these conglomerates are more interested in increasing their asset base, building revenues, and enhancing their brand names than serving their clients’ best needs by improving performance. As a result, he finds that the relative returns of funds managed by private companies were better than 70% of their peers [see also Berkowitz and Qiu (2002)].

**Hypothesis 2: the number of funds that a fund managing firm has under management, has a negative effect on fund performance.**

The number of funds within a family might have an impact on fund performance. Bogle (2005) talks about the “marketingization” of the mutual fund industry. He argues that most major firms now create and market whatever funds they can sell. Massa (1998) also argues that fund proliferation can be seen as a marketing strategy, used by a fund family to make inter-fund comparisons harder and consequently influence the way investors “perceive” funds. Another reason for a fund family to increase the number of funds might be to “steal” market share from rivals and discourage new entrants. Other arguments for the increasing proliferation of funds offered include increasing the probability of creating a star fund instead of a good performance for each fund in the family [Nanda et al. (2006)] or allocating resources unevenly between funds within the same family and thereby treating some funds preferentially over others [Guedj and Papastaikoudi (2005)]. Bogle (2005) finds that the funds of U.S. firms focused on offering a few funds almost always outperformed the funds of firms that manage a lot of funds. However, in large fund families there may be greater possibilities to distort the incentives of fund managers. Indeed, Gaspar et al. (2006) find evidence for cross-fund subsidization. These observations lead to our next hypothesis.

**Hypothesis 3: the size of a fund management firm has a positive influence on fund performance**

The size, in terms of assets under management, of the mutual fund family might have an impact on the performance of its funds. Literature from the U.S. finds contradictory results for the influence of size on fund performance and is still inconclusive about this effect. Several papers argue that large families can benefit from economies of scale, such as lower transaction costs, more resources for R&D, and lower expense ratios. Chen et al. (2004) find evidence supporting the presence of economies of scale in the mutual fund industry. Their results show that family size (measured in terms of assets under management) has a (moderate) positive impact on fund performance. Ivkovic (2002) also reports a (moderately) statistically significant positive relationship between risk-adjusted fund performance and family size. Others believe that a large asset base erodes fund performance because of trading costs associated with liquidity or price impact. They argue that bureaucracies and related coordination costs cause diseconomies in large organizations. Consequently, belonging to a large fund family might negatively impact fund performance. In accordance with this argument, Massa and Zhang (2009) study the impact of hierarchical structure on performance. They find results suggesting that funds belonging to less hierarchical structures outperform their peers. Likewise, Dermine and Röller (1992) find significant diseconomies for larger fund management firms in the French mutual fund Industry. However, since the European fund industry is still lagging the U.S. in terms of assets, with average firm and funds sizes being smaller in Europe, we would expect economies of scale to be more important in the case of Europe.

**Hypothesis 4: the degree of focus of a fund management firm positively influences fund performance.**

The strategy of the fund management firm could have an impact on the performance of its funds. Some firms choose to offer a diversified range of funds while others choose to focus on offering funds in a few (niche) strategies. A high degree of product focus may allow a firm to specialize in particular products, resulting in higher quality and/or lower costs. Sigkelkow (2003) investigates this issue for the U.S. market and finds results indicating that the performance of a mutual fund improves with the fund family's
degree of focus in that fund’s category. U.S. mutual funds that belong to more focused fund providers outperform similar funds offered by more diversified providers. On the other hand, offering funds in several different categories can improve the visibility of the firm in the market and thereby increase possible inflows into its funds. Fund families that pursue such a differentiation strategy could be more concerned with gaining market share than achieving good performance. A higher degree of differentiation might then imply lower performance [see, for example, Khorana and Servaes (2007)]. Another reason why fund families choose to follow a diversified strategy might be to increase the possibility of creating a star fund. Nanda et al. (2004) show that families with higher variation in investment strategies across funds (a diversified strategy) are more likely to generate star performance. But their results also show that portfolios of fund families with high variation in investment strategies across funds significantly underperform portfolios of fund families with low variation in investment strategies across funds (focus strategy).

Fund performance
Data
We have gathered net return data for open-ended equity funds domiciled in the E.U. for the period 1998-2008 with at least 5 years of data. We chose the period 1998-2008 to get a sufficient sample size and sufficient data. We have divided the sample into two periods, 1998-2008 and 2004-2008 to see whether the results are consistent over time. Morningstar Direct is the primary source of data at the fund level. The domicile countries for which we gathered data are the 11 early entrants of the EMU plus the U.K. and Switzerland. In terms of Net Assets, these countries together represent 94% of the European mutual fund industry. Because funds often have several share classes that are supported by the same underlying portfolio, we have included only the oldest share class of each fund in our sample. The dataset gathered from Morningstar includes: fund name, firm name, net fund returns (in euro), fund domicile, and fund category. We have only included funds with at least five years of data to get reliable (and significant) values for the alphas. This selection results in 5,167 funds belonging to 579 firms. Our sample represents about 42% of the current number of open-end equity funds in Europe listed in Morningstar. To correct for regional differences in market and factor movements and thus, alpha generation, we have divided the sample over nine regions in which the funds invest. Table 1 displays the summary statistics of the fund data for the 5- and 11-year period per region.

<table>
<thead>
<tr>
<th>Region</th>
<th>With data from Jan 2004</th>
<th>With data from Jan 1998</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of funds</td>
<td>Average return (%)</td>
<td>Standard deviation (%)</td>
</tr>
<tr>
<td>Asia excluding Japan</td>
<td>204</td>
<td>0.30% 5.8%</td>
</tr>
<tr>
<td>Asia including Japan</td>
<td>127</td>
<td>-0.02% 5.0%</td>
</tr>
<tr>
<td>Japan</td>
<td>286</td>
<td>-0.22% 5.0%</td>
</tr>
<tr>
<td>Emerging markets</td>
<td>186</td>
<td>0.51% 6.6%</td>
</tr>
<tr>
<td>Europe including U.K.</td>
<td>1,115</td>
<td>0.04% 4.6%</td>
</tr>
<tr>
<td>Europe excluding U.K.</td>
<td>1,070</td>
<td>0.16% 4.9%</td>
</tr>
<tr>
<td>U.K.</td>
<td>434</td>
<td>-0.24% 4.7%</td>
</tr>
<tr>
<td>Global</td>
<td>1,185</td>
<td>-0.17% 4.2%</td>
</tr>
<tr>
<td>U.S.</td>
<td>560</td>
<td>-0.43% 4.3%</td>
</tr>
<tr>
<td>Total</td>
<td>5,167</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 - Monthly net fund return summary statistics
This exhibit displays the summary statistics for the monthly fund return data. All data in this exhibit are monthly data. The returns are net including all dividends. Average returns are per month over the period from 1-01-2004 to 31-12-2008 and the period from 1-01-1998 to 31-12-2008.

The number of funds that have existed over the past 11 years is only half of the number of funds that have existed during the past 5 years. This is in line with the doubled size of the European industry, in terms of assets under management, over the past 11 years.

---

4 See Elton et al. (2001) for a comparison between the accuracy of Morningstar and the CRSP database for the U.S. They find that even the CRSP is not completely bias-free because it suffers from omission bias.

5 The countries included in the sample are: Austria, Belgium, France, Germany, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Switzerland, and the U.K. We do not include Finland, since its fund industry size, in terms of net assets, is negligible.

6 The requirement that there are at least five years of data may induce some survivorship bias in the alphas and may bias the estimates in Table 1 upwards. However, as long as the bias is not a function of the fund family characteristics this will not bias the slope coefficients in the regression of the alphas on these characteristics.
Methodology
To calculate fund performance we employ the Fama-French (1993) three-factor model:

$$ R_{it} - R_f = \alpha_i + \beta_{i1}(R_{mt} - R_f) + \beta_{i2}\text{SMB}_t + \beta_{i3}\text{HML}_t + \epsilon_{it} $$

(1)

where $R_{it}$ is the return of fund $i$ at time $t$, $R_f$ is the risk-free rate, $R_{mt}$ is the return on the market, $\text{SMB}_t$ is the return on the size factor, and $\text{HML}_t$ is the return on the book-to-market factor. The market and factor movements are not the same for every region (e.g., a fund that invests in emerging markets is likely to have low correlations with the U.S. factors). We, therefore, construct separate factors for each of the nine regions in which the funds invest. We construct $R_{mt} - R_f$, $\text{SMB}$, and $\text{HML}$ factors using monthly return data from DataStream. The risk-free rate is the monthly three-month Euribor rate.\(^7\) The market factor is the MSCI index for that specific region. The $\text{SMB}$ factor is constructed by subtracting the regions’ MSCI Large-cap index returns from the regions’ MSCI Mid-cap returns\(^8\). The $\text{HML}$ factor is constructed in a similar manner, by subtracting the regions’ MSCI Value index returns from the regions’ MSCI Growth index returns. All these factors are calculated using net returns, including dividends.

Results
Using the monthly fund returns and the monthly factor data from above, we have run regression (1) for each individual fund in the sample. We run this regression once for all the funds in each region with at least 5 years of data and again for all the funds in the sample with at least 11 years of data. This results in a range of 5-year fund-specific alphas and a range of 11-year fund-specific alphas. We use these alphas as a measure of risk-adjusted performance. The results are summarized in Table 2.

For the 5-year sample, the monthly alpha is -0.111%, this is -1.33% annually. In the 11-year sample, the monthly alpha is -0.033% or annually -0.398%. The coefficient for the market factor is always positive and highly significant (a total sample average of 0.971 for the 11-year period and 0.978 for the 5-year period). The $\text{SMB}$ coefficient is also positive and often significant. The $\text{HML}$ factor is almost always significant and negative for the 11-year period (a significant total sample average of -0.101) but not for the 5-year period (an insignificant total sample average of -0.083). For both periods the adjusted $R^2$ values are relatively close to 1, which indicates that the model can explain a large part of the variations in the fund returns. The alphas gathered from these regressions are used as a measure of fund performance. We have also performed regressions including the Carhart (1997) momentum factor. Since the momentum factor was almost never significant (except for the U.S. sample) we do not report those results here. Table 3 summarizes the sample of alphas per region for the 5-year period. For the 11-year period the results are not reported here but available upon request from the authors. For the 11-year period, the percentage of positive alphas is higher in all regions (41% for the total sample) and more significant (T-stat of 0.841 for the total sample). As can be seen, there are, on average, more negative alphas (total sample average of 71% for the 5-year period and 59% for the 11-year period) and they are more statistically significant than the positive alphas. These

\(^7\) Because all the funds in our sample are European-domiciled funds, the Euribor is a good proxy for the risk free rate.

\(^8\) We use the Mid-cap index and not the Small-cap index because the latter returns are not available for every region and often represent only very small companies. In this case the Mid-cap index is a better proxy.

