Stressing Bank Profitability for Interest Rate Risk*
Preliminary and Incomplete

Valentin Bolotnyy
Harvard University

Rochelle M. Edge
Federal Reserve Board

Luca Guerrieri
Federal Reserve Board

June 7, 2014

Abstract

We consider the ability of an array of time series models to forecast net interest margins (NIMs) both for an aggregate of large U.S. banks and for a number of these banks individually. The forecast models that we consider are those that in other applications have been found to be near the frontier of forecast performance. Of the models that we consider that are similar to those used in the macro-banking literature, the only ones to deliver forecasts that improve on a random walk—and here only marginally—are those derived from an average of the forecasts of many simple bi-variate models (that include NIMs and Treasury-yields of a number of different maturities). A few other models—such as, those specified in first differences—also improve on a random walk but here too the improvement is marginal. As a result even our best performing models have very large forecast errors. Consequently, even stressful interest rate risk scenarios, such as those used in the Federal Reserve's 2013 Dodd-Frank Act stress tests, imply NIM outcomes that are statistically indistinguishable from those obtained under the stress test's baseline scenario. This degree of uncertainty around the paths of key performance metrics for banks may limit the ability of stress-test results to maintain confidence in the banking sector, especially during periods of crisis.

Key Words: Net Interest Margins, Interest Rate Risk, Forecasting, Stress Tests.

JEL Classification: C53, E47, G21

* We thank Christina Wang for comments on an earlier draft of this paper and Jon Faust, Barbara Rossi, and Jeremy Rudd for helpful conversations. Bolotnyy acknowledges support from the NSF and the Paul & Daisy Soros Fellowship for New Americans. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System. Contact Information: vbolotnyy@fas.harvard.edu; rochelle.m.edge@frb.gov; luca.guerrieri@frb.gov.

1 Introduction

We examine how well an array of time-series models can forecast net interest margins (NIMs), the main source of revenue for "traditional" banks. The models and the techniques that we consider to forecast NIMs—defined as the ratio between net interest income (NII) and interest earning assets—have, in other applications, been found to be near the frontier of forecast performance. Our focus in this analysis is on conditional NIM forecasts, where the variables that we condition on are key macroeconomic variables. Our motivation for focusing on NIM forecasts that condition on macro variables is based mainly on two considerations. First, one of the key practical applications of our analysis is scenario-based stress testing—such as those required for the largest U.S. bank holding companies by the 2010 Dodd-Frank Act (DFA) and under the Federal Reserve Board's 2012 Capital Planning Rule—which conditions on a set of macroeconomic variables. Second, focusing on conditional forecasts allows us to separate the task of modeling and forecasting NIMs from that of forecasting interest rates, which is itself a sizable topic—see Duffee (2012).

Our results are not encouraging for NIM forecasts. Although all the models that we consider improve on a random walk *in-sample*, far fewer forecasts improve on a random walk *out-of-sample*—and here only marginally. Among the models that are similar to types of models that have been used in the macro-banking literature to model NIMs, the only ones that can improve on a random walk out-of-sample are those derived from an average of the forecasts of many simple bi-variate models (that include NIMs and Treasury-yields of a number of different maturities). A few other models—such as, those specified in first differences—also improve on a random walk, though again only marginally, and ultimately all of these models yield forecast errors that are very large and on the order of the variability of NIMs themselves. As a result, even dire scenarios, such as the stressful interest rate risk scenarios used by the Federal Reserve in the 2013 DFA stress tests, imply NIMs that are statistically indistinguishable from those obtained under the stress test's baseline scenario.

As noted one of our key motivations for studying NIMs' conditional forecast performance is the prominence of macroeconomic stress testing and capital planning as one of the key post-crisis reforms to U.S. capital regulation. Most significantly, stress testing and capital planning introduce forward-looking considerations to the assessment of bank capital adequacy by requiring that so-called *pro forma* capital ratios—that is, capital ratios projected to obtain under some *future* specified stressful scenarios—also satisfy regulatory minimum capital ratios.[1] Assessing bank capital adequacy based on stressed *pro forma* capital ratios entails (i) formulating stressful macroeconomic and financial scenarios upon which *pro forma* capital ratios are calculated and (ii) developing econometric models that can translate these stressful scenarios into revenues, losses, and ultimately bank *pro forma* capital ratios. It is the second of these tasks that concerns us in this paper and more specifically for revenues generated from banks' borrowing and lending activities

There are several reasons for our focus on revenue generation and on net interest income (NII), its largest component.[2] First, of all of the variables that need to be projected in conducting stress tests, revenue generation is where the state of modeling is at its earliest stages. Second, while it is typical to think of stressful scenarios as those in which credit-risk intensifies and large loan losses materialize, stressful scenarios in which revenue generation plummets or even turns negative as a result of the configuration of interest rates are equally important. Indeed, depressed or negative NIMs resulting from inverted yield curves have been a central feature of a number of banking crises and as such it is important that the revenue generation models used in stress tests can also capture these types of developments. Third, even

[1]The consequences of one or more of a bank's *pro forma* capital ratios falling below its regulatory minimum is that the bank would not be approved to pay dividends or make other capital distributions.

[2]In the decade prior to the zero interest-rate environment interest income accounted for about two-thirds of large banks' total income and about 40 percent of expenses (excluding provisions).

for scenarios that are primarily oriented to adverse credit risk outcomes, projections of bank revenues just as an important in calculating stressed *pro forma* capital ratios as are loan charge-offs and loan-loss provision projections. We elaborate on these motivations below.

As noted, of all of the variables that need to be projected in conducting stress tests, revenue generation is where the state of modeling is least advanced. Some of this reflects the fact that projecting credit losses through probability of default (PD), loss given default (LGD), and exposure at default (EAD) models has had a much longer tradition both in the industry and in the research literature given the central role these parameters play in assessing credit risk both for banks' own risk-management practices and under that Basel II internal ratings-based approach. Indeed it is PD, LGD, and, in some cases, EAD models, estimated on loan-level data, that are used to project charge-offs and loan-loss provisions conditional on the scenarios in the DFA stress tests, whereas it is more aggregate bank-level income and expense type models—similar to those used in this paper—that are used to project revenues.[3] To be sure, the recent importance of stress testing has led to the development of a number of models that link NIMs to macroeconomic variables and here, for the U.S., Covas, Rump, and Zakrajsek (2012) and Hirtle, Kovner, and Vickery (2013) are two notable examples. This research, however, has placed relatively less emphasis on evaluating the conditional forecast performance of these models, which we consider to be important given these models' ultimate use for generating *pro forma* capital ratios under stressed scenarios.[4]

While it is typical to think of stressful scenarios as those in which credit risk

[3]See Saidenberg and Schuermann (2003) for a discussion of PD, EAD, and LGD modeling and FRB (2014) for a description of the models used to project charge-offs, loan-loss provisions, and also revenue variables for the DFA stress tests.

[4]Covas, Rump, and Zakrajsek (2012) do present some forecast evaluation results in order to demonstrate the benefits of focusing on quantile projections. However, they only report results for net chargeoffs and pre-provisioning net revenue and their focus is on density forecasts generated by their quantile regression model.

intensifies and large loan losses materialize, stressful scenarios in which revenue generation plummets or even turns negative are equally important and should therefore be a development that models used for stress testing can capture. To be sure, losses that arise from this source of risk—called, interest rate risk—did not feature in the most recent crisis in which bank losses mounted as a result of loan defaults. But there are ample examples of other earlier banking crises in which adverse developments in NIMs have played a significant role. The most familiar of these from the U.S. perspective is the Savings and Loans (S&L) crisis that started in the early 1980s with short-term interest rates rising above long-term interest rates—in part, as a result of the Volcker disinflation—and resulted in interest expenses rising above interest income, NII and NIMs turning negative (in the thrift sector), and sizable losses and ultimately a substantial number of bank failures. Likewise, elevated short-term interest rates that depressed NII and NIMs contributed to the Nordic banking crises of the late 1980s and early 1990s (in Finland, Sweden, and Norway), as well as to the U.K.'s secondary banking crisis of the early 1970s.

Even for scenarios that are primarily oriented to adverse credit risk outcomes, projections of bank revenues are just as an important part of *pro forma* capital estimation as loan charge-offs and loan-loss provision projections. Indeed, in times of stress, the ability of a bank to remain viable depends just as much on its ability to generate revenues as it does on its losses on current assets. This is because bank revenues reflect a bank's ability to replenish its capital following losses.[5] One way to see this is to consider the magnitudes of the projected losses, revenue, and net income in the severely adverse scenario of the most recent DFA stress tests.[6] In this scenario, cumulatively through 2015:Q4, generated pre-provision net revenue (PPNR) is \$316 billion, resulting total losses are \$533 billion, and implied net income before taxes in -\$217 billion. Over this period, the aggregate tier one com-

[5]See Governor Tarullo's April 2012 speech "Developing Tools for Dynamic Capital Supervision" for an articulation of this view.