<table>
<thead>
<tr>
<th>Number of funds</th>
<th>5-year average</th>
<th>11-year average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire sample averages</td>
<td>5,167</td>
<td>2,216</td>
</tr>
<tr>
<td>Alpha</td>
<td>Coefficient</td>
<td>Abs. T-Statistic</td>
</tr>
<tr>
<td>% significant positive</td>
<td>-0.111%</td>
<td>1.191</td>
</tr>
<tr>
<td>% significant negative</td>
<td>0.81%</td>
<td>2.53%</td>
</tr>
<tr>
<td>Coefficient</td>
<td>17.51%</td>
<td>12.59%</td>
</tr>
<tr>
<td>SMB</td>
<td>Coefficient</td>
<td>Abs. T-Statistic</td>
</tr>
<tr>
<td>% significant positive</td>
<td>0.259</td>
<td>2.390**</td>
</tr>
<tr>
<td>% significant negative</td>
<td>44.09%</td>
<td>60.51%</td>
</tr>
<tr>
<td>Coefficient</td>
<td>6.50%</td>
<td>4.11%</td>
</tr>
<tr>
<td>HML</td>
<td>Coefficient</td>
<td>Abs. T-Statistic</td>
</tr>
<tr>
<td>% significant positive</td>
<td>-0.083</td>
<td>1.384</td>
</tr>
<tr>
<td>% significant negative</td>
<td>6.14%</td>
<td>11.10%</td>
</tr>
<tr>
<td>Coefficient</td>
<td>17.90%</td>
<td>35.33%</td>
</tr>
<tr>
<td>SVM</td>
<td>Coefficient</td>
<td>Abs. T-Statistic</td>
</tr>
<tr>
<td>% significant positive</td>
<td>0.978</td>
<td>22.728***</td>
</tr>
<tr>
<td>% significant negative</td>
<td>100%</td>
<td>30.273***</td>
</tr>
<tr>
<td>Coefficient</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 2 - Monthly four-factor results for the period 2004-2008 and 1998-2008
This exhibit summarizes the Fama-French regression results for all regions. All data in this exhibit are monthly data. Superscripts *, **, *** denote statistical significance at 10%, 5%, and 1% respectively.
results corroborate the findings of Carhart (1997), Ferris and Yan (2008) and Chen et al. (2004).

Firm characteristics

Data

For each fund in the sample we have gathered the relevant family characteristics to investigate the impact of these variables on fund performance. When we look at the firm characteristics we consider the fund family. For the number of funds per family, we take the total number of open-end (so not only equity but also real-estate, money market, and bond funds), funds that a firm has registered in Morningstar Direct. According to EFAMA and ICI statistics, the total number of mutual funds at the end of 2008 was approximately 68,500. In Morningstar, there are 63,172 open-end funds registered. This means that Morningstar covers about 92% of the world’s mutual fund industry. Data on whether a firm is public or private is gathered from Bloomberg. If a firm is listed on any stock exchange according to Bloomberg it is considered public. If a company went public (or private) during the sample period, it will be marked with the label that it had during the longest period in the sample. For example, ABN AMRO is labeled public because it was listed during the largest part of the sample. The size of a firm is measured in terms of assets under management (AUM) in 2004 and is gathered from Lipper FMI. The year 2004 is in the middle of the sample-period and the last date at which a firm in the sample could have been incorporated. Consequently, each firm in the sample should at least have AUM data from 2004 onwards. To measure a firms’ degree of focus we employ a Herfindahl-like index, considering the number of categories that a firm covers and the number of funds it offers in each category:

\[ H = \sum_{i=1}^{N} s_i^2 \quad s_i = \frac{\text{Number of funds in category } j \text{ by firm}}{\text{total number of funds held by firm}} \]  

By using this measure, we not only consider the number of funds per category versus the total number of funds, but also the number of categories in which a firm operates. Unfortunately, Morningstar does not provide net asset data for each fund. We cannot, therefore, measure the asset-weighted degree of focus. To determine the number of categories in which a firm operates, we use the Morningstar Global categories. In total, Morningstar identifies 62 global categories. The distribution of the degree of focus for our sample is presented in Figure 1. Where the degree of focus is 100% (approximately 7% in our sample) the firms have only one category and/or fund. These firms are inherently as focused as possible, thus 100%.

Table 3: Monthly alpha summary statistics

<table>
<thead>
<tr>
<th>5-year averages</th>
<th>Positive alphas sample</th>
<th>Negative alphas sample</th>
<th>Significant alphas sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percentage</td>
<td>Average</td>
<td>T-stat</td>
</tr>
<tr>
<td>Asia excluding Japan</td>
<td>21%</td>
<td>0.130%</td>
<td>0.703</td>
</tr>
<tr>
<td>Asia including Japan</td>
<td>28%</td>
<td>0.279%</td>
<td>0.710</td>
</tr>
<tr>
<td>Japan</td>
<td>11%</td>
<td>0.146%</td>
<td>0.414</td>
</tr>
<tr>
<td>Emerging markets</td>
<td>24%</td>
<td>0.146%</td>
<td>0.501</td>
</tr>
<tr>
<td>Europe including U.K.</td>
<td>35%</td>
<td>0.174%</td>
<td>0.725</td>
</tr>
<tr>
<td>Europe excluding U.K.</td>
<td>35%</td>
<td>0.208%</td>
<td>0.695</td>
</tr>
<tr>
<td>U.K.</td>
<td>35%</td>
<td>0.150%</td>
<td>0.707</td>
</tr>
<tr>
<td>Global</td>
<td>25%</td>
<td>0.170%</td>
<td>0.683</td>
</tr>
<tr>
<td>U.S.</td>
<td>20%</td>
<td>0.133%</td>
<td>0.565</td>
</tr>
<tr>
<td>Total sample</td>
<td>29%</td>
<td>0.176%</td>
<td>0.687</td>
</tr>
</tbody>
</table>

This exhibit summarizes the sample of 5-year alphas. All data in this exhibit are monthly data for the period of 1-01-2004 to 31-12-2008. Percentage is the percentage of positive/negative alphas, average is the average coefficient for alpha within the sample of positive/negative alphas and T-stat is the average T-statistic in the sample of positive/negative alphas. Superscripts *,**,*** denote statistical significance at 10%, 5% and 1% respectively.

9 This was the number of open-end funds registered in Morningstar Direct at 16-07-2009.
Since several studies have indicated that the European fund industry is still lagging the U.S. fund industry, we have also gathered data for country of origin of the firm to see whether European funds underperform non-European funds. We distinguish between three groups of funds: Non-European, U.K., and Continental European. Table 4 displays the summary statistics for the firm-specific data.

There is a large difference between the mean (63) and the median (13) number of funds offered by a firm, which indicates the sample distribution is skewed. Consistently, most master groups in the sample are relatively small, offering a low number of funds. There are, however, several large financial conglomerates in the sample that offer many funds and have an enormous amount of assets under management. Guedj and Papastaikoudi (2005) find an average number of funds per family of 5 for a sample of 479 families in 2002 for the U.S. This is only a fraction of the average of 63 that we have found for Europe. The fact that European companies offer a higher number of funds has already been reported by Otten and Bams (2002).

Another interesting observation is the high percentage of private firms in the sample (73%). This is in contrast with the U.S., where only a very small part of the fund management industry is still private. Bogle (2005) has a sample of 54 asset managers in the U.S., of which only 13 are private (see also Ferris and Yan (2009)).

In Europe the consolidation of the industry is only just getting started. In the European fund industry there are still many small and private independent asset managers present. In the U.S. most small/private asset managers have been taken over by large corporations or were listed on a stock exchange. This is consistent with Pagano and Röell (1998), who argue that in Europe, fewer companies go public than in the U.S. The market share (in terms of number of funds) of public companies, however, is quite similar in both markets. Ferris and Yan (2008) examine a sample of U.S. mutual funds for the period 1992-2004. In their sample, 65% of the funds belong to public companies. In our sample this percentage is 60.5%.

Assets under management exhibit the same pattern as the number of funds, where the difference between the mean and the median is quite substantial. Khorana et al. (2005) find a global mean AUM (in U.S.$ billions) of 189.50 (±211.52) and a median of 9.31 (±10.39) in 2001; also, a quite substantial difference between the mean and the median.
Do firm characteristics influence mutual fund performance?
An empirical study for European mutual funds

When considering the focus of the master groups, most firms in our sample are relatively focused. These findings are all an indication of the fact that the European fund industry is different from the U.S. and may be less consolidated, as has already been reported in several studies [Otten and Schweitzer (2002); Otten and Bams (2002); Ramos (2009)].

Furthermore, 11% of the firms that operate in our sample are non-European. These are often the larger U.S. fund managers who offer so-called offshore funds in Europe.

Methodology
With the fund-specific alphas and the firm-specific data for the firm that the fund belongs to, we estimate the impact of firm characteristics on fund performance, using the following cross-sectional regression:

\[
\alpha_{ij} = \beta_0 + \beta_1 \text{Nfunds}_j + \beta_2 \text{Dpublic}_j + \beta_3 \text{DegreeFocus}_j + \beta_4 \text{DNon-European}_j + \beta_5 \text{DUK}_j + \epsilon_{ij}
\]

Where, \(\alpha_{ij}\) is the four-factor alpha for fund \(i\) belonging to firm \(j\), \(\text{Nfunds}_j\) is the logarithm of the number of funds offered by firm \(j\), \(\text{Dpublic}_j\) is a dummy variable equal to 1 if the firm is public, \(\text{AUM}_j\) is the logarithm of the assets under management of company \(j\) in 2004, \(\text{DegreeFocus}_j\) is a measure of the degree of focus of company \(j\), \(\text{DNon-European}_j\) is a measure of the degree of focus of company \(j\), and \(\text{DUK}_j\) is a dummy variable equal to 1 if the master group’s headquarters are located outside Europe, and \(\epsilon_{ij}\) is a dummy variable equal to 1 if the master group’s headquarters are located in the U.K. Before running this regression, we have checked the cross-correlations between the factors. The results from this check showed that the number of funds and assets under management were too correlated to be included in the same regression. To overcome these multicollinearity issues, we have split the regression above in two separate regressions. Table 5 displays the results for the 5- and 11-year samples. As the errors terms of the regressions could be correlated within the same firm, we have used cluster-robust standard errors. The T-statistics reported in Table 5 are therefore cluster-robust.\(^{10}\)

Results
The coefficient for log number of funds is positive in all of the regressions and often significant. On a monthly basis the coefficient for log number of funds with 5-year alphas is 0.00036, which implies that for a firm with 100 funds the monthly alpha increases by 0.078% (0.938% annually) compared to a one-fund firm. An explanation for the fact that other U.S. studies have found the reverse relationship to the number of funds could be that the U.S. fund industry is more consolidated than Europe’s. In the U.S., the average fund size is much larger than in Europe but the total number of funds offered is much smaller. In the U.S., several large conglomerates dominate the market whereas in Europe there is a large number of small independent fund managers. The fact that in

\(^{10}\) To calculate cluster robust standard errors we have used the approach of Cameron et al. (2006) for one-way clustering.
Europe a higher number of funds is preferred might indicate that the average fund size in Europe is too small. The firms offering a small number of funds might simply be too small and therefore inefficient. This issue could be solved by offering more funds or by merging one fund into another, thereby increasing the fund’s asset base.