[6]See page 27 of FRB (2014) for the figures that follow.

mon ratio drops from 11.5 percent at the start of the stress test simulation period (2013:Q3) and to its lowest point of 7.6 percent, which since risk-weighted assets are $8,374 billion at the start of the simulation period and $8,656 billion at the end, means that tier one common equity drops from $963 billion to $658 billion. This $305 billion drop in tier one common equity, we interpret to equal net income losses combined with taxes, the phased-in effects of other comprehensive income losses, and the DFA stress tests' assumed path of dividend payments. Roughly speaking, a 0.22 p.p. higher or lower NIM—which is approximately the standard deviation of NIMs over the sample period we consider in this paper and is substantially less than many of our forecast models' root mean-squared (forecast) errors—implies a 6.6 percent higher or lower level of net interest income, a 10 percent higher or lower level of PPNR, and—assuming the same loan losses, taxes, and dividend payments—a 0.4 p.p. higher or lower *pro forma* capital ratio. This 0.4 p.p. capital ratio difference is roughly on the same order of magnitude as what CCAR 2012 BHC proposed capital distributions implied for bank capital, indicating that the magnitude of our forecast errors are large.[7]

Ultimately—and from a microprudential perspective—one of the key purposes of including forward-looking and stressed *pro forma* capital ratios as part of the regulatory capital regime is to demonstrate to bank creditors and counterparties that the banks with which they are transacting are able to withstand a severe macroeconomic outcome and, in this environment, continue to meet their obligations.[8] The

[7] The Fed has not published the implications of BHCs undertaking proposed capital distributions versus no capital distributions since CCAR 2012, so for this reason we can only compare our forecast-error magnitudes with these results. In brief, in CCAR 2012, the aggregate tier one common ratio dropped from 10.1 percent at the start of the stress test simulation period (2011:Q3) and to its lowest point of 6.2 percent, when firms undertook their proposed capital distributions, and to 6.8 percent, when firms undertook no capital distributions. See page 23 of FRB (2012) for these figures.

[8] From a macroprudential perspective, the purpose is also to demonstrate that the banking sector is able to withstand a severe macroeconomic outcome and, in this environment, is continue to serve its key credit-intermediation function in the macroeconomy, thereby not creating any additional source of drag to the weak economy.

results of such programs can, however, only be effective in maintaining creditor and counterparty confidence in banks if observers are convinced of the ability of supervisors to assess the consequences to bank capital ratios of their scenarios. It is therefore important to consider the ability of forecast methods at the frontier of efficiency to disentangle the implications of baseline and stress scenarios, since their not being able to do so dramatically detracts from the ability of stress-test programs to maintain confidence in the banking system.[9]

To consider this the paper proceeds as follows. Section 2 provides a review of the literature on NIMs with an eye to cataloging variables that have been found useful to explain the evolution of NIMs. Section 3 turns to the models that we develop to forecast NIMs and the approaches that we will use to evaluate and understand relative forecast performance. Section 4 describes the data used to estimate the forecast models and Section 5 reports our forecast results. We document that of the models that are similar to those used in the macro-banking literature, the only forecasts that improve on the random walk are those that average the forecasts of many simple (NIM and Treasury-yield) bi-variate models and here they do so only marginally. Models specified in first differences also improve on a random walk but again only marginally. As such, our out-of-sample our forecasts errors are large, especially relative to the variability of NIMs. Section 6 uses two of the paper's models to consider the path of NIMs under the scenarios published by the Federal Reserve as part of the 2013 Dodd Frank Act stress tests. We find that because our forecast errors are so large, the NIM outcomes that we obtain under the stress scenarios are are statistically indistinguishable from those obtained under the baseline and argue that this degree of uncertainty around the paths of key BHC variables may compromise one of the key goals of stress testing and its

[9]Note that this type of stress test program is not limited to the U.S., where bank stress testing has been undertaken annually since 2011 as well as during the financial crisis in early 2009: Stress testing of banks based on scenarios was undertaken during the Nordic banking crises to shore up confidence in the ability of banks to withstand the crisis and more recently and more extensively by the European Banking Authority for the European Union in 2009, 2010, and 2011.

associated capital planning, which is to maintain the confidence of bank credits and counterparties including in periods of severe financial sector stress. Section 7 concludes.

2 Related literature and variable selection

There are two strands in the literature on modeling NIMs. The first strand emphasizes the link between interest rates of various maturities and NIMs and lies more in the macro-banking tradition, while the second strand focuses on the optimal margin set by banks and lies more in the micro-banking tradition. Our paper lies in the first strand of the literature and so we focus our discussion here.

Covas, Rump, and Zakrajsek (2012) and Hirtle, Kovner, and Vickery (2013) are two recent papers in the macro-banking tradition that both link the level of NIMs to the level of the slope of the yield curve and the level of a short-term interest rate. This choice of variables reflects the two key services supplied by banks that their earnings from interest income reflect; specifically, maturity transformation services and deposit transactions services. The slope of the yield curve reflects the return to banks from maturity transformation and is thus is expected to enter NIM models with a positive sign. The short-term market interest rate in NIM models reflect the fact that bank deposit rates, while typically lower than market rates (since they provide transactions services), are constrained by the zero lower bound. As such, the short rate places an upper limit on what banks can earn from the provision of their transactions services. As a result, the short rate is expected to enter the model, also, with a positive sign.

Both Covas, Rump, and Zakrajsek (2012) and Hirtle, Kovner, and Vickery (2013) measure the slope of the yield curve in their respective NIM models with the 10-year to 3-month Treasury term spread and short-term market rates with the 3-month Treasury bill rate, as too do English (2002), English, den Heuvel, and Zakrajsek (2012), and Alessandri and Nelson (2012)—other papers in this tradition. All of these papers estimate NIM models with slightly different purposes in mind but all use essentially the same interest-rate variables to summarize the slope and the level of the yield curve.

The NIM models of Covas, Rump, and Zakrajsek (2012) and Hirtle, Kovner, and Vickery (2013) are developed for stress testing and are part of much larger aggregative models of a number of components of banks' financial statements. These models are both estimated on panel data with the most notable difference between these models being the estimation techniques they employ: Least squares regression in the case of Hirtle, Kovner, and Vickery (2013) and quantile regression in the case of Covas, Rump, and Zakrajsek (2012).

Like our paper, English (2002) focuses exclusively on NIMs, although he does so for a number of countries (that is, all G-7 countries except for France, and for Australia, Norway, Sweden, and Switzerland), and, additionally, he considers asset yields and liability yields separately—the two measures that when differenced imply NIMs. The main focus of English (2002) is understanding how well the aggregate banking sectors in the countries that he studies appear to have managed the interest rate risk that they face on their earnings, as such he uses aggregate data (rather than firm-level data) and does not focus on the conditional forecasting performance of his model.

English, den Heuvel, and Zakrajsek (2012) consider the relationship between NIMs and the slope of the yield curve and short-term rates for a panel of banks, however, in contrast to the other studies they allow for the specific coefficients estimated for each bank to vary as linear functions of bank balance sheet ratios that one *a priori* might think would be relevant to net interest income. In particular, these ratios include the maturity gap between bank assets and liabilities, the deposit share of liabilities, and the loan share of assets. Again, since English, den Heuvel, and Zakrajsek (2012) are focused on understanding the determinants of NIMs they do not consider conditional forecast performance. Alessandri and Nelson (2012), who use data for the U.K., also considers the relationship between NIMs and the slope of the yield curve and short-term rates for a panel of banks, with an ultimate interest in investigating the possibility of hedging between banks' income generation from their lending and deposit-taking activities and their trading book activities, which they document to be respond to interest rates in the opposite direction to NIMs. Here too, however, the authors do not consider forecast performance.

The macro-banking papers that we have just discussed rely on two summary measures of the yield curve to model NIMs—in particular, the level and slope of the yield curve—and all focus on modeling the level of NIMs. The literature modeling the yield curve also considers curvature as a key summary measure of the yield curve, and so to the extent that yields at more than just the 3-month and 10-year points on the yield curve are relevant for banks' net interest income, we also consider yield-curve curvature.

A separate strand of the NIM modeling literature comes from the micro-banking tradition, which focuses on the determinants of the loan rates and deposit rates that banks set as implied by the maximization of their profits. Papers in this tradition—such as, Ho and Saunders (1981), Angbazo (1994), and Saunders and Schumacher (2000)—emphasize a very different set of variables for modeling NIMs; in particular, the degree of competition facing banks in deposit and loan markets, banks' degree of risk aversion to finding themselves with deposit supply without corresponding loan demand or loan demand without corresponding deposit supply, and the volatility of interest rates. We intend (in future drafts) to add some of these variables from the micro-banking literature to our forecasting models for robustness analysis, especially since the current draft leaves us with the unanswered question of what—at the macro level—does forecast NIMs, given that interest rates do not seem to do so.

3 Data

We perform our forecast evaluation analysis of aggregate NIMs over two sample periods 2000q1 to 2008q2 and 2000q1 to 2013q4. We use the sample that ends in 2008q2 as our baseline (and preferred) sample period due the complications posed by the last half decade of data. The most notable of these is, of course, the zero interest rate environment, which poses significant challenges for many models of interest rates. Additionally, the full repeal of Reg. Q in the DFA, which now allows interest to be paid on business checking accounts, could also be a complication, although perhaps less so to date given the low interest rate environment. We do, however, extend our analysis to 2013q4 for robustness analysis and to compare (in future drafts) our model forecasts with those of equity analysis, which only exist from 2008 onwards.

3.1 Net interest margins

3.1.1 Aggregate net interest margins

For our forecast analysis of aggregate NIMs we use data from the quarterly Consolidated Reports of Condition and Income (Call Report) that every national, state member, and insured nonmember bank is required to file on the last day of each quarter by the Federal Financial Institutions Examination Council (FFIEC). The Call Report data used in this analysis are adjusted for bank mergers and acquisitions, using structure data from the National Information Clearinghouse (NIC) on mergers and acquisitions. Foreign entities are excluded and domestic subsidiaries are aggregated up to the parent, bank-holding-company (BHC), level.[10] To get an aggregate banking sector measure of NIMs, we aggregate NIMs for the top 25 BHCs, as ranked by total assets, where this is assessed quarterly.

[10]NIM data are adjusted for mergers between commercial banks by comparing balance sheet values of interest income, interest expenses, and interest-earning assets at the end of the quarter with those at the beginning of the quarter, accounting for amounts acquired or lost during the period because of mergers – see the appendix in English and Nelson (1998).

The top panel of figure 1 shows the historical time series of NIMs while the middle panel shows the two series—Interest Income divided by Interest Earning Assets and Interest Expenses divided by Interest Earning Assets—that when differenced imply NIMs. The historical time series of NIMs shows a very large and peculiar spike in 1988q4. This large spike stems from late payments from Brazil during the Latin American debt crisis and it is for this reason that our baseline sample starts in 1989q1 (rather than at the start of the series).

In the latter part of the sample period that we consider in this paper there were two banking-sector structural changes that have implications for measured NIMs. The first change was the Fed starting to pay, in the fourth quarter of 2008, interest on excess reserves. This change meant that excess reserves, which were previously not part of interest earning assets, became part of interest earning assets, so increasing the denominator of NIMs. As the same time the value interest income also increased, although, only by a small amount given the very low interest rate paid on excess reserves. As such, the overall effect on NIMs of the Fed paying interest on excess reserves, is to push it down. The amount by which this structural change pushes down NIMs can be seen by comparing the black solid and black dashed lines in the top panel of figure 1. We also show in the middle panel how the interest income and expense components of NIMs change as a result of making this adjustment. The second change was Financial Accounting Statements (FAS) 166/167 coming into effect at the beginning of 2010. These new standards resulted in banks needing to bring nearly $362 billion in assets and liabilities back onto their balance sheets of which nearly 90 percent were consumer loans.[11] Of these consumer loans a sizable fraction were credit cards, which have much higher interest margins than other types of loans. Thus as a result of FAS 166/167 NIMs jumped notably in the first quarter of 2010, as can be seen by comparing the black solid and black dotted lines in the

[11] See "A jump in consumer loans," *Economic Synopses* 2010-18, by Hoda El-Ghazaly and Yadav Gopalan, St Louis Fed.

top panel of figure 1. We also show in the middle panel how the interest income and expense components of NIMs change as a result of making this adjustment. The NIM series that we use when we extend our analysis our to the end of 2013 combines the two adjustments shown in the top panel of figure **??**.

3.1.2 Firm-level net interest margins

For the firm-level analysis (which we plan to undertaken in future drafts) the data that we use is from the quarterly Consolidated Financial Statements for bank holding companies, called the Y-9-C. For each BHC in our sample—which is roughly the 30 BHCs that will be in the 2014 DFA stress tests—mergers are accounted for using a so-called *pro forma* approach. That is, a historical time series for each BHC is constructed assuming that all of the institutions that are now part of the BHC always were part of the BHC. Our current data set starts in 1996q1, which is when the *pro forma* Y-9-C data that we use for this analysis starts. Y-9-C data, however, does extend back to 1991, so our sample period for the firm-level analysis should be able to be extended back to then as well.

3.2 Treasury yields

The yields data are derived using a smoothing technique from Gurkaynak, et al. (2007), based on Nelson and Siegel (1987) and Svensson (1994), which allows for daily measures of an off-the-run Treasury yield curve. We use quarterly yields for twelve maturities in our models: 3-month, 6-month, 9-month, 1-year, 2-year, 3-year, 5-year, 7-year, 10-year, 15-year, 20-year, and 30-year.[12] The yields, which are plotted in the lower panel of figure 1, are quarterly averages of daily yields.

[12]Daily yields are published with a two-day lag under the base mnemonic SVENY at www.federalreserve.gov/econresdata/researchdata/feds200628_1.html.

3.3 Variables suggested by the micro banking literature

The two variables from the micro-banking literature that we (plan to) add to our forecasting models are the degree of competition facing banks in deposit and loan markets and the volatility of interest rates. The degree of competition that banks face is captured by the relative size of the shadow banking industry. This is measured by the volume of assets in the shadow banking industry relative to the size of the total banking industry, that is, both traditional and shadow banking (as measured by the combined assets of these these two industries). We prefer to use the size of the traditional banking industry as our measure competition—rather than measures of competition within the banking sector itself, like Herfindahl Hirschman indices—for the reason that competition from the shadow banking sector has been the main source of increased competition faced by the traditional banking sector that has over time compressed interest margins. Importantly, competition from the shadow banking sector adversely affects both sides of traditional banks' balance sheets; that is, it puts upward pressure on the deposit rate that traditional banks must pay to attract deposits—since the shadow banking sector offers alternatives to deposits like money market mutual funds—and it puts downward pressure on the lending rate that traditional banks can earn—since the shadow banking sector offers alternatives to bank loans like loans from finance companies. The relative size of the shadow banking industry can be easily calculated from the U.S. Financial Accounts (previously known as the Flow of Funds Accounts), where the traditional banking sector consists of commercial banks, savings institutions, and credit unions and the shadow banking sector consists of broker-dealers, ABS issuers, finance companies, mortgage pools, and funding corporations. This series is plotted in the top panel of figure 2.

We capture the volatility of interest rates by the Merrill Lynch Option Volatility Estimate (MOVE) index, which is a market-based estimate of future Treasury bond yield volatility. The MOVE Index reports the average implied volatility across a wide range of outstanding options (with expiry dates of approximately one month) on two-year, five-year, 10-year, and 30-year U.S. Treasury securities. This series is plotted in the middle panel of figure 2.

3.4 Bank equity-analysis forecasts and Blue Chip Treasury rate forecasts

The last two pieces of data that we mention are bank equity-analyst forecasts of individual BHC NIMs, which we will use for comparison against our model-based BHC-specific NIM forecasts, and Blue Chip (Financial) Treasury rate forecasts, which we will use (in future drafts) to condition our model-based BHC-specific NIM forecasts. The financial-information firm SNL Financial LC has since 2007q4 collected bank equity analysts' forecasts for a number of key variables in BHC financial statements. For any quarter SNL begins collecting forecasts about two years in advance, however, at this horizon only a very small number of analysts report forecasts. Forecasts are more widely reported by analysts a year in advance of the quarter, which is what we use for our analysis.

The BHCs for which (in future drafts) we consider analyst forecasts are as close as possible to the 30 BHCs that were included in the 2014 DFA stress tests. Ultimately, this ends up being 19 BHCs (though, *not* the same BHCs that were part of the original SCAP and original CCAR stress tests). The BHCs that we cannot make comparison with are either those for which we do not have historical BHC data on which to estimate our bank-specific NIM models (such as, Ally, American Express, Discover, Goldman Sachs, HSBC North America, Morgan Stanley, and Santander/Sovereign) and those for which we do not have SNL forecasts (such as, BBVA, BMO, RBS, and UnionBanCal). In all cases we use only the average across all analysts for the BHC for the quarter since this was all that was available to us. For all of our BHCs we have about 20 different equity analysts forecasts per BHC, although for several BHCs we have notably more, particularly for the larger BHCs, where we typically have close to 30.

Importantly, equity analysts do not have realized interest rates available to them in forming their forecasts, which is something that in all of our other model-based NIM forecast evaluation analysis we take as being available. So as not to give our model-based NIM forecasts an unfair advantage in this forecast comparison exercise we generate our model-based NIM forecasts conditional on Blue Chip (Financial) Treasury-rate forecasts and, more specifically, the rate on the 3-month, 6-month, and 1-year Treasury bills, the rate on 2-year, 5-year, and 10-year Treasury notes, and 30-year Treasury bonds. Blue Chip releases these forecasts on the first day of each month. We therefore generate all of our NIM-model forecasts—conditional on Blue Chip Treasury rate forecasts—as if we were doing so on the first day of the quarter; that is, January 1, April 1, July 1, and October 1. Given this timing of our NIM-model forecasts, the SNL forecasts that we use for comparison are those that were recorded on January 4, April 4, July 4, and October 4 (or the earliest day thereafter). Choosing these dates for the SNL forecasts means that the equity analysis have as much information as we have about interest rates in generating their forecasts.

4 Models for forecasting NIMs and methods for comparing performance

4.1 Models for forecasting NIMs

We construct forecasts from a wide array of standard methods taken from the forecasting literature. We use data with quarterly frequency and consider a forecast horizon up to ten quarters long. For all but our benchmark model, we forecast NIMs conditional on Treasury yields or conditional on some constructed variables—that is, yield-curve factors—that summarizes yields. We denote the variables upon which we condition our forecast by I and denote our s-step ahead NIM forecast conditional on I by $NIM_{t+s/I}$. When we compare (as we will in future drafts) our BHC-specific model-based forecasts with SNL forecasts, we do—as just described—condition on Blue Chip (Financial) Treasury-rate forecasts. When we perform scenario simulations—as we do in section 6—the variables that we use to condition our forecasts are the paths of Treasury yields or implied factors given in the scenarios. Under our preferred specification Treasury yields or yield-curve factors enter our forecasting models with one or more lags. We decided not to use contemporaneous Treasury yields or yield-curve factors to condition our forecasts because doing so raises questions the about exogeneity of our regressors and using these terms instead does not improve forecast performance (likely reflecting the fact that interest rates on securities do not reprice, while those, even on floating-rate loans and on all forms of non-market funding reprice with some lag).