In terms of firm size, as measured by assets under management, we have found a significant positive relationship with fund performance. This implies that the funds belonging to firms with large assets under management have performed better. Interestingly, while we obtain clear and often significant results for the impact of size on fund performance, the literature is still inconclusive about this impact. We have found a coefficient for log assets under management, its monthly alpha increases by 0.00013 for the 5-year alphas. This finding is consistent with the findings of Ferris and Yan (2009), Chen et al. (2004), and Ivkovic (2002).

As stated in hypothesis 4 and consistent with previous research, funds belonging to focused firms have performed better than those that belong to less focused and more diversified firms, though the coefficient is small and only moderately significant. For the sample of all 5-year alphas the average coefficient over the two regressions for the degree of focus is small and only significant in the first regression. In the case of the first regression, when a firm has a 50% degree of focus the monthly alpha increases by 0.0150% (0.19% annually). Other studies use a similar Herfindahl-like index and also find a positive impact of focus on fund performance [Siggelkow (2003); and Khorana and Servaes (2007)]. Nanda et al. (2004), who use variations in investment strategies across funds as proxy for focus, also find a similar Herfindahl-like index and also find a positive impact of focus on fund performance [Siggelkow (2003); and Khorana and Servaes (2007)].

In addition to the variables used to test our four hypotheses, we have also included two dummy variables to see whether there is a significant difference between the performance of funds managed by Anglo-Saxon (non-European and U.K.) companies and those managed by European companies. A company is labeled “non-European” if the headquarters of the firm are located outside of continental Europe and are labeled “U.K.” when the companies’ headquarters are located in the U.K. Interestingly, the coefficient for the non-European dummy variable is positive and very significant.
This indicates that funds belonging to European firms have performed worse than their non-European peers, predominantly U.S.-based institutions. In the case of U.K. firms, we also find that they perform better than European firms, though the effect is smaller and only moderately significant.

When we investigate this issue further and run the regressions separately for each region we find that, even for funds domiciled in Europe and investing in Europe, the dummy for non-European is still significant and positive. Furthermore, the dummy for non-European is also positive and often significant for every other region except Japan, implying that non-European managers outperform European managers in every region except Japan (Table 6).

Conclusions

Using a sample of 5,163 open-end European-domiciled mutual funds belonging to 579 firms (families), we find that firm characteristics indeed influence fund performance. With a Carhart four-factor model for nine separate regions, we calculate fund alphas as a measure of fund performance. With the resulting sample of alphas we have performed several cross-sectional regressions including each fund's firm characteristics. Our results indicate that, in contrast with U.S. research, in Europe a larger number of funds per family improves fund performance. This might indicate that in Europe the average firm is not large enough in terms of number of funds or that the average fund size is simply too small to benefit from economies of scale. This is reflected in the large number of small independent asset managers in our sample. In line with previous research in the U.S. and Canada, the funds of privately managed companies have performed better in Europe as well. Whereas existing literature for the U.S. finds contradicting results, we have found a clear positive impact of family size on fund performance. Possible explanations are a larger amount of resources available for research, cost benefits with transaction costs, more visibility in the market, and hence an improved ability to attract skilled managers. Consistent with previous U.S. research, we have found a positive (but only moderately significant) effect of focus on fund performance. The results imply that firms who have a set of related funds have better performance than firms who offer a broad range of different products and/or cover many categories. Interestingly, we have found that the funds of non-European firms have performed significantly better than the funds of European firms. This might indicate that the European industry is not yet mature and there is still room and need for growth, to benefit from economies of scale.

References

Guedj, I., and J. Papastaioudi, 2005, “Can mutual fund families affect the performance of their funds?,” working paper
Abstract
Fund transfer pricing (FTP) is fundamental for evaluating the profitability of deposits and loans. Following the global banking crisis, this paper seeks to draw attention to five issues that have been previously ignored: rationing on the interbank market, the creation of a Basel III contingency liquidity buffer, the necessity to adjust FTP to the credit riskiness of specific assets of the bank, the need to include a liquidity premium in the case of long-term funding, and finally, the choice of a consistent methodology to incorporate the credit spread on the bank's own debt due to the perceived risk of bank default.

1 This paper significantly expands on the first paper on fund transfer pricing (Dermine (2012)).
Introduction
Fund transfer pricing (FTP) is used by bankers to evaluate the profitability of deposits and loans and for pricing. It is used by academics and antitrust authorities to evaluate the degree of competition in banking markets. The challenge, as far as on-balance sheet banking is concerned, is as follows. When one evaluates the profitability of deposits, the costs are known, namely the interest paid on deposits and the operating expenses associated with deposits, such as employee time and IT. However, determining the return on deposits is more problematic because deposits are used to finance various types of assets: consumer loans, corporate loans, interbank assets, bonds, and fixed assets. Revenue — known as the fund transfer price — must be determined to remunerate deposits. For loans, the problem is symmetrical: one knows the return on loans, that is, the interest income net of expected bad debt expense, but not its funding costs. The reason is that banks use several sources of funds to finance assets: demand deposits, savings deposits, time deposits, corporate deposits, interbank deposits, subordinated debt, and equity. Again, there will be a need for a specific fund transfer price to evaluate the cost of funding loans. Appropriate identification of the FTP is fundamental for pricing commercial products, performance evaluation, bank strategy design, and assessment of the level of competition in banking markets.

The foundation approach, used throughout the banking world, is presented below. It covers two cases: products with fixed and undefined maturities. We argue that, as a result of the global financial crisis, attention should be given to five issues that have previously been ignored: rationing on the interbank market, the creation of a Basel III contingency liquidity buffer, the necessity to adjust fund transfer price to the credit riskiness of specific assets of the bank, the need to include a liquidity premium in the case of long-term funding, and finally, the choice of a consistent methodology to incorporate the credit spread on the bank’s own debt due to the perceived risk of bank default. We conclude that an advanced approach to FTP must be adopted by banks.

Fund transfer pricing, foundation approach
The foundation approach to FTP is presented first. It includes two cases: deposits or loans with fixed maturity and products with undefined maturity.

![Figure 1 - The separation theorem](image-url)

Foundation approach: products with fixed maturity (FTP1)
The foundation approach to FTP is represented in Figure 1. The horizontal line represents the market rate, that is, the interest rate observed on the interbank market (LIBOR). The line is horizontal as the interest rate is set on the major international markets and is independent of the volume of transactions initiated by the bank. The two other lines represent the marginal income on loans and the marginal cost of deposits. As a bank wishes to increase its loan portfolio, the expected income from an additional dollar of loan — the marginal or incremental income — will go down because the bank needs to reduce the interest rate to attract the additional dollar of loan, or because the bank accepts a loan of lower quality. Similarly, the cost of collecting an additional dollar of deposits — the marginal or incremental cost of deposits — will go up because the bank needs to raise the deposit rate to attract the additional dollar of loan, or because the bank needs to open more expensive branches in remote areas. In Figure 1, the optimal volume of deposits, D_opt, is reached when the marginal cost of deposits is equal to the interbank market rate. One would not want to go beyond D_opt because the bank needs to raise the deposit rate to attract the additional dollar of deposits or because the bank needs to open more expensive branches. In the case of loans, the optimal volume, L_opt, is reached when the marginal revenue from loans is equal to the marginal investment return, the market rate. One would not want to increase the loan portfolio beyond L_opt because the incremental income on the new loan would be lower than the return available on the money markets.

2 In some countries with illiquid interbank markets, the relevant market rate is the interest rate on government bonds.
3 We ignore reserve requirements with Central Bank which reduce the revenue earned on deposits.
should correspond to the maturity of the fixed-term product. For short maturities of up to one year, one frequently uses the interbank market rates, and for longer fixed-rate maturities, one uses the swap rates. Note that there is a separation between the lending and funding decisions. The separation theorem states that loans and deposits must be priced with reference to the market rate and these decisions are independent of one another. The difference between the optimal volumes of deposits and loans (DOPT - LOPT) is the net position in treasury assets, bonds or interbank assets. In Figure 1, it is positive, with deposits exceeding the volume of loans. The bank is a net lender in the money market. But it could be negative when the bank is a net borrower, as illustrated in Figure 2. In this case, the difference between the volume of loans and deposits (LOPT - DOPT) must be funded in the money markets.

Foundation approach: products with undefined maturity (FTP2)
In the foundation approach to FTP, the relevant maturity for the marginal return is the maturity of the deposit or loan. A two-year deposit should be priced against the two-year matched maturity market rate. There are several well-known cases, such as demand or savings deposits, for which the contractual maturity (very short, as withdrawable on demand) is different from the effective economic maturity. Indeed, many deposits are fairly sticky, with a longer effective maturity. An ad hoc method is to divide the stock of deposits into permanent core deposits and volatile deposits. The effective maturity is a weighted average of the two maturities, long and short. As shown below, dynamic considerations have to be taken into account to reflect the impact of the current volume of deposits on future volumes and interest margins.

Consider the case where the volume of deposits in Year 2, \( D_2 \), is a function of not only the deposit rate paid in that period, \( d_2 \), but also of the volume of deposits collected in Year 1, \( D_1 \).

Given the dynamic consideration, \( b_1 \) and \( b_2 \) denoting the market rates in Year 1 and Year 2, the present value of future profits over two years, evaluated at end of Year 1 is equal to:

\[
\text{Present Value of Profits} = \left( b_1 \cdot d_1 \right) x D_1 + \frac{\left( b_2 \cdot d_2 \right) x D_2 \cdot \left( D_2 \cdot D_1 \right)}{1 + b_2}
\]

The marginal income on one dollar of deposits collected in Year 1 is equal to:

\[
\text{Marginal Income} = \left( b_1 \cdot d_1 \right) + \frac{\left( b_2 \cdot d_2 \right)}{1 + b_2} \cdot \frac{\partial D_2}{\partial D_1} - d_1
\]

With \( \alpha \) = persistence factor = \( \frac{\partial D_2}{\partial D_1} \).