All of our forecast models, with the exception of the benchmark model, are autoregressive in NIMs, which means that forecasts two or more periods ahead can be generated either iteratively and directly. We consider both of these ways.

We also consider versions of our models specified with both NIMs and yields (or yield-curve factors) in levels and with both NIMs and yields (or yield-curve factors) in first differences. In the former case our models always contain at least one lag of NIMs and one additional lag of yields (or yield-curve factors). Specifying the model in this way is important given the possibility of unit roots in both NIMs and yields and so if we are to specify our models in levels we must—as one of the standard approaches for addressing the possibility of spurious regressions—include lags of both our dependent and explanatory variables. Of course, another standard approach for dealing with the possibility of spurious regressions is to estimate the model in first differences and this we do as well. Our first model, however, which serves as our benchmark model, is the same throughout.

1. No-change forecast. This model, which forecasts NIMs as a random walk without a drift at all horizons s, is given simply by: $NIM_{t+s/t} = NIM_t$.

4.1.1 Iterative levels models

Our first three iterative models—specifically, observed factors with forecast combination, dynamic factor model with forecast combination, and principal component analysis (PCA) factors with forecast combination—represent NIMs as a function of one lag of itself and two lags of the three factors that are typically viewed to summarize the yield curve; that is, the level, L, slope, S, and curvature, C, of the yield curve.[13] All that differs between these three factor-based forecast-combination models is how we obtain these factors, as we will describe below. What is identical between these three approaches is our general approach, which is (i) to regress NIMs in three separate equations on its own first lag and on the lags of each one of the three factors $F \in \{L, S, C\}$:

$$NIM_t = c_f + \rho_f NIM_{t-1} + \sum_{j=1}^{2} \gamma_{f,j} F_{t-j} + \eta_{f,t}, \tag{1}$$

(ii) to produce three different forecasts of NIMs extending out s quarters and each conditional on one of the three factors $F \in \{L, S, C\}$ through period $t+s-1$, denoted by $NIM_{f,t+s/f}$, and (iii) to combine these three forecasts with equal weights—that is, average them—so as to obtain:

$$NIM_{t+s/\{L,S,C\}} = \sum_{f \in \{L,S,C\}} \frac{NIM_{f,t+s/f}}{3}.$$

Clearly, a variant to the above approach would be to include all three factors in a single NIM equation, which we have tried but invariably found to imply poorer forecast performance.

2. Observed factors with forecast combination. In this model the three factors used in equation 1 are calculated as simple (additive) functions of a small

[13]Experimentation with longer lag structures led to no improvement in the forecast results reported below.

number of interest rates of different maturities. In particular, the level factor L_t is set equal to the average of the 3-month, 2-year and 10-year Treasury yields, the slope factor S_t is the difference between the 10-year and the 3-month yields, and the curvature factor C_t is given by 2 times the 2-year yield minus the the sum of the 3-month and 10-year yields. These factors follow the "observed factor" definitions used by Diebold and Li (2006).

3. Dynamic factor model with forecast combination. In this model the three factors used in equation 1 are obtained following Diebold and Li (2006). The dynamic factor model that we consider takes the form:

$$\begin{pmatrix} L_{t+1} - \mu_L \\ S_{t+1} - \mu_S \\ C_{t+1} - \mu_C \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} L_t - \mu_L \\ S_t - \mu_S \\ C_t - \mu_C \end{pmatrix} + \begin{pmatrix} \eta_{Lt} \\ \eta_{St} \\ \eta_{Ct} \end{pmatrix}$$

$$r(\tau) = \begin{pmatrix} 1 & \frac{1-e^{-\lambda\tau}}{\lambda\tau} & \frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau} \end{pmatrix} \begin{pmatrix} L_t - \mu_L \\ S_t - \mu_S \\ C_t - \mu_C \end{pmatrix} + e(\tau)_t,$$

We use the smoothed estimates of the factors L_t, S_t, and C_t from this model, which we estimate and condition on all information up until period T.

4. Principal component analysis (PCA) factors with forecast combination. In this model the three factors used in equation 1 are obtained from principal component analysis of our twelve different yields.

5. Single partial least squares (PLS). This model is similar to the PCA factor model in that it represents NIMs as a function of one lag of itself and lags of three factors. What is different is that while PCA selects factors that explain the largest fraction of the variance of Treasury yields, PLS selects factors as those that explain the largest fraction of the covariance between NIMs and Treasury yields. We calculate our PLS factors following the approach laid out by Groen and Kapetanios

(2009) for when the equation being estimated includes lagged dependent variables. In particular, we need to control for the effect of lagged values of the dependent variable on the covariances between NIMs and lagged Treasury yields before we calculate the PLS factors. We first regress demeaned NIMs and each lagged Treasury yield individually on lagged demeaned NIMs. We then take the residuals from each of these regressions and using the algorithm described by Groen and Kapetanios (2009) calculate our PLS factors. Unlike the other factor models we do not use forecast combination methods. Lagged demeaned NIMs and the PLS factors are all, by construction, orthogonal to each other, so estimating the model with all variables in the same equation will not give different answers to taking the forecast combination approach.

6. Treasury yields with forecast combination. This model, which like the first three iterative models follows a forecast combination approach, (i) regresses NIMs in twelve separate equations on its own first lag and lags of each Treasury yield $r_{\tau,t-1}$ where $\tau \in \{0.25, 0.5, \cdots, 30\}$:

$$NIM_t = c_\tau + \rho_\tau NIM_{t-1} + \sum_{j=1}^{2} \gamma_{\tau,j} r_{\tau,t-j} + \eta_{\tau,t}, \tag{2}$$

(ii) produces twelve different forecasts of NIMs extending out s quarters and each conditional on one of the twelve Treasury yields through period $t+s-1$, denoted by $NIM_{\tau,t+s/\tau}$, and (iii) combines these twelve forecasts with equal weights—that is, average them—so as to obtain:

$$NIM_{t+s/\{0.25,0.5,\cdots,30\}} = \sum_{\tau} \frac{NIM_{\tau,t+s/\tau}}{12}.$$

7. 3-month and 10-year Treasury yields with forecast combination. This model is identical to the one described immediately above with the exception that it only uses two yields—the 3-month yield and the 10-year yield—so that $\tau \in \{0.25, 10\}$. As such, this approach averages only over two models.

8. VAR with 3-month and 10-year Treasury yields. The VAR includes NIMs and the 3-month and 10-year Treasury yields. Forecasts conditional on the factors are obtained using the Kalman filter, as in Clarida and Coyle (1984). We take as conditional forecasts the filtered (that is, one-sided) estimates of NIMs conditional on the pseudo-out-of-sample observations of the remaining variables in the VAR. This procedure efficiently considers the covariance between reduced-form residuals in the VAR.

4.1.2 Direct levels models

As noted earlier generating forecasts directly means specifying and estimating separate equations for each step-ahead forecast. Note that whereas the lag of the dependent variable changes in the equation specified for each step-ahead forecast, the timing of the yields or factors upon which we condition the forecast do not change.

2. Observed factors with forecast combination, 3. Dynamic factor model with forecast combination, and **4. Principal component analysis (PCA) factors with forecast combination.** For our three factor-based forecast-combination models generating direct forecasts means specifying the following ten different versions of equation 1 for each factor $F \in \{L, S, C\}$:

$$NIM_t = c_{f,1} + \rho_{f,1} NIM_{t-1} + \sum_{j=1}^{2} \gamma_{1,f,j} F_{t-j} + \eta_{f,t}, \text{for the one step-ahead forecast}$$

$$NIM_t = c_{f,2} + \rho_{f,2} NIM_{t-2} + \sum_{j=1}^{2} \gamma_{2,f,j} F_{t-j} + \eta_{f,t}, \text{for the two step-ahead forecast}$$

$$\dots$$

$$\dots$$

$$NIM_t = c_{f,10} + \rho_{f,10} NIM_{t-10} + \sum_{j=1}^{2} \gamma_{10,f,j} F_{t-j} + \eta_{f,t}, \text{for the ten step-ahead forecast}$$

All other steps for generating forecasts from these models—such as, calculating the factors and combining the forecasts implied by each factor—are the same as for the iterative forecasts.

5. Single partial least squares (PLS). This model also requires specifying the ten different versions of our PLS NIM equations. Note that the PLS factors calculated for each forecast horizon will be different, since in each equation we need to control for the effect of a different lagged value of the dependent variable on the covariances between NIMs and lagged Treasury yields before calculating the PLS factors.

6. Treasury yields with forecast combination, and **7. 3-month and 10-year Treasury yields with forecast combination.** Similar to what we do for the three factor-based forecast-combination models, generating direct forecasts for the Treasury yields combination models requires specifying ten different versions of equation 2 for each yield $\tau \in \{0.25, 0.5, \cdots, 30\}$ for model 6 and $\tau \in \{0.25, 10\}$ for model 7. All other steps for generating forecasts from these models—such as, combining the forecasts implied by each factor—are the same as for the generating the iterative forecasts.

Finally, note that we do not have a direct forecasting counterpart for the VAR with 3-month and 10-year Treasury yields.

4.1.3 Iterative first differences models

Our iterative changes models are the same as those given in subsection 4.1.1 with the exception that first differences of NIMs, first differences of factors, and first differences of yields are used in place of their level counterparts. Note also that in the first differences models we have no lags of the first difference of NIMs and only one lag of the first difference of factors or the first difference of yields.

4.2 Methods for comparing performance

We use root mean squared (forecast) errors (RMSEs) to gauge a model's out-of-sample forecast performance and we use the Diebold-Mariano-West (DMW) test to determine the statistical significance of differences in forecast performance across models. Comparing forecast errors across models tells us *whether* one model forecasts NIMs better or worse than another out-of-sample but does not tell us *why*. There are two main reasons why a model may forecast better or worse out-of-sample. One is that the model that forecasts better out-of-sample captures the data over history better and thereby has better in-sample predictive content. The other is that the model that forecasts better out-of-sample has been less overfit in-sample. We can get some sense of the relative importance of these explanations by simple ocular comparisons of relative in-sample forecast performance across models with relative out-of-sample forecast performance, although doing so is imprecise. A more rigorous approach to considering this issue is by following the methodology developed by Rossi and Sekhposyan (2011). This approach decomposes the out-of-sample loss differential of the DMW test into three independent and interpretable components: marginal predictive content, over-fitting, and time-variation. To explain how this is done (in loose terms), let $\epsilon^2_{out,m,t+h|t,t+h-1}$ denote the squared h-step ahead conditional forecast error for model m, which is estimated on data extending out to quarter t and for which the forecast is conditional on NIM data out to quarter t and yields data out to quarter $t+h-1$, and let $\delta_{out,X,Y,t+h|t,t+h-1} = \epsilon^2_{out,X,t+h|t,t+h-1} - \epsilon^2_{out,Y,t+h|t,t+h-1}$ denote the time series of the out-of-sample h-period ahead conditional forecast loss differential for models X and Y.

Employing the same estimated models as were used to generate each $t+h$-step ahead NIM forecast (from which we then calculated $\epsilon_{out,m,t+h|t,t+h-1}$) we can generate in-sample forecasts, also h periods ahead. We can then calculate the squared h-period ahead conditional forecast error for the *very last* h-quarters of the estimation period—that is, $\epsilon^2_{in,m,t|t-h,t-1}$—and from this we calculate a time-series of in-sample h-step ahead loss differentials: $\delta_{in,X,Y,t|t-h,t-1} = \epsilon^2_{in,X,t|t-h,t-1} - \epsilon^2_{in,Y,t|t-h,t-1}$.

One part of Rossi and Sekhposyan (2011) decomposition involves (in loose terms) regressing $\delta_{out,X,Y,t+h|t,t+h-1}$ on $\delta_{in,X,Y,t|t-h,t-1}$, saving the estimated coefficient, β, from the regression, and then decomposing the full evaluation period out-of-sample loss differential of the DMW test, $E[\delta_{out,X,Y,t+h|t,t+h-1}]$, into $B = \beta \cdot E[\delta_{in,X,Y,t|t-h,t-1}]$ and $U = E[\delta_{out,X,Y,t+h|t,t+h-1}] - \beta \cdot E[\delta_{in,X,Y,t|t-h,t-1}]$, which they show are independent. They also develop test statistics for these two terms—denoted $\Gamma^{(B)}_P$ and $\Gamma^{(U)}_P$—that examine these terms' statistical significance. Importantly, these two terms allow one to understand two possible reasons *why* one model forecasts better or worse than the other. In particular, if the B term is positive and significant it means that the model that forecasts better out-of-sample also does so in-sample. Consequently, we can attribute the reason for that model's better out-of-sample forecast performance to better in-sample predictive content.[14] If the U term is significant it means that the model that forecasts better out-of-sample does so because the other model is over-fit in-sample.

Rossi and Sekhposyan (2011) also consider the time-variation in the out-of sample loss differentials, $\delta_{out,X,Y,t+h|t,t+h-1}$. That is, rather than just focusing on the average of $\delta_{out,X,Y,t+h|t,t+h-1}$ over the full out-of-sample evaluation period, they look at averages of $\delta_{out,X,Y,t+h|t,t+h-1}$ over rolling windows in the evaluation period and—through the test statistic $\Gamma^{(A)}_P$ that they formulate—provide a method to test the significance any observed time variation in these rolling-window averages. This component of the decomposition is then informative as to whether time variation accounts for relative forecast performance.

In summing up this discussion one important qualification is needed for the

[14] Note that it is possible for the B term to be negative and significant, which means that relative in-sample forecasting performance is predictive but misleading about out-of-sample forecast performance.

Rossi-Sekhposyan decomposition and this is that it can only be used to compare the relative forecasting performance across direct forecasting models—that is, the models in subsection 4.1.2—it can not be used to the relative forecasting performance across iterative forecasting models—that is, the models in subsections 4.1.1 and 4.1.3. For iterative models getting a sense of my one model forecasts better than another still requires an ocular comparison of in-sample and out-of-sample relative forecast performance. We would also note that the Rossi-Sekhposyan decomposition can be used when forecast models are estimated both over expanding windows and over rolling windows. That being said, the decomposition is more straightforward when models are estimated over rolling windows and for this reason we focus on results from models estimated over rolling window.

5 Forecasting results

5.1 Aggregate results

To compare the performance of the aggregate NIM models, we consider out-of-sample forecasts and we present our results that use 10-year rolling windows for estimation. (We have also generated the same results for recursive or expanding windows with broadly similar results.) The first 10-year window spans 1989q4 to 1999q4 and the assessment window spans 2000q1 to 2008q3. We focus on forecasts that go out to 10 steps ahead for the reason that the DFA Stress Tests extend out 9 quarters. In constructing the out-of-sample forecasts, we condition on the observation of the right-hand-side variables in each model, with the exception of lagged NIMs. That is, we condition on the observed value of Treasury yields or factors based on the Treasury yield curve. If lagged NIMs are required, as they are for forecast two or more steps ahead in our iterative forecast models, we use the NIM forecasts for preceding periods. This set-up means that when we say that forecasts are some given number of steps ahead we are referring to the NIM term.

We also generate in-sample forecasts to help understand relative out-of-sample forecast performance. Note also that the way we generate RSMEs for our in-sample forecast errors is the same way that Rossi and Sekhposyan (2011) do to perform their decomposition. That is, for each rolling window model for which we generate an out-of-sample forecast of some horizon and an associated out-of-sample forecast error, we also generate an in-sample forecast for that same horizon and an associated in-sample forecast error that is right at the end of the sample.

Figure 3 shows root mean squared errors (RMSE) for in-sample forecasts and out-of-sample forecasts for our iterative forecasting models specified in levels. Table 1 reports the numbers underlying these results along with the VAR model. (The VAR model has been dropped from the figure given how large its forecast errors are relative to all the other forecast methods.) As is evident from the in-sample results, all of the iterative levels models (with the exception of the VAR model), have lower RMSEs across all forecast horizons than Model 1—the no change forecast. Note also that the standard deviation of NIMs over the forecast evaluation period (currently not shown in the figures) is 0.23. This means that at most horizons and for most models the RMSE for NIMs are smaller than the standard deviation of NIMs, which suggests that the in-sample forecasts of NIMs are helpful in predicting future values of NIMs. These results do not for the most part translate to the out-of-sample forecast errors. Out-of-sample, only Model 6—yields with forecast combination—has a lower RMSEs than Model 1—the no change forecast. Moreover, beyond the one year horizon very few models have RMSEs that are smaller than the standard deviation of NIMs. These findings are notable given that most models of NIMs in the macro-banking literature—including those oriented to stress testing—use levels relationships between NIMs and interest rates.

Figure 4 shows RMSE for in-sample forecasts and out-of-sample forecasts for our direct forecasting models specified in levels and table 2 reports the numbers for these charts. For the most part these results are qualitatively very similar to those obtained for the iterative forecasting models. Indeed, out-of-sample the relative forecasting performance of the different direct forecasting models is practically the same as those obtained for the iterative forecasting models. Forecasting performance, unsurprisingly, also deteriorates in going between the in-sample and out-of-sample results, although for the direct forecasting models the deterioration is not so pronounced.

The Rossi-Sekhposyan decomposition of relative forecast performance can be used to compare the forecasts implied by these direct forecasting models and tables 3 show the results for these comparisons. Note that here the model that we are comparing all the others against is Model 6—yields with forecast combination—and not Model 1—the no change forecast. We use Model 6 to compare all our other models against because it is the model that yields the lowest RMSEs, moreover it puts our comparison of models on a more equal footing given that Model 1 really only uses information up until period $t - h$, whereas Model 6 and all of the other models use information up until period $t - 1$.

For all models with the exception of Model 5—the PLS model—Model 6 forecasts significantly better out-of-sample at the 8 step-ahead and 10 step-ahead horizons. In all of the cases for the 10-step ahead horizon and in all but one case for the 8-step ahead horizon, Model 6 forecasts better than the other four models for the reason that it suffers from less in-sample overfitting. At the 8-step ahead horizon Model 6 forecasts better than Model 4—PCR with forecast combination—for the reason that the former has better in-sample predictive content. At the 6-step ahead horizon Model 6 forecasts significantly better than only two models, Model 2—observed factors with forecast combination—and Model 7—3-month and 10-year yields with forecast combination—and as was the case with most of the 8 step-ahead and 10 step-ahead horizon results this is because Model 6 suffers from less in-sample overfitting.