The marginal income on one dollar of deposits collected this year includes two parts: the market rate in Year 1 earned on this dollar of deposits and the present value of the profit in Year 2 resulting from the persistence of that additional unit of deposit. The existence of a lag in the deposit or loan function resulting from loyalty of clients or the rigidity of interest rate (due to marketing reason and cost of changing the pricing menu) forces one to analyze the FTP over multiple periods.

It should be noted that the choice of a fund transfer price in a multi-period setting applies not only to demand or savings deposits

---

4 Matching maturities not only meets intuition in the search for a relevant opportunity cost, it also protects the commercial units against interest rate (or currency) fluctuations. Interest rate (or currency) mismatches are transferred to the ALM department which is in charge of managing these sources of risk (Dermine (2009)).

5 The swap rate gives the long-term cost of a rollover of short-term interbank funding that is hedged with a swap. This is likely to differ from the actual cost of funding the long-term asset with a long-term debt that would include a liquidity or credit spread. The use of a swap rate is appropriate when the bank performs the traditional function of maturity transformation, funding long-term assets with short-term debt. The case of maturity matching – long-term assets funded with long-term debt – is analyzed separately later.

6 Note that we ignore the cost of the action needed to attract the extra dollar of deposits, such as additional marketing expenses or increase of the deposit interest rate.

7 As shown formally in Dermine (2009), it is only in the case of a fixed rate on deposit (\( d_1 = d_2 \)) and a persistence factor of \( \alpha = 1 \) that the fund transfer price is equal to a two-year maturity market rate.
The use of a matched-maturity market rate forms the core of the foundation approach to FTP. In the case of dynamic considerations, such as lag in the supply/demand function or interest rate rigidity, one must use a dynamic multi-period approach. The foundation approach has been used extensively around the world.

During the economic expansion observed in many countries from 2003 to 2007, banks expanded their loan books massively, relying on market funding when the loan books exceeded the deposit base. The loan-to-deposit ratio exceeded 100% in several countries, including the U.K., Greece, Portugal, Spain, Brazil, Peru, and Vietnam. The underlying assumption was that market-based funding, such as interbank deposits or certificates of deposit (CDs), would always be available to finance their illiquid loans. The global financial crisis, which started in July 2007, proved this assumption wrong. The interbank markets froze after Bear Stearns’ announcement that it had to refinance two of its structured investment vehicles. Banks rediscovered liquidity risk and the need to manage and price this source of risk. This has led to a review of the FTP methodology.

**Fund transfer pricing, advanced approach**

The global financial crisis has drawn attention to the fact that an advanced fund transfer pricing system is needed in specific circumstances. Five cases are analyzed below: rationing on the interbank market, the need to build up a Basel III contingency liquidity buffer, the need to adjust to the credit-riskiness of specific assets of the bank, the liquidity premium on the term structure of interest rates in the case of long-term funding, and the need to take into account the bank’s credit spread when its default risk is not trivial.

**FTP and rationing on the interbank market (FTP3)**

In Figure 2, the optimal volume of loans is larger than deposits. This implies net borrowing on the interbank market. The recent crisis has raised our awareness of the fact that interbank funding is volatile. In countries facing a sovereign debt crisis, such as Greece, Ireland and Portugal, local banks have been shut out of the international money markets. In such a situation, loans cannot exceed the volume of deposits. It must be noted that, although rationing on the interbank market is quite novel in OECD countries, it has frequently been observed in emerging markets with less developed liquid interbank markets.

In Figure 3, the starting position is identical to that of Figure 2. If deposits and loans are chosen with reference to the market rate, the optimal volume of loans exceeds that of deposits, the shortage of funds being met on the interbank market. As this will not be feasible in a situation in which no lender is willing to provide funds at a reasonable market rate, the bank needs to increase the volume of its deposits and/or reduce the volumes of its loans. The optimal point is achieved where the marginal income on loans line intersects with the marginal cost of deposits line. As this will not be feasible in a situation in which no lender is willing to provide funds at a reasonable market rate, the bank needs to increase the volume of its deposits and/or reduce the volumes of its loans. The optimal point is achieved where the marginal income on loans line intersects with the marginal cost of deposits line. As this will not be feasible in a situation in which no lender is willing to provide funds at a reasonable market rate, the bank needs to increase the volume of its deposits and/or reduce the volumes of its loans. The optimal point is achieved where the marginal income on loans line intersects with the marginal cost of deposits line. As this will not be feasible in a situation in which no lender is willing to provide funds at a reasonable market rate, the bank needs to increase the volume of its deposits and/or reduce the volumes of its loans.9

In a situation of rationing on the interbank market, which is closed to the bank, or in a situation of a binding loan-to-deposit ratio, a relevant fund transfer price can be found by grossing up the market rate. It creates internal incentives to raise additional deposits and reduce the volume of loans, ensuring the desired equality of loans and deposits. As, in reality, the location of the deposit and loan marginal lines are not known with certainty, some experimentation will be needed. The setting of a higher fund transfer price meets the intuition of better remunerating the collection of stable deposits, and charging a higher funding cost for illiquid bank loans. It preserves the logic of maximization of profit on loans evaluated against an adjusted fund transfer price, and the maximization of profit on deposits. An alternative to raising the FTP would be to directly set deposit and loan volume targets at the level where the volume of deposits is equal to the loans.

In this first advanced case, there was rationing on the interbank market, that is, banks cannot fund themselves in this market, often

---

8 If central banks provide liquidity in the short-term adjustment period, they will require rapid self-sufficiency in funding the assets of the bank.
9 An alternative financial strategy is to sell or securitize assets to reduce the funding deficit.
due to a fear of insolvency of the bank or sovereign risk resulting from a fear of insolvency of a country. In the second case, analyzed below, banks have access to an interbank market, but they must fund a Basel III contingency liquidity buffer.

FTP and Basel III contingency liquidity buffer (FTP4)
Consider the following balance sheet. At the initial stage, a loan portfolio (L) is funded partly by deposits (D) and interbank funding (I). The interest rate on interbank deposits is denoted by \( i \).

Loans L  
Deposits D  
Interbank deposits I (I)  
Liquid bonds B (b)  
Long-term debt F (f)

During the global liquidity crisis, that started in July 2007, the interbank market dried up and banks turned to governments or central banks to obtain liquidity. For example, in 2010, 18% of funding of Greek banks was obtained from the European Central Bank [Financial Stability Review (2010)]. As a result, bank regulators and the Basel Committee (2011) decided to put in place a new regulation to ensure self-sufficiency of liquidity in future crises, whereby banks must create a portfolio of contingency liquid bonds (B), funded with long-term funds (F, with B = F).10 The return on liquid bonds and the cost of long-term funding are denoted by \( b \) and \( f \). The objective is that the banks can sell the bonds (or use them as collateral) in the event that the interbank market dries up.

It is assumed that the cost of long-term funding (\( f \)) is higher than the return on liquid assets (\( b \)). The higher cost of funding is not related to the risk of bank default, a case analyzed explicitly below, but is caused by a liquidity premium. Corporate finance theorists [Holmström and Tirole (2011)] attribute the existence of a liquidity premium to imperfections in the capital markets and an absence of complete pledgeability of a firm’s future income. As a result, bondholders might be afraid that the bank’s insiders (management or shareholders) take actions that are detrimental to bond holders. An example is asset substitution, with an increase in the riskiness of bank’s assets after the terms of the bond have been fixed.11

Another related explanation of a liquidity spread is that investors are concerned that liquidity in the corporate bond markets, the ability to trade at reasonable cost in the future, might disappear.

The contingency portfolio of liquid bonds must be able to meet a deposit outflow of \( a \% \) of stable deposits \( D \) and \( \beta \% \) of more volatile interbank deposits \( I \) (with \( \beta > a \)). As long-term funding is expensive, the cost of the liquidity buffer, the difference between the cost of long-term funding and the return on contingent liquidity assets, is equal to: \( (f - b) \times B \).

Assume that \( B = F = (a \times D) + (\beta \times I) \).

As shown in the appendix, banks need two fund transfer prices, one for loans and one for deposits. They take into account not only the impact of collecting deposits and loans on the interbank position, but also the impact on the cost of the contingency liquidity buffer:

\[
\text{Loan} \cdot \text{FTP} = \text{marginal cost of loan} = i + [(\beta \times (f - b))] = \text{interbank rate} + \text{liquidity cost}
\]

\[
\text{Deposit} \cdot \text{FTP} = \text{marginal income on deposit} = i + [(\beta \cdot a) \times (f - b)] = \text{interbank rate} + \text{liquidity revenue}.
\]

The intuition is as follows. In the first step, an increase of one dollar in the loan portfolio requires additional interbank funding as

---

10 The report of the Financial Services Authority into the failure of Royal Bank of Scotland [FSA (2011, p. 106)] shows that the FSA liquidity rules did not cover the firm’s dependence on non-sterling denominated wholesale funding nor did they capture off-balance sheet liquidity risk, for example as a result of committed liquidity facilities. For sterling deposits, banks had to be able to meet a five-day stress scenario. A comment on a survey by the Financial Stability Institute [Grant (2011, p. 6)] states that: “Most banks included in the survey lacked a Liquidity Transfer Pricing policy.”

11 One way to reduce the liquidity premium on long-term funding is to issue covered bonds, bonds collateralized by safe assets.
there is a lack of deposits. The first part of the fund transfer price is the usual interbank rate. But, in a second step, the increase in interbank funding requires, due to the new liquidity regulation, an increase in the costly contingency buffer. This is the added liquidity cost \((\beta \times (f - b))\), the cost of carrying a liquidity cushion. Applying similar reasoning, an increase in stable deposits allows the bank to reduce interbank funding. The first part of the fund transfer price of deposits is, thus, the usual interbank rate. But, due to the contingency liquidity regulation, an increase in stable funding accompanied by a decrease in volatile interbank funding leads to a reduced contingency liquidity buffer \((\beta - \alpha)\) and additional revenue (the reduced cost of the liquidity buffer \(((\beta - \alpha) \times (f - b))\)) is added to the fund transfer price for deposits. It should be noted that a higher fund transfer price for loans will lead to higher lending rates. This is justified in order to correctly price the cost of liquidity risk.

Liquidity risk has drawn attention to the need to adjust fund transfer pricing to take into account rationing on the interbank market or the need to fund a Basel III contingency liquidity portfolio. The global crisis has also brought to public attention a third special case that creates the need for an advanced approach to FTP, whereby the credit riskiness of a specific asset of the bank differs from the average credit riskiness of the asset portfolio.12

**FTP and bank’s asset-risk-adjusted fund transfer price (FTP5)**

In many situations, a bank’s credit approval committee will consider a loan, the riskiness of which differs from the average riskiness of the bank’s assets. Consider the following example: the assets of a bank are funded with deposits, interbank debt and equity. The rating of the bank is A, with a cost of interbank debt of 5.6%.

<table>
<thead>
<tr>
<th>Assets</th>
<th>Deposits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interbank debt (A-rated, 5.6%)</td>
<td>Equity (10%)</td>
</tr>
</tbody>
</table>

Evaluate the following loan proposal: the funding of a very safe (AAA-equivalent) loan of 100 yielding an expected return of 5.5%. The new, very safe asset is funded with 2 parts of equity and 98 parts of interbank debt. The overall cost of equity is 10% and the corporate tax rate is 40%. In capital markets, the expected return on AAA-rated corporate bonds is 5%.

If we use the overall cost of debt and cost of equity of the bank, the foundation economic profit used to evaluate performance would be measured as follows:

\[
\text{Foundation economic profit} = \text{profit} - \text{cost of allocated equity} = (1 - 0.4) \times (5.5\% \times 100 - 5.6\% \times 98) \times (10\% \times 2) = -0.1928
\]

The economic profit being negative, the loan would not be accepted.

In the above example, the foundation economic profit would underestimate the value created by the safe loan because it would charge a too high average cost of interbank debt and equity. There are two flaws in this approach. The first, and more significant, is that it fails to recognize that taking a very safe (AAA-equivalent) asset (partly funded with equity) reduces the overall risk of the debt of the bank. The marginal cost of debt funding should be less than 5.6%. The second is that the overall cost of equity of 10% does not recognize the specific risk of the loan.

If it is not careful, the bank could find itself in a vicious circle, avoiding safe loans and funding risky loans that later lead to a further downgrade. One can argue that the funding of a safe asset would not change the market perception of the riskiness of the bank in the short term. This may well be true as the opacity of accounts could hide the change in the riskiness of the bank. However, opacity disappears in the longer term and the market will realize the improved soundness.13 Proper bank corporate governance rules call for long-term value creation (Dermine (2012)) and one therefore needs a FTP methodology that recognizes changes in the riskiness of the bank. The advanced economic profit approach makes it possible to respond to these flaws.

---

12 In our analysis of the impact of the contingency liquidity buffer, an investment in low-risk liquid assets will reduce the average riskiness of the bank's assets. If this effect might go unnoticed in the short term, it should, with more transparency, lead to a higher rating of the bank and a reduction in its cost of funds. This effect will reduce the net cost of the contingency liquidity buffer. A methodology discussed in the next case analyzes the case of a change in the riskiness of bank’s assets.

13 An official Swiss enquiry found that one of the reasons for the problems that UBS experienced during the crisis was the evaluation of very risky assets booked in New York with a fund transfer price that reflected the high rating of the overall bank (Swiss Federal Banking Commission (2008)).
An intuitive introduction to asset-based FTP is as follows. It is based on a bank valuation model developed by Dermine (2009). The traditional balance sheet of the new loan is rewritten as the sum of two equity-funded positions.

<table>
<thead>
<tr>
<th>Traditional balance sheet</th>
</tr>
</thead>
<tbody>
<tr>
<td>New safe loan 100 (5.5%)</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

In the traditional balance sheet, the loan is funded by debt and equity. One refers to levered equity as some debt is used in the funding of the asset. This traditional balance sheet can be expressed in a revised manner:

<table>
<thead>
<tr>
<th>Revised balance sheet (sum of unlevered 100%-equity financed positions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New safe loan 100 (5.5%)</td>
</tr>
<tr>
<td>Debt 98 (5% - ε)</td>
</tr>
<tr>
<td>(Short position)</td>
</tr>
</tbody>
</table>

Under the revised balance sheet, the leveraged loan is identical to the sum of a 100% equity-funded loan and a short position in debt (indeed, the debt of a company, the obligation to pay cash out, is equivalent to shareholders having a short position in a bond). The opportunity cost for the first position – equity funding the safe loan – is the expected return on similar-risk corporate bonds, 5%. Indeed, instead of making the safe loan, the bank can return the funds to shareholders who can buy similar-risk corporate bonds on the market. The expected cost of the marginal debt used to fund the loan is 5% - ε. Protected by the equity cushion of 2, the expected marginal cost of debt funding should be less than the expected return required on the new asset, 5%.

In the revised balance sheet (the sum of two 100% equity-financed positions), the advanced economic profit is equal to:

\[
\text{Advanced economic profit on loan: } [(1-0.4) \times 5.5\% \times 100 - 5\% \times 100] + [- (1-0.4) \times (5\% - \varepsilon) \times 98 - ((5\% - \varepsilon) \times 98)] \approx 0.26
\]

The first part represents the economic profit on the loan, funded 100% with equity (the after-tax return on the loan net of the cost of equity, the return available to shareholders on the corporate bond market). The second part represents the economic profit on the debt (that is, the after-tax cash outflow on the debt net of the opportunity cost of equity (the borrowing opportunity cost available to shareholders on the market)).

In the advanced approach, we ignore the overall cost of debt and the overall cost of equity, but rather focus on the marginal/incremental risk-adjusted specific opportunity rate. The advanced economic profit being positive, the loan should be accepted.

One could argue that financial markets are not so well informed and that they will not recognize the change in riskiness of the bank when it funds a safe asset. As stated above, this might well be the case in the short term, but boards of banks should be concerned with long-term value creation and accept that the market will soon grasp the change in riskiness of the bank. As this might take time, it is advisable to conduct two evaluations of the economic profit of a transaction: a short-term evaluation based on foundation economic profit (based on current average cost of funds) and a long-term advanced economic profit that recognizes the marginal risk of the new transaction.

**FTP and matched-maturity funding (FTP6)**

In the case of the Basel III Contingency Liquidity Buffer analyzed above, it is assumed that an extra dollar of loan would be funded on the short-term money market, a case of maturity transformation. Due to the new Basel III regulation on the contingency liquidity buffer, this prompted us to analyze the additional marginal effect on the liquidity buffer of granting an extra dollar of loan or collecting an extra unit of deposit. Funding of some assets on the short-term interbank market might not be feasible if a second Basel III liquidity regulation, the Net Stable Funding Ratio (NSFR), is binding [Basel Committee (2011)]. This regulation, which is to be applied in 2018, demands that on- and off-balance sheet positions that the bank must keep over one year be funded with funds with more than one-year maturity. In the event that this regulation is binding, a long-term loan can

---

14 In this example, it is assumed that large shareholders face the same cost of funds as the bank for the marginal funding of the new asset, 5% - ε. The calculation of the advanced economic profit is an approximation as the very small ε has been ignored.

15 An alternative approach to the sum of unlevered 100%-equity financed positions is to choose a funding structure (mix of debt and equity) that keeps the cost of bank debt constant [Dermine (2009)]. In that case, the marginal cost of debt is equal to the overall cost of debt of the bank.

16 Two financial techniques can be used by banks to ease the short-term effect resulting from opacity. The first is securitization of the safe asset. Securitization will force the market to evaluate the riskiness of the specific safe asset. A second technique is to use collateralized funding. The safe asset being used as a collateral should allow the bank to access lower-cost funding.
no longer be funded with short-term market funding (with the resulting impact on the contingency liquidity buffer analyzed above), and will have to be funded with long-term funding, such as a long-term bond. The cost is likely to be higher as it could incorporate a liquidity spread. In the event that long-term loans must be funded by long-term funding, the relevant fund transfer price will be the long-term market rate on bank’s debt. The search of an adequate fund transfer price must take into account the liquidity constraints imposed by the bank or the regulator.

So far we have ignored the risk of bank default. It has been assumed implicitly that the cost of interbank or bond funding would be met. Since the global banking crisis, however, the risk of bank default has increased significantly and the spreads on bank debt and credit default swaps (CDS) have become significant, more than 6% in several European countries. The risk of bank default is often the result of a risk of country default (sovereign risk) when banks hold a large volume of domestic country bonds or when the market anticipates that a country in default could nationalize its banking system to access domestic deposits. Figure 4 illustrates the very strong correlation between spreads on government and bank debt. Again, this is a new situation for banks from OECD countries, that has frequently been met in emerging markets.

Two methodologies to incorporate the bank’s credit spread in FTP are discussed next.

FTP and bank’s credit risk spread: conditional and unconditional approach (FTP7)

In this seventh and final specific case of FTP for deposits and loans, we analyze the situation in which the risk of bank default has created a credit spread on the bank’s funding. Should this spread be incorporated in the fund transfer price?

To ease the intuition, we consider the case of a one-year-to-maturity risky asset A, with contractual return a. The distribution is binomial: there is a probability (p) that the loan is repaid, and probability (1-p)

![Figure 4 - Changes in sovereign and banking sector CDS premia for a sample of European countries (22 November 2010 to 22 November 2011)](source: Financial Stability Report (2011)).

The cost of equity is the CAPM risk-free rate plus a risk premium. It is an unconditional cost of equity, giving the expected return on shares over both states of default and no default. The unconditional cost of equity has to be distinguished from the conditional cost of equity introduced below. The conditional cost of equity gives the return on shares in the case of no default. It is assumed that the asset default puts the bank into default. In that case, bond holders take control of the bank and recover the asset, REC.

The payoffs on the asset, bonds, and equity are given below for the two states of the world, default and no-default.

### Asset A (a) Corporate bonds C (contractual cost c) Equity E (unconditional cost of equity R_E)

The payoffs on the asset, bonds, and equity are given below for the two states of the world, default and no-default.

<table>
<thead>
<tr>
<th>State</th>
<th>Asset</th>
<th>Bond</th>
<th>Equity</th>
</tr>
</thead>
<tbody>
<tr>
<td>No default (probability p)</td>
<td>(aA + A)</td>
<td>(cC + C)</td>
<td>(R_E conditional on no-default + E)</td>
</tr>
<tr>
<td>Default (probability (1-p))</td>
<td>REC</td>
<td>REC</td>
<td>0</td>
</tr>
</tbody>
</table>

17 As explained above, corporate finance theorists attribute the existence of a liquidity spread to agency problems between outsiders (bondholders) and insiders (bank’s managers or shareholders). A second reason for a higher spread is a pure bank default risk spread. This last case is analyzed next.

18 As discussed in the case of the asset credit-risk approach, two methods can be used: a loan funded with a leverage structure (long-term funding and equity), or a portfolio of two unlevered positions: a loan funded 100% with equity (and as a cost of equity, the return expected on same-maturity risk corporate bonds) and a short position in debt.