Interestingly, Model 6 also forecasts significantly better than Model 7 at the

2-step ahead and 4 step-ahead horizons, which is perhaps not what one would immediately expect from a comparison of these two model's RMSEs in figure 4. This is particularly puzzling given that other models that have notably larger RMSEs at the 2 step-ahead and 4 step-ahead horizons, such as Model 2—Observed factors with forecast combination—are not found to have significantly poorer forecast performance than Model 6. What ultimately accounts for Model 6 forecasting significantly better than Model 7—despite the relatively small differences in RMSEs—is that quarter-by-quarter the difference in squared errors between models is very stable and this implies quite a high DMW test statistic. In contrast, other models, such as Model 2, that have larger average differences in forecast errors, have more reversals in terms of which model forecasts better. Note that this feature can also be seem from the relative magnitudes of $\Gamma_P^{(A)}$ for Models 2 and 7 in table 3. In brief, although $\Gamma_P^{(A)}$, which measures forecast performance stability, is insignificant for both models at the 2 step-ahead and 4 step-ahead horizons, it is notably higher for Model 2. In summing up our discussion of the Rossi-Sekhposyan decomposition we would note that for the most part Model 6—yields with forecast combination—forecasts better than the other models for the reason that it suffers less from in-sample overfitting. The only model for which this is not the case is Model 4—PCR with forecast combination—since in this case Model 6 forecasts better because it has better in sample predictive content.

Figure 5 shows RMSE for in-sample forecasts and out-of-sample forecasts for our iterative forecasting models specified in changes. These results do differ somewhat from those that we obtained from our models specified in levels. Although, the in-sample finding that all models forecast better than the no-change forecast at all horizons carries through, with the changes models, the improvement is (admittedly unsurprisingly) smaller. Out-of-sample, however, all models continue to forecast better than the no change model, which is, clearly, not what we found for the majority of the levels models. This finding is the same result that Campbell and Perron (1991) documented, which is that even when it is inconclusive as to whether a unit root is present in model series, assuming one and using a first-difference specification typically leads to better forecast performance. Among the changes models, Model 6—yields with forecast combination—no longer has the best forecast performance. Indeed, while out-of-sample it forecasts better than the no change forecast, of all the models we consider, it has the largest RMSEs. The other models are virtually indistinguishable from each other. The PLS and DFM model are marginally the best performing models over up until the 6 step-ahead forecast horizon. After that, however, the DFM model performs slightly better.

Figure 6 is similar to figure 3, the key difference being that the analysis extends from 2000 to 2012:Q3. [Currently we are not using the NIM series with adjustments for IOER and FAS 166/167.] As is evident our results do switch around a bit when we extend the sample to 2012:Q3. Now, no single model consistently performs better than the no change forecast: Model 4—PCR with forecast combination—forecasts better than the no change forecast up until the 5 step-ahead horizon, no model forecasts better between the 6 step-ahead and 8 step-ahead horizon, and Model 6—yields with forecast combination—forecasts better at the 9 and 10 step-ahead horizons. These differences are, however, minuscule. These poorer results may reflect the challenges for some of the approaches that we use for extracting factors from a large number of interest rates in the zero lower bound environment. [Though it might also reflect the fact that we a not using a NIM series that adjusts for some fairly significant structural changes at the end of the sample, so we will return to this issue once we have addressed that.]

5.2 BHC-level results

- To be completed.

6 Scenario analysis

We return now to our aggregate analysis and consider what our best-performing forecast models would imply for the paths of NIMs under the different 2013 Dodd Frank Act stress test scenarios.[15] We focus on last year's scenarios (rather than the 2014 DFA stress test scenarios) for the reason that both the 2013 scenarios featured developments in interest rates that on balance seems more stressful to banks net interest income.[16] These scenarios are shown in figure 7. As can be seen from the figure, the severely adverse scenario—which also featured a severe global recession—implies a downward shift in the profile of the yield curve (which at its largest is about 175 b.p. relative to baseline) accompanied by a modest flattening. The adverse scenario—which was motivated by a sudden jump in inflation that accompanies a moderate recession—implies a sizable upward shift of the yield curve (which at its maximum is on the order of about 250 b.p. relative to baseline) as well as a flattening in the slope of the yield curve (which at its maximum implies a yield curve that is about 100 b.p. flatter than baseline). In terms of changes in the yield curve, these two scenarios capture vastly different configurations.

In considering the effects of the scenarios on NIMs we focus on two models: Model 6—yields with forecast combinations—for the iterative levels models and Model 5—PLS—for the iterative changes models. We choose these models because relative to the other models that use similar set-ups, these models are the best performing. The published DFA stress test scenarios only contain two Treasury yields—specifically, the 3-month and the 10-year yields—whereas our forecast mod-

[15]See, "2013 Supervisory Scenarios for Annual Stress Tests Required under the DFA Stress Testing Rules and the Capital Plan Rule" available at http://www.federalreserve.gov/bankinforeg/bcreg20121115a1.pdf.

[16]The interest rate developments in the 2014 DFA stress test severely adverse scenario are reasonably similar to those of the 2013 severely adverse scenario, although because long rates at the scenario's jumping off point were so much higher, the 2014 scenario did feature a larger flattening. The 2014 DFA stress test adverse scenario was very different to 2013 adverse scenario since it featured a large steepening in the yield curve. Such a scenario while stressful for banks' unrealized capital gains on securities, would not be stressful for their net interest income.

els require many more. Christensen and Lopez (2014) have developed a model for obtaining a full Treasury yield curve (and other corporate yield curves as well) from the relatively small number of interest rate variables included in the published DFA stress test scenarios and so for the other rates that we need for our forecast models we used the Treasury yields that they have constructed using their model. (We can do something similar with the dynamic factor model that we described in section 4.1.1, which, just for internal consistency with the rest of the paper, we will likely do this in future drafts.)

Figure 8 plot the paths of NIMs under the three DFA stress tests scenarios given the two models that we use. The paths of NIMs in the baseline and severely adverse scenarios are shown to the left and the paths of NIMs in the baseline and adverse scenarios are shown to the right. The implications of the point forecasts appear sensible. The severely adverse scenario is unambiguously unfavorable to NIMs, since the yield curve is flatter—albeit only by a small amount—which reduces banks' returns from maturity transformation, and the yield curve is lower, which reduces banks' returns from the provision of transactions services. This outcome is very evident from from the left-hand panels of figure 8, which for both models show a lower path of NIMs under the severely adverse scenario.

The adverse scenario has more ambiguous implications for NIMs, since the yield curve is flatter, which reduces banks' returns from maturity transformation, but is also higher, which boosts banks' returns from the provision of transactions services, since now banks have greater scope to pay deposit rates below that of short-term market rates. The right-hand panels of figure 8 do seem to suggest forces pulling in different directions. For the iterative levels version of Model 6—yields with forecast combinations—the path of NIMs in the adverse scenario is marginally higher than the baseline over the first year of the scenario and marginally lower over the second year. For the iterative changes version of Model 5—PLS—the path of NIMs is higher throughout the scenario, indicating that the positive impact on NIMs from the provision of transactions services outweighs the negative impact from maturity transformation.

Figure 8 also consider the uncertainty of the forecasts. In addition, to showing point estimates of the paths of NIMs, the figure also shows bands surrounding the central forecast, which are constructed by adding and subtracting to the point forecast the RMSEs for the forecast combination model. Strikingly, the size of the uncertainty bands dwarfs the variation implied by the scenarios for NIMs and as such the two severe scenarios are statistically indistinguishable from the baseline and from each other. This degree of uncertainty around the paths of key BHC variables could compromise one of the key goals of stress tests, which is to maintain the confidence of bank creditors and counterparties in periods of financial sector stress by generating forward-looking estimates of bank capital under stressful economic conditions.

7 Conclusion

Over the evaluation period 2000q1 to 2008q3, a few models outperform the no change model in forecasting NIMs. These include—among the levels models that we consider—the yields with forecast combination model, and—among the changes versions of the models that we consider—the DFM with forecast combinations model and the PLS model. Over the evaluation period 2000q1 to 2012q3, the models that outperform the no change model are (depending on the horizon) PCR with forecast combination model and the yields with forecast combination model. But these victories are only small, as the variance of the forecast error is large, especially when compared to the observed variation of NIMs. Accordingly, even stress test scenarios that have vastly different implications for the configuration of interest rates contain relatively little information to forecast NIMs. This degree of uncertainty around the paths of key BHC variables could compromise one of the key goals of stress tests, which is to maintain the confidence of bank creditors and counterparties in periods of financial sector stress by generating forward-looking estimates of bank capital under stressful economic conditions.

Our findings have implications beyond the study of net interest margins. Scenario-based stress tests have been used widely to maintain confidence in the capital adequacy of banks, including in periods of crises. Yet, they are predicated on the assumption that the unfavorable macro scenarios used in the stress tests capture the unfavorable outcomes that could materialize for key balance sheet variables. An array of time series methods found by others to be near the frontier of forecast performance produce projections conditional on radically different scenarios for yields that are statistically indistinguishable.

Going forward we plan to investigate whether other macro variables may be relevant for forecasting NIMs and may be able to stress NIMs more effectively. For example, we plan to consider some of the variables emphasized by the micro-banking literature in explaining NIMs, such as, the degree of competition and the degree of volatility of interest rates. We also plan to consider other macro variables such as those included, in addition to Treasury yields, in the DFA stress test scenarios, to see whether stressful outcomes for these variables might also imply stresses for NIMs. Clearly, if we find that these other possible macro variables do not provide any way to stress NIMs, it may suggest the need to consider other ways to introduce stresses to NIMs. Being able to obtain stressful outcomes for NIMs is necessary if stress-test results are to leave bank creditors and counterparties confident that when banks balance sheets come under stress, banks will be able to continue to meet their obligations.

References

Alessandri, P. and B. Nelson (2012). Simple banking: profitability and the yield curve. Technical report, Bank of England.