19 The case of a continuous asset distribution for a risky bank is developed in Dermine (1986).
A project creates value when the unconditional expected return on the asset is higher than the unconditional expected cost of funds:

\[ p \times (aA + A) + (1-p) \times REC > [p (cC + C) + (1-p) REC] + [(1-p) \times (R_{E, conditional on no default} E + E) + p \times 0]. \]

This is the approach adopted in Dermine (1986) for a risky bank or by Cooper and Davydenko (2007) for risky corporate finance projects. It is an unconditional approach calculated for two states of the world, default and no default. Another, conditional approach, is possible. One observes that the asset return in case of default (recovery REC) is identical to the payoffs to the bond holders in case of default (they take over the bank and the assets). One can thus ignore the state of default and focus exclusively on the case of no default. In the conditional approach (the no-default case), the asset return has to be larger than the conditional cost of funds:

\[ (aA + A) > (cC + C) + (R_{E, conditional on no default} E + E). \]

In this last approach, the conditional case of no default, the cost of debt is the contractual rate paid by the bank (including the credit spread), and the cost of equity is the return demanded by shareholders in the case of no default, when they know that there is a risk of receiving nothing in the case of default. Where we introduce the risk of bank default, two approaches to FTP are feasible and fully consistent with one another: an unconditional approach over two states of the world (default or no default), and a conditional approach (the case of no default). We favor the second conditional approach, with the state of no default. Indeed, for the communication of a fund transfer price internally, it is easier to communicate the current contractual cost of debt which includes a credit risk premium. The communication of an expected cost of bank debt would be more difficult. But, one must note that in the conditional approach the relevant cost of equity is the conditional cost of equity, the return expected by shareholders in the case of no default. One must realize that the conditional cost of equity is higher than the unconditional CAPM cost of equity, as there is a need to remunerate shareholders for the risk of receiving zero return in the case of default.

Until recently, the consequence of a risk of bank default for FTP and the cost of equity had been ignored in the OECD banking world as the risk of bank default was low. Currently, much higher credit spreads in a number of countries have underlined the need to correctly incorporate the risk of bank default into the cost of funding, debt and equity. The bank must choose between an unconditional or conditional approach to FTP.

**Conclusion**

Fund transfer pricing is fundamental for evaluating the profitability of on-balance sheet products, for pricing and the design of bank strategy. It is fundamental for assessing the degree of competition in banking markets. This paper has drawn attention to five issues that have been previously ignored: rationing on the interbank market, the creation of a Basel III contingency liquidity buffer, the necessity to adjust FTP to the credit riskiness of specific assets of the bank, the need to include a liquidity premium in the case of long-term funding, and finally, the choice of a consistent methodology to incorporate the credit spread on the bank's own debt due to the perceived risk of bank default.

The seven cases which have been identified require a specific FTP methodology:

**Foundation approach**
- Products with fixed maturity (FTP1)
- Products with undefined maturity (FTP2)

**Advanced approach**
- Rationing on the interbank market (FTP3)
- Basel III contingency liquidity buffer (FTP4)
- Specific asset risk (FTP5)
- Long-term funding constraint (FTP6)
- Credit spread on bank's own debt (FTP7)

Following the global banking crisis, there is a pressing need to revise the methodology used to calculate FTP for deposits and loans. Sound corporate governance that focuses on long-term value creation calls for a distinction between short-term profit evaluation, likely to be observed in the case of opacity, and longer-term profit evaluation, when the market becomes transparent.
Appendix: FTP and Basel III contingency liquidity buffer

Here is the bank’s balance sheet, $R_E$ denoting the cost of equity:

| Loan: $L$ ($l$) | Deposits: $D$ ($d$) |
| Liquid bonds: $B$ ($b$) | Long term funding: $F$ ($f$) |
| Equity: $E$ ($R_E$) |

Balance sheet constraint is: $L + B = D + I + F + E$

Contingency liquidity constraint is: $B = (\alpha x D) + (\beta x I) = F^{20}$

The last constraint implies a revised balance sheet constraint $L = D + I + E$

One would expect $\alpha < \beta$ (interbank deposits, being more volatile, need a higher liquidity contingency buffer), and $f > b$ (liquidity premium raising the cost of long-term funding).

Economic profit = $[lL + bB - dD - iI - fF - R_E E]$

= $L × (l + bβ - i - fβ) + D × (bα - bβ - d + i - fα + fβ) + (bβ + fβ + i - R_E E)$

One observes that the separation theorem is restored. In this case, at the optimum, the marginal income on loans should equal the marginal funding cost given by the Loan Fund Transfer Price (FTP-L):

Marginal income on loan = FTP-L = $i + [f × (f - b)]$

At the optimum, the marginal cost of deposits must be equal to the marginal return given by the Deposit Fund Transfer Price (FTP-D):

Marginal cost of deposit = FTP-D = $i + [(\beta - \alpha) × (f - b)]$

The marginal return on deposits is equal to interbank funding rate plus the net cost reduction from a reduced contingency liquidity requirement (when $\alpha < \beta$).

Notice that the transfer price for loans and deposits is no longer identical because the impact of an extra unit of loans and deposits on the contingency liquidity requirement differs. One extra dollar of loan demands one dollar of interbank funding and a contingency liquid buffer. One extra dollar of deposits requires one dollar less of interbank deposits, with an impact on the contingency buffer which is a netted effect (smaller buffer for stable retail deposits, and larger buffer for volatile interbank deposits).

References


Dermine, J., 1986, “Deposit rates, credit rates and bank capital, the Klein-Monti Model revisited,” Journal of Banking and Finance, 10, 99-114


20 It is assumed the contingency liquid assets are funded with long-term debt because the cost of equity $R_E$ is more expensive.
The cultural revolution in risk management

Frank De Jonghe
Partner, Ernst & Young LLP, and University of Ghent

Mark Edelsten
Director, Ernst & Young LLP

Isabel Xavier
Senior Consultant, Ernst & Young LLP

Abstract
Following much debate among regulators and in society more broadly, it is now widely acknowledged that inadequate risk culture was a key contributor to the global financial crisis of 2007–08 and more recent corporate banking scandals. While there is now a growing consensus that something must be done to address behavioral risks, uncertainty remains around what exactly is meant by “risk culture” and how to “strengthen” risk culture. In this paper, we provide some clarity around the concept of risk culture, and propose a model and approach for assessing and strengthening it. The basis of our model is that risk culture is not static – it can be managed and shaped to provide a competitive advantage, allowing a company to achieve its objectives within the stated risk appetite.
The cultural revolution in risk management

Introduction
Risk culture is a leading indicator of how an organization can be expected to behave when under stress. During periods of economic uncertainty or organizational change, therefore, it is the organization’s risk culture that will drive decisions and behaviors, rather than risk policy. This was most recently demonstrated by the financial crisis, when market turmoil revealed the latent cultural issues that had been developing under the preceding period of economic growth. Inappropriate risk-taking behavior is now commonly recognized as one of the root causes of the recent financial crisis, and corporate scandals resulting from risky behaviors continue to be reported. However, despite all the controversy that has come to surround the concept of risk culture, the fact is that risk taking is integral to the success of a business. This is expressed by Thomas Stewart, a prominent management thought leader, thus: “Risk – let’s get this straight up-front – is good. The point of risk management is not to eliminate it; that would eliminate reward.”

The relationship between risk and reward is proportionate – the greater the risk a company takes, the greater the potential reward. The purpose of risk management is to allow organizations to optimize the risk-reward trade-off, to mitigate risks while enhancing potential rewards.

Characteristics of a strong risk culture include a committed leadership that models the right behaviors and provides clear and consistent communication around risks and effective governance structures, with clear roles and responsibilities for risk management. An open culture that encourages and responds to challenge will also assist in ensuring timely escalation of risk issues and learning from past mistakes. The outcome of a strong risk culture – a common understanding and respect for the role of risk management and appropriate risk behavior – provides a competitive advantage, allowing a company to execute its strategy within its defined risk appetite.

In the remainder of this paper we provide some further background on why risk culture has become a topical issue. We also provide a definition of risk culture and outline the defining features of a strong risk culture. Finally, we present a model that can be used to assess the current risk culture and guide broad intervention measures that aim to strengthen it.

Background
Regulators responded to the 2007–08 financial crisis through a number of interventions aimed at, for example, strengthening solvency, capital, and liquidity requirements and addressing operational risk issues. Bankers’ bonus structures were identified early on as one of the root causes of the crisis, and were an obvious target for regulatory intervention. Following the lead of the Financial Stability Board and the G20, who endorsed measures for sound compensation practices, a number of regions and countries implemented measures to “correct” bonus structures, with the aim of reducing incentives toward excessive risk taking (e.g., deferral of cash bonuses into equity, introduction of risk-based performance measures, and introduction of malus and clawback clauses).

This initial response shows how entrenched the concept of the self-interested “rational agent” remains in the thinking of society in general, and financial regulation in particular, since it offers a reasonable approximation for individual behavior within capital market institutions. In recognition of this notion that individual decision-making is motivated by the opportunity for personal rewards, institutions typically design incentive structures to align the interests of the institution with those of the individual, although at the same time encouraging and reinforcing individualistic, short-term behaviors.

Many criticisms of the rational agent model itself exist, however, and can be found in many areas of the social sciences, with the most recent criticism from behavioral economists and psychologists challenging the notion that individuals consistently act rationally. When making decisions in situations with an uncertain outcome (i.e., where there is an element of risk), one implicitly has to assess the odds of success of the different options. Research has consistently demonstrated that human beings are not good at this task, relying heavily on heuristics (i.e., mental shortcuts) in decision-making, which is subject to personal biases (for example, under the mental anchoring heuristic, an investor selling a stock will often take into account...

the original purchase price to assess future prospects since there will be a mental association between the price and the stock). As such, rather than acting as rational agents, individual decisions are influenced by personal biases, formed as a result of previous experiences, which may lead to an outcome that is not in the individual’s best interests.

By 2010, it was becoming clear that the problems large systemically important market institutions faced were more complex than previously envisaged. Regulatory interventions did not seem to be having an impact on behavior within the institutions they were targeting, and the focus on bonus structures was too narrow to address the problem at hand. As such, regulators started to broaden their focus to look at organizational culture. The Financial Services Authority (FSA) stated: “Regulators should recognize culture as a legitimate area of intervention... I particularly encourage boards to have a structured process for reviewing their firm’s culture, identifying its drivers, and the behaviors and outcomes it delivers... cultural change is essential if the industry is to minimize the probability and severity of the next crisis and regain the trust of society.”

The Prudential Regulation Authority (PRA) confirmed: “The PRA will expect firms’ governing bodies to embed and maintain a firm-wide culture that supports the safe and sound management of the firm. The PRA will not have any “right culture” in mind when making its assessments; rather it will focus on whether a firm is achieving the right regulatory outcomes.”

The Basel Committee on Banking Supervision supported the important role of culture: “An effective governance framework, set by the board, is critical to achieving a risk-focused culture within a financial institution. This, in turn, provides the foundation for the successful implementation of sound remuneration practices across the firm.”