Angbazo, L. (1994). Commercial bank net interest margins, default risk, interest rate risk, and off-balance sheet banking. *Journal of Banking and Finance 21*(6), 55–87.

Campbell, J. and P. Perron (1991, January). Pitfalls and opportunities: What macroeconomists should know about unit roots. *NBER Macro Annual*, 141–220.

Christensen, J. and J. Lopez (2014). Constructing yield curves from macro scenarios of bank stress tests. Technical report, Federal Reserve Bank of San Francisco.

Clarida, R. and D. Coyle (1984, May). Conditional projection by means of kalman filtering. Technical report, NBER.

Covas, F., B. Rump, and E. Zakrajsek (2012, September). Stress-testing u.s. bank holding companies: A dynamic panel quantile regressions approach. Technical report, Federal Reserve Board.

Diebold, F. X. and C. Li (2006, February). Forecasting the term structure of government bond yields. *Journal of Econometrics 130*(2), 337–364.

Duffee, G. R. (2012, July). Forecasting interest rates. Economics Working Paper Archive 599, The Johns Hopkins University,Department of Economics.

English, W. B. (2002, December). Interest rate risk and bank net interest margins. BIS Quarterly Review 2002-26, Bank for International Settlements.

English, W. B., S. J. V. den Heuvel, and E. Zakrajsek (2012, May). Interest rate risk and bank equity valuations. Finance and Economics Discussion Series 2012-26, Board of Governors of the Federal Reserve System (U.S.).

FRB (2012). Comprehensive capital analysis and review 2012: Methodology and results for stress scenario projections. Report, Board of Governors of the Federal Reserve System (U.S.).

FRB (2014). Dodd-frank act stress test 2014:supervisory stress test methodology and results. Report to congress, Board of Governors of the Federal Reserve System (U.S.).

Groen, J. and G. Kapetanios (2009, September). Revisiting useful approached to data-rich macroeconomic forecasting. Technical report, Federal Reserve Bank of New York.

Hirtle, B., A. Kovner, and J. Vickery (2013, September). The capital and loss assessment under stress scenarios (class) model. Technical report, Federal Reserve Bank of New York.

Ho, T. S. Y. and A. Saunders (1981). The determinants of bank interest margins: Theory and empirical evidence. *Journal of Financial and Quantitative Analysis 16*(04), 581–600.

Rossi, B. and T. Sekhposyan (2011). Understanding models' forecasting performance. *Journal of Econometrics 164*, 158–172.

Saidenberg, M. and T. Schuermann (2003, September). The new basel capital accord and questions for research. Technical report, Wharton, Financial Institutions Center.

Saunders, A. and L. Schumacher (2000, December). The determinants of bank interest rate margins: an international study. *Journal of International Money and Finance 19*(6), 813–832.

Table 1: RMSEs – Iterative Forecasts, Regressions of Levels on Levels, and 2000q1 to 2008q3 Evaluation Window

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9	Step 10
In-Sample RMSEs										
1. No-Change Forecast	0.11	0.14	0.15	0.18	0.20	0.22	0.23	0.23	0.25	0.27
2. Observed Factors with F. Combination	0.10	0.12	0.12	0.13	0.15	0.16	0.16	0.16	0.16	0.16
3. DFM with F. Combination	0.09	0.11	0.12	0.12	0.14	0.15	0.15	0.15	0.15	0.15
4. PCR with F. Combination	0.10	0.14	0.15	0.17	0.19	0.20	0.20	0.20	0.20	0.20
5. PLS	0.12	0.15	0.17	0.19	0.21	0.21	0.21	0.22	0.21	0.20
6. Yields with F. Combination	0.09	0.11	0.12	0.13	0.14	0.15	0.15	0.15	0.14	0.14
7. 3M and 10Y with F. Combination	0.09	0.11	0.12	0.12	0.14	0.14	0.14	0.14	0.13	0.13
8. VAR on 3M and 10Y Yields	0.17	0.27	0.35	0.38	0.43	0.46	0.47	0.49	0.50	0.50
Out-of-Sample RMSEs										
1. No-Change Forecast	0.11	0.14	0.15	0.18	0.20	0.22	0.24	0.24	0.27	0.29
2. Observed Factors with F. Combination	0.12	0.16	0.19	0.25	0.30	0.35	0.40*	0.44*	0.50*	0.55*
3. DFM with F. Combination	0.11	0.16	0.19	0.23	0.27	0.31	0.35	0.38*	0.42*	0.46*
4. PCR with F. Combination	0.11	0.14	0.17	0.20	0.23	0.27	0.30	0.32	0.36*	0.39*
5. PLS	0.12	0.15	0.18	0.21	0.23	0.25	0.27	0.28	0.30	0.32
6. Yields with F. Combination	0.11	0.14	0.14	0.16	0.18	0.20	0.21	0.21	0.22	0.23
7. 3M and 10Y with F. Combination	0.11*	0.14	0.17*	0.20	0.23	0.25	0.28	0.29	0.32	0.34
8. VAR on 3M and 10Y Yields	0.17*	0.24*	0.29*	0.30*	0.30*	0.31*	0.33	0.36	0.39	0.42

'*' denotes significance at the 95% level of the difference in RMSEs relative to the no-change forecast.

Table 2: RMSEs – Direct Forecasts, Regressions of Levels on Levels, and 2000q1 to 2008q3 Evaluation Window

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9	Step 10
In-Sample RMSEs										
1. No-Change Forecast	0.11	0.14	0.16	0.18	0.20	0.22	0.23	0.23	0.25	0.27
2. Observed Factors with F. Combination	0.10	0.12	0.13	0.15	0.18	0.20	0.20	0.19	0.19	0.18
3. DFM with F. Combination	0.10	0.11	0.12	0.13	0.15	0.17	0.18	0.17	0.16	0.16
4. PCR with F. Combination	0.11	0.14	0.16	0.18	0.20	0.21	0.22	0.22	0.22	0.22
5. PLS	0.10	0.12	0.15	0.18	0.21	0.24	0.27	0.28	0.25	0.20
6. Yields with F. Combination	0.10	0.12	0.13	0.14	0.17	0.18	0.19	0.18	0.19	0.19
7. 3M and 10Y with F. Combination	0.10	0.12	0.13	0.14	0.16	0.18	0.19	0.18	0.18	0.18
Out-of-Sample RMSEs										
1. No-Change Forecast	0.11	0.14	0.15	0.18	0.21	0.23	0.24	0.25	0.27	0.29
2. Observed Factors with F. Combination	0.12	0.16	0.20	0.25	0.30*	0.34*	0.35*	0.36*	0.39*	0.41*
3. DFM with F. Combination	0.11	0.16	0.20*	0.23	0.26	0.29	0.30	0.31	0.34*	0.37
4. PCR with F. Combination	0.11	0.14	0.17	0.20	0.23	0.26	0.28	0.28	0.31*	0.34
5. PLS	0.11	0.16	0.18	0.20	0.23	0.25	0.26	0.26	0.26	0.27
6. Yields with F. Combination	0.11	0.13	0.15	0.17	0.19	0.20	0.21	0.21	0.22	0.23
7. 3M and 10Y with F. Combination	0.11	0.15	0.18*	0.20	0.24	0.26	0.27	0.27	0.29	0.31

'*' denotes significance at the 95% level of the difference in RMSEs relative to the no-change forecast.

Table 3: Rossi-Sekhposyan Decomposition and Tests Against Model 6, Yields with Forecast Combination

		Model 2. Obs. Factors†	Model 3. DFM†	Model 4. PCR†	Model 5. PLS	Model 7. 3M & 10Y†
2 step ahead	DMW	1.208	1.816	0.895	0.914	2.216*
	$\Gamma_P^{(A)}$	7.306	5.515	6.353	5.752	6.128
	$\Gamma_P^{(B)}$	0.421	-0.517	3.662*	0.297	-0.976
	$\Gamma_P^{(U)}$	1.186	1.857	0.339	0.881	2.370*
4 steps ahead	DMW	1.777	1.677	1.036	0.816	2.948*
	$\Gamma_P^{(A)}$	7.095	4.765	6.471	6.125	5.086
	$\Gamma_P^{(B)}$	-0.266	-0.647	3.632*	1.349	1.642
	$\Gamma_P^{(U)}$	1.851	1.692	0.302	0.770	2.453*
6 steps ahead	DMW	2.976*	1.816	1.772	1.013	3.366*
	$\Gamma_P^{(A)}$	5.402	5.948	5.653	6.625	6.619
	$\Gamma_P^{(B)}$	1.229	-0.793	3.573*	1.764	1.028
	$\Gamma_P^{(U)}$	2.887*	2.033*	1.729	0.344	3.228*
8 steps ahead	DMW	3.410*	2.924*	2.193*	0.886	2.929*
	$\Gamma_P^{(A)}$	6.617	5.516	6.488	6.851	5.623
	$\Gamma_P^{(B)}$	0.482	-0.556	3.074*	1.800	-0.788
	$\Gamma_P^{(U)}$	3.399*	3.690*	1.631	0.303	3.335*
10 steps ahead	DMW	3.342*	5.785*	3.646*	0.610	3.563*
	$\Gamma_P^{(A)}$	4.646	5.932	5.268	5.037	4.783
	$\Gamma_P^{(B)}$	0.100	0.681	1.907	-0.535	-0.936
	$\Gamma_P^{(U)}$	4.458*	6.287*	3.283*	0.687	3.727*

'*' denotes significance at the 95% level of the difference in RMSEs relative to the no-change forecast. † Models employ forecast combination