The industry concurred, with the Institute of International Finance (IIF) stating: “... part of the management challenge of creating and sustaining a strong risk culture is to make explicit what is going on tactically, to correct the negative aspects, and to enhance and entrench the strong aspects already in place...”

So, the challenge for financial institutions is to understand what regulators now require them to do in response to their new demands. Can risk culture be clearly defined? And, if so, how can companies create and sustain a strong risk culture?

To understand why risk culture is a worthwhile approach, it is necessary to explore the development of risk management as a concept. Traditional approaches to risk management tacitly assume that individuals act rationally and are risk-averse to losses, harking back to Adam Smith and John Stuart Mill. However, if individuals do not act rationally or are prepared to take excessive risks, the risk management model starts to break down. Hence, traditional risk management frameworks, processes, and controls have come to be seen as necessary but not sufficient to allow an organization to achieve its strategic objectives within its defined risk appetite.

There has been recognition in the last decade that a broader, balanced perspective was required in order to adequately capture all types of risk. This can be seen in the growth of enterprise risk management (ERM) frameworks. ERM frameworks typically divide risks into categories, such as operational, financial, and strategic. To manage risk at a company level, it is common to see ERM frameworks further categorize financial risks into liquidity risk, credit risk, market risk, etc. These frameworks, however, fail to sufficiently take into account the real-world problem of behavioral risks.

The financial crisis has demonstrated that operational risk reviews are incomplete without assessing the risks created by people’s behavior. Processes, controls, and systems may be flawless, but if the people operating them take excessive risks, or behave unpredictably or irrationally, operational risks will emerge. The financial crisis has also taught us that behavioral issues can create systemic risks, and hence are strategically important and should be incorporated into ERM frameworks.

---

5 The PRA is to be established in early 2013 as a division of the Bank of England, spun out of the current FSA. The PRA’s remit will include the assessment of the quality of a firm’s risk management and governance, including culture.
7 Basel Committee on Banking Supervision, 2011, Range of methodologies for risk and performance alignment of remuneration.
8 Institute of International Finance, 2009, Reform in the financial services industry: strengthening practices for a more stable industry.
Importantly, understanding risk culture is not the same thing as measuring people risk, or the risks arising from specific organizational cultures. There are already multiple tools and approaches for understanding organizational culture. In this article, our purpose is to understand risk culture: that is, behaviors and attitudes to risk in large organizations, requiring specific tools and approaches. But first, we must clarify the concept of “culture.”

The concept of culture
The concept of organizational culture emerged in the early 1980s, loosely defined within the HR management profession as “the way we do things around here,” and more specifically defined by Schein, a leading expert in this field, as: “A pattern of shared basic assumptions that the group learned as it solved its problems of external adaptation and internal integration that has worked well enough to be considered valid and, therefore, to be taught to new members as the correct way to perceive, think, and feel in relation to those problems.”9

Schein identified various layers of organizational culture, with fundamental assumptions and values forming the core, and patterns of behavior and symbols forming the outer layers. These, often unwritten, rules were found to have a strong influence on employees’ behavior within the workplace.

A strong culture is defined as one in which there is a consistent and pervasive set of beliefs, values, assumptions, and practice across employees. A strong culture has been shown to support the achievement of strategic objectives by enabling an organization to act in an integrated and coordinated manner. In a simplistic sense, an organization’s strategy determines the “what,” while the culture determines the “how.” Importantly, despite the common perception that culture is a soft, implicit concept that cannot be measured or managed, research indicates that organizational culture can be assessed and shaped to provide a competitive advantage.10 To be successful, however, any change in culture requires a clear vision of the desired end state, strong leadership to model the desired changes, and a significant and long-term investment of time and resources to implement and embed the change across the entire organization.

Culture and risk are closely linked: “Culture is key to risk management... You can't rely on people looking at the rules. They are conditioned by culture and how the rules are enforced.”11 In order to understand the risks created by people's behavior, one must first understand how to characterize and analyze a given culture. Many risk culture models are generic and include factors such as people, processes, organization design, and rewards. Typically, however, these models are not organization-specific and not geared toward answering the question of how people behave with respect to risk taking and risk management. In addition, large organizations often maintain that they do not have a single culture, but rather, multiple cultures, relating to the different geographies or business lines of the organization. A new approach is, therefore, needed to address risk culture, a model that is organization-specific and focuses on behaviors relating to risk within a specific context.

Risk culture process model
There have been several attempts to adapt organizational culture models to describe how a company manages and responds

---

11 Ernst & Young LLP and Tapestry Networks, 2012, Progress on the risk governance journey, but key challenges remain.
to risk. The Institute of Risk Management (IRM) has recently developed a model for how risk culture develops, using the traditional organizational culture literature as a basis. The IRM proposes that “the culture in an organization arises from the repeated behavior of its members. These behaviors are shaped by the underlying values, beliefs and attitudes of individuals, which are partly inherent but are also themselves influenced by the prevailing culture in the organization.”12 Thus, risk culture is influenced and reinforced through employee attitudes and behaviors, in a continuous, iterative loop (Figure 1).

Different risk cultures may develop within an organization to support different business strategies. However, since risk culture describes attitudes to organization-specific risk, a slightly different approach may be required, balancing organization-specific and individual behavioral analysis.

Risk culture – what is it?
In brief, risk culture describes how an organization manages and responds to risk. The FSA provides the following definition of risk culture: “the general awareness, attitude and behavior of [the organization’s] employees and appointed representatives to risk and the management of risk within the organization.”13 The IIF provides this definition of risk culture: “the norms and traditions of behavior of individuals and of groups within an organization that determine the way in which they identify, understand, discuss and act on the risks the organization confronts and the risks it takes.”14 On the basis of its research in the financial services sector, the IIF found there is a close link between the risk culture of an organization and its risk appetite, defined as: “the amount and type of risk that a company is able and willing to accept in the pursuit of its business objectives.”15 The IIF emphasizes the importance of having a clear definition of risk appetite since “the statement of risk appetite balances the needs of all stakeholders by acting as both a governor of risk and a driver of current and future business activity.”

The risk appetite and the risk culture reinforce each other, with the defined appetite influencing behaviors and the culture influencing how well the appetite is embedded. Both work together to influence planning and decision-making. To some extent, one could say that the risk appetite is a formalized representation of attitudes toward risk.

What defines a strong risk culture?
A strong risk culture creates an environment in which risk is “everyone’s business” – there is a shared understanding and acceptance of the company’s risk appetite, and decisions are made in line with the risk appetite, even in the absence of a defined process or policy. A strong risk culture provides a competitive advantage by supporting the organization in capitalizing on opportunities without exposing the company to unacceptable levels of risk. Companies with a weak risk culture will either avoid risk due to a lack of clarity around what risks are acceptable or, more likely, expose themselves to excessive levels of risk due to insufficient risk awareness and low levels of compliance with risk policies and controls.

Recent research conducted by the IIF on how financial services companies have strengthened their risk management frameworks, showed that the need to build a consistent and unified risk culture is a critical area of focus for senior management teams. Fifty-eight percent reported increased attention to risk culture in the past 12 months, although this result was more pronounced among those companies that were most impacted by the financial crisis.

In addition, Ernst & Young LLP and Tapestry Networks conducted a study on risk governance in 2011, capturing the views of directors, chief risk officers (CROs) and supervisors. The results of the study identified four components that are necessary for a strong risk culture:16

- **Consistent tone at the top:** leadership (i.e., the board and senior management team) is responsible for setting the tone – communicating the firm’s values, strategy, and risk appetite, and modeling appropriate behaviors. A clear and consistent tone from the top is the foundation for a strong risk culture. As one director put it: “Risk is not a function, it’s an attitude.”
It must start with the board, the CEO, and then you get down into metrics."

- **Appropriate metrics that are regularly monitored**: one of the firm’s biggest challenges in embedding the risk appetite throughout the organization was identifying appropriate operational risk metrics, including measures of risk culture (i.e., instances of risk limits being exceeded without prior approval, percentage of self-reported control or risk problems).

- **Proper escalation processes and an open culture**: employees need to feel comfortable raising concerns and identifying issues to management. Once identified, managers must escalate issues in a timely manner. This openness is created when boards and managers are tolerant of mistakes and of honest attempts to do the right thing.

- **Consistent enforcement**: consistent enforcement was highlighted as critical for driving the right behaviors — for example, traders who exceed their limit should be sanctioned the same way, regardless of whether the behavior leads to a profit or a loss. However, there needs to be a balance between penalizing poor behaviors and encouraging openness.

As mentioned previously, the strong risk culture is not aimed at avoiding risk, but at ensuring that risks are being taken in a controlled way. As one CRO of a global bank told us: “I want transparency of the facts, an open and intellectually honest analysis, and then a commercially viable decision.”

### Measuring risk culture

The organizational culture literature demonstrates that culture can be measured and molded, and a number of models have been developed to achieve this (e.g., Hofstede’s (1980)\(^{17}\) “cultural dimensions theory,” and O’Reilly et al’s (1991)\(^ {18}\) “organizational profile model”). By extension, then, an organization’s risk culture can also be shaped to support the organizational strategy and risk appetite. Indeed, regulators and practitioners alike are beginning to acknowledge that risk culture can, and should, be measured. In response to growing demand, models and survey tools have been developed to assess employees’ attitudes and behaviors around identifying, taking, managing, and escalating risks. Any tool seeking to assess risk culture needs to acknowledge that the values and code of conduct communicated and endorsed by the company may not be reflected in the attitudes and behaviors of its employees. This is because there are multiple factors that directly or indirectly influence employees’ attitudes and behaviors, which may lead to a disconnect between what is explicitly communicated by leadership and what is understood (explicitly and implicitly) by employees. Models of risk culture aim to capture these factors, while the tools aim to identify where gaps exist between the desired risk culture and the actual risk culture.

### A potential model of risk culture

Risk culture cannot be adequately measured using existing tools for culture assessments. Most off-the-shelf culture models and tools are generic, typically designed to look at large samples across organizations or sectors, and focusing on individual behaviors, ignoring the organizational context. At the same time, risk models typically focus on risk strategy, appetite, or process, at the expense of individual perspectives and people factors.

An alternative approach is to balance the two perspectives and assess individual risk behaviors within an organizational context. Such a model works “top down” — the organization is the best place to focus initial efforts, as this is what leaders can control. Once the organizational factors have been addressed, the individual factors fall into place, even though this may be a drawn-out process in time.

At an organizational level, there are three primary factors that impact risk culture:

- **Leadership and communication**: how clearly leadership sets expectations around risk behavior; may include assessment of risk strategy and appetite, expectations, processes, and procedures.

- **Resources**: how supported people are to comply with risk policies; may include assessment of systems and tools, escalation mechanisms, training, etc.

- **Incentives**: whether people are incentivized to manage risk; may include assessment of bonus structures, accountability, and consequences.

The organizational factors will impact the following factors at an individual level:

- **Competencies**: what employees can do.

---


Motivation: what employees want to do.  
Application: what employees actually do.

It is important to take a holistic view of risk culture, as the factors are interrelated, and changes in one area may be ineffective without changes in another area; for example, changes in the firm’s risk policies must be supported by timely and effective training.

This is just one approach to identifying the factors that are likely to impact risk culture in an organization. Before embarking on any cultural assessment project, the organization’s leadership should review and select a model of risk culture (one which has been tested and validated with similar organizations) and refine the model for its organizational context. It is important to gain input from both HR and the risk and internal audit functions to help ensure all relevant factors are captured.

Tools and methodologies to assess risk culture
Once the organization has developed an adequate model of risk culture that captures the factors that may influence employees’ risk behavior, tools and methodologies can be designed to measure risk behaviors and attitudes to risk within the organization against these factors. In order to assess the gap between the desired risk culture and the actual risk culture, the assessment needs to capture and compare what the organization claims to be doing (and what leadership may think is happening) with what is actually happening “on the ground.”

An assessment of risk culture, therefore, should start with interviews or a focus group with leadership, to establish the desired risk culture. The aim of the interviews is to get leadership’s views on the organization’s strategy, risk appetite, governance framework, and behavioral expectations. Simultaneous with the interviews, there might be a review of the formal communications, including the organizational values, policies, and processes. Once the desired risk culture has been established, the organization needs to assess the actual risk culture. This can be achieved via interviews or workshops with selected employee representatives or via a survey administered to larger groups. Leadership and other project sponsors need to identify survey participants – the organization may choose to survey only “risk takers,” or both risk takers and “risk controllers.” Survey questions need to be developed (perhaps with the assistance of external consultants) to assess each of the factors in the model. In global organizations, it will be important to understand the latent cultural norms within each jurisdiction before assessing the risk culture.

The interview and survey methods outlined previously provide an insight into employee attitudes and perceptions around risk with the aim of assessing the tendency or likelihood of risk behavior within the organization. This type of assessment, which is naturally subjective and dependent on the integrity of individual reports, could be supplemented with analysis of more objective data sources available within financial institutions. Data such as records of compliance breaches, customer complaints, minutes of committee meetings, and preparatory files for major transactions could be used to analyze the incidence of actual risky or non-compliant behavior.

Once interview outcomes or survey responses have been analyzed, as well as any other data gathering, the organization can design and implement necessary culture change interventions. The organization needs to develop a clear action plan, based on priorities, impact and frequency, cost benefit, timelines, and dependencies, and outline specific activities to effect change, including responsibilities and measurements for success.

Strengthening risk culture
An effective way to influence behaviors in the workplace is through an organization’s HR processes (i.e., selection and recruitment, onboarding, performance assessment, and exit processes). Consequently, an assessment of risk culture must consider the risk framework and HR processes simultaneously. There needs to be alignment between risk and HR at a strategic and operational level. For example, the risk strategy and appetite should align with the organizational mission and values and acceptable risk behaviors should be reinforced through incentives.

At a more granular level, interventions aimed at addressing areas highlighted in the assessment phase may fall into one of the following categories of organizational factors:

Leadership and communication
The risk appetite should reflect the firm’s strategy and vice versa. Many organizations develop the firm’s vision, strategy, and
business plan in isolation from risk. Risk should be a key input into the business strategy and planning decisions. This may require organizational leadership to re-engineer their business planning processes, with input from risk and HR functions.

Consistent and frequent communications are critical, in order to embed the risk appetite within the organization. This includes explicit communications from leadership, setting out the firm’s vision, strategy, and risk appetite, as well as implicit communications that reinforce leadership’s message (e.g., staff training, recruitment process, performance management framework, and key performance indicators). Communications should highlight the benefits to the company (and therefore employees) of an effective risk appetite framework, explaining how the framework translates into behaviors at all levels and how these behaviors benefit the organization.

It is a commonplace view among risk managers that risk attitudes – or tone from the top – is a key ingredient of the risk management framework. However, it is worth noting that risk behaviors – defined by IRM as “observable risk-related actions, including risk-based decision-making, risk processes, risk communications” – can also influence attitudes via their impact on culture. Within a large retail bank, for example, branch bank managers have a key role in influencing frontline employee behaviors. Branch managers have a primary role in people management, sales and customer satisfaction, risk management, financial performance, and so on. Typically, branch bank managers, therefore, have a major impact on the culture of a retail bank. Hence, the tone from the middle can be as important as the tone from the top. The overall attitudes in a retail bank will be strongly influenced by this, especially where the senior executives have a long tenure, and gained much of their management experience in branch banking roles.

Resources
Risk governance and risk management frameworks should provide clear guidance on how risk issues should be escalated, with clarity and understanding around the roles and responsibilities for risk management throughout the organization. There needs to be frequent interaction between the first and second lines of defense, with adequate oversight and challenge from the second line of defense. The risk function needs to be given appropriate status within the organizational structure, and resources to fulfill its duties effectively.

There must be infrastructure for sharing risk information on a regular basis, which is not just limited to significant incidents. Data systems must allow the aggregation of risk data to provide the board with a view of the level of risk the organization is adopting overall.19

Risk policies and procedures should be reviewed regularly by the owner and should facilitate the completion of day-to-day tasks. All expectations around risk behavior should be supported by frequent, organization-wide training to ensure there is a common understanding and competence around risk management.

Incentives
Reward programs are a powerful driver of employee behavior. However, currently, the design of reward programs adjusts reward levels to take into account risks that have been identified (for example, by using risk adjusted measures), rather than seeking to prevent risk issues from occurring in the first place (i.e., by building risk management into role profiles). For example, our research on U.K. banks shows that risk measures are taken into account in determining bonus pools and individual bonus payments (ex-post adjustment). However, risk measures are, typically, not embedded into performance management processes (ex ante approach) and performance targets are generally based around profit, creating an imbalance between risk and reward.

We propose that, in order to influence behavior proactively and prevent breaches from occurring, risk needs to be considered prospectively and embedded into competency frameworks and role profiles (i.e., ex ante). Role profiles should include risk accountabilities and targets, which are then assessed as part of the performance management framework. A balanced scorecard approach is a useful tool for translating the organization’s vision and strategy into a clear and balanced set of financial and non-financial objectives. Traditional key performance indicators (KPIs) should be accompanied by key risk indicators ( KRIs) that allow an organization to plan, measure and monitor risk.

Recognition and reward outcomes should be directly linked to the performance management appraisal, and should be determined by both contributions and behaviors (including risk behaviors, such

---

The Journal of Financial Perspectives

as compliance with risk processes and openness around reporting risk issues). Outcomes must allow for both positive and negative reinforcements, thus reinforcing good risk behavior and penalizing poor risk behavior. People must be held accountable for their risk decisions and behavior. This depends on having robust risk measures (hard measures around compliance and soft measures around values) that are built into the performance management framework.

Finally, while there is a focus on cash bonuses in driving behavior, it is important to recognize that behavior can be influenced by multiple recognition and reward systems, and companies should use everything in their arsenal for maximum impact (e.g., career progression, training, and development opportunities). There is now wide acceptance that the disproportionate focus on cash incentives in financial services has driven pursuit of short-term gains at the expense of long-term sustainability. Most regulators, however, acknowledge there is a role for annual cash bonuses, but that these should be designed to reflect the organization’s risk appetite and profile.

Conclusion

This article has sought to address a wide range of new ideas within a short span. We have attempted to show that elements of people risk and risk behaviors may be measured and monitored to a greater degree than is currently imagined.

Analytical qualitative tools could be much more widely used by banks’ management and boards and by regulators, to understand what factors are driving unacceptable employee risk behavior. These tools offer potentially better ways to regulate banks. Currently limited to reward information, regulators may draw inappropriate conclusions on current culture. By widening the scope of their information sets, regulators would have much more useful information to develop future regulatory approaches.

As those institutions evolve, analytical tools of this nature could be used to monitor the improvements in risk culture within banks. Realistically, culture change of the type sought by regulators and politicians will take time and may involve a generational shift. This implies the need for longitudinal studies to track progress over time. Such studies, measuring employee attitudes and behaviors in the workplace, already exist in the academic world (e.g., the British Workforce Employee Relations studies), and could be used as a model for regulators.

In discussing these issues with practitioners and clients, we find there is great interest but, to a large extent, a lack of experience and expertise in this area, even in the largest global organizations. We hope that this article helps to clarify some of these ideas, and provides some guidance to practitioners.

Our key learnings

We have been able to distill the following key learnings from our research and experience in this area:

- **Leadership support**: risk culture is the responsibility of the leadership team. Boards and senior management must visibly and consistently demonstrate appropriate risk awareness, attitudes, and behaviors and unwavering commitment to compliance. Any cultural change must be driven by the board and requires visible and tangible support from leadership.
- **Phased approach**: an effective way of gaining acceptance of the risk culture model and assessment approach is to pilot the approach within a single function or business unit before rolling out to other areas of the business.
- **Change management**: a significant investment of time and resources will be required to effect and sustain change, with a strong change management strategy to help coordinate change activities.
- **Action plan**: creating simple, easy-to-understand outputs with clear business application will help to gain commitment to the results and action plan. The action plan should include “quick wins,” with tangible benefits, to secure commitment to the change process.
- **Information gathering and data**: information is key to allowing the business to identify current weaknesses in risk culture and measure the success of interventions.
- **Re-assess**: the risk culture measurement process should be repeated at regular intervals (at least annually) to assess the effectiveness of any cultural change initiatives.
About Ernst & Young
Ernst & Young is a global leader in assurance, tax, transaction and advisory services. Worldwide, our 167,000 people are united by our shared values and an unwavering commitment to quality. We make a difference by helping our people, our clients and our wider communities achieve their potential.

Ernst & Young refers to the global organization of member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. For more information about our organization, please visit www.ey.com.

© 2013 EYGM Limited.
All Rights Reserved.
EYG No. C00069

The articles, information and reports (the articles) contained within The Journal are generic and represent the views and opinions of their authors. The articles produced by authors external to Ernst & Young do not necessarily represent the views or opinions of EYGM Limited nor any other member of the global Ernst & Young organization. The articles produced by Ernst & Young contain general commentary and do not contain tailored specific advice and should not be regarded as comprehensive or sufficient for making decisions, nor should be used in place of professional advice. Accordingly, neither EYGM Limited nor any other member of the global Ernst & Young organization accepts responsibility for loss arising from any action taken or not taken by those receiving The Journal.