Table 4: RMSEs Iterative Forecasts, Regressions of Level on Levels, 2000q1 to 2012q3 Assessment Window

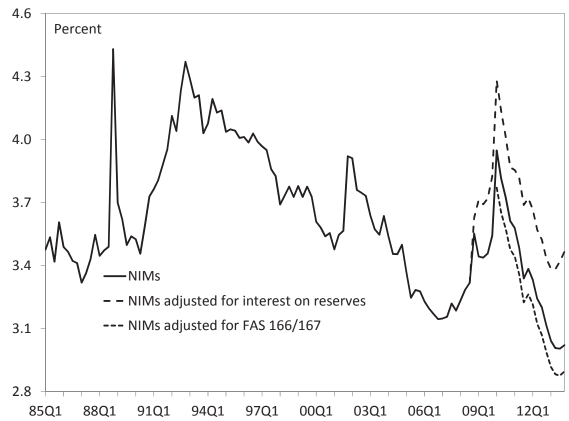
	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8	Step 9	Step 10
1. No-Change Forecast	0.11	0.15	0.17	0.20	0.22	0.24	0.26	0.26	0.27	0.28
2. Observed Factors with F. Combination	0.10	0.12	0.13	0.14	0.14	0.15	0.15	0.15	0.16	0.16
3. DFM with F. Combination	0.10	0.11	0.12	0.13	0.13	0.14	0.15	0.14	0.15	0.15
4. PCR with F. Combination	0.11	0.14	0.15	0.17	0.18	0.19	0.19	0.19	0.19	0.20
5. PLS	0.12	0.15	0.16	0.18	0.19	0.19	0.20	0.20	0.20	0.20
6. Yields with F. Combination	0.10	0.12	0.13	0.14	0.14	0.15	0.16	0.15	0.15	0.15
7. 3M and 10Y with F. Combination	0.10	0.12	0.13	0.14	0.14	0.15	0.15	0.15	0.15	0.14
8. VAR on 3M and 10Y Yields	0.18	0.25	0.31	0.34	0.38	0.40	0.41	0.42	0.43	0.43
Out-of-Sample RMSEs										
1. No-Change Forecast	0.11	0.15	0.17	0.20	0.22	0.25	0.27	0.28	0.29	0.31
2. Observed Factors with F. Combination	0.12	0.16	0.19	0.23	0.27	0.32	0.36	0.39	0.43	0.46
3. DFM with F. Combination	0.12	0.16	0.19	0.22	0.26	0.30	0.32	0.35	0.38	0.41
4. PCR with F. Combination	0.12	0.14	0.16	0.19	0.22	0.26	0.28	0.29	0.32	0.34
5. PLS	0.12	0.16	0.18	0.21	0.23	0.26	0.28	0.29	0.30	0.31
6. Yields with F. Combination	0.12	0.16	0.20	0.23	0.26	0.28	0.29	0.29	0.29	0.29
7. 3M and 10Y with F. Combination	0.12*	0.15	0.19*	0.21	0.25	0.27	0.30	0.30	0.33	0.34
8. VAR on 3M and 10Y Yields	0.18*	0.24*	0.29*	0.29*	0.29*	0.29	0.30	0.31	0.33	0.35

Table 5: Out-of-Sample RMSE for the 1-Step Ahead Forecast: Bank-Specific Models and 2007q4 to 2013q1 Evaluation Window

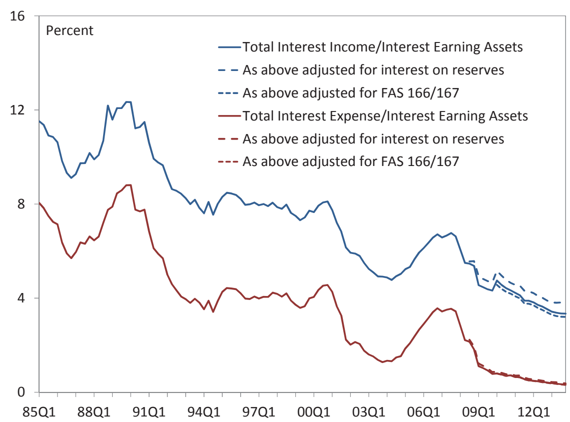
	Recursive			Rolling				
	1. F. Comb	4. PCR	4. PLS	1. F. Comb	3. PCR	4. PLS	7. No Change	SNL
JPMORGAN CHASE & CO	0.17	0.18	0.18	0.18	0.18	0.17	0.15	0.37
BANK OF AMER CORP	0.24	0.24	0.25	0.25	0.24	0.30	0.26	0.29
CITIGROUP	0.22	0.24	0.23	0.29	0.24	0.30	0.21	0.46
WELLS FARGO & CO	0.14	0.14	0.16	0.15	0.14	0.12	0.12	0.54
U S BC	0.10	0.13	0.12	0.11	0.13	0.12	0.10	0.10
PNC FNCL SVC GROUP	0.21	0.22	0.22	0.21	0.22	0.20	0.18	0.17
BB&T CORP	0.14	0.18	0.16	0.14	0.18	0.12	0.13	0.15
SUNTRUST BK	0.25	0.21	0.16	0.14	0.21	0.13	0.10	0.17
CAPITAL ONE FC	1.06	0.90	0.91	1.13	0.90	1.08	0.53	0.91
REGIONS FC	0.16	0.19	0.14	0.16	0.19	0.14	0.12	0.11
FIFTH THIRD BC	0.30	0.32	0.32	0.30	0.32	0.29	0.34	0.22
KEYCORP	0.47	0.44	0.45	0.50	0.44	0.48	0.49	0.44
M&T BK CORP	0.11	0.14	0.15	0.12	0.14	0.13	0.12	0.11
COMERICA	0.16	0.20	0.17	0.17	0.20	0.19	0.17	0.09
HUNTINGTON BSHRS	0.23	0.13	0.10	0.27	0.13	0.18	0.11	0.10
ZIONS BC	0.23	0.19	0.22	0.22	0.19	0.21	0.20	0.21
BANK OF NY MELLON CORP	0.32	0.31	0.32	0.33	0.31	0.29	0.37	0.27
STATE STREET CORP	0.22	0.26	0.25	0.22	0.26	0.26	0.20	0.18
NORTHERN TR CORP	0.11	0.10	0.12	0.11	0.10	0.13	0.12	0.15

Figure 1: Interest Income, Interest Expenses and Short-term Treasury Rates

Net Interest Margins



Interest Income/Interest Earning Assets and Interest Expenses/Interest Earning Assets



Treasury Yields at Different Maturities

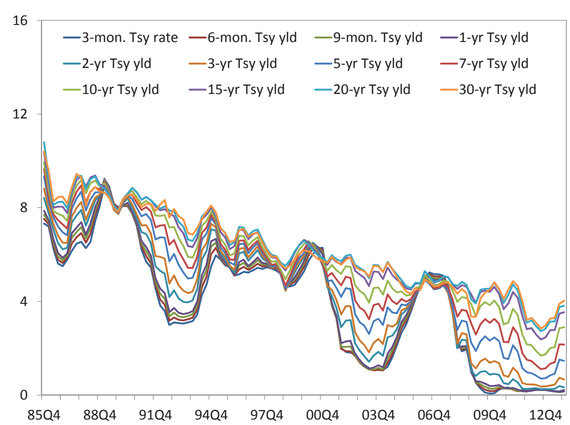
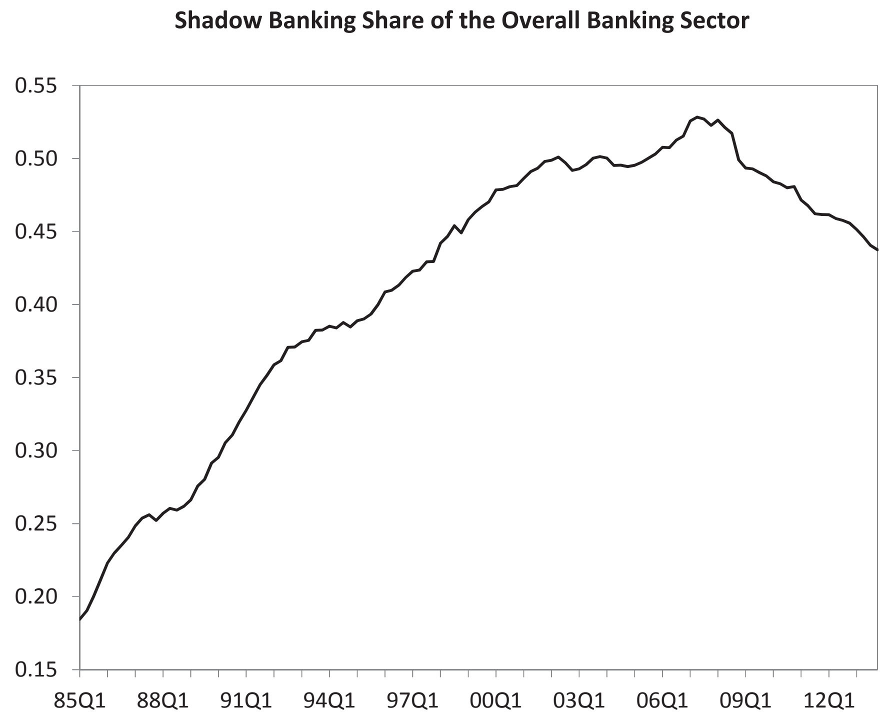


Figure 2: Variables from the Micro-banking Literature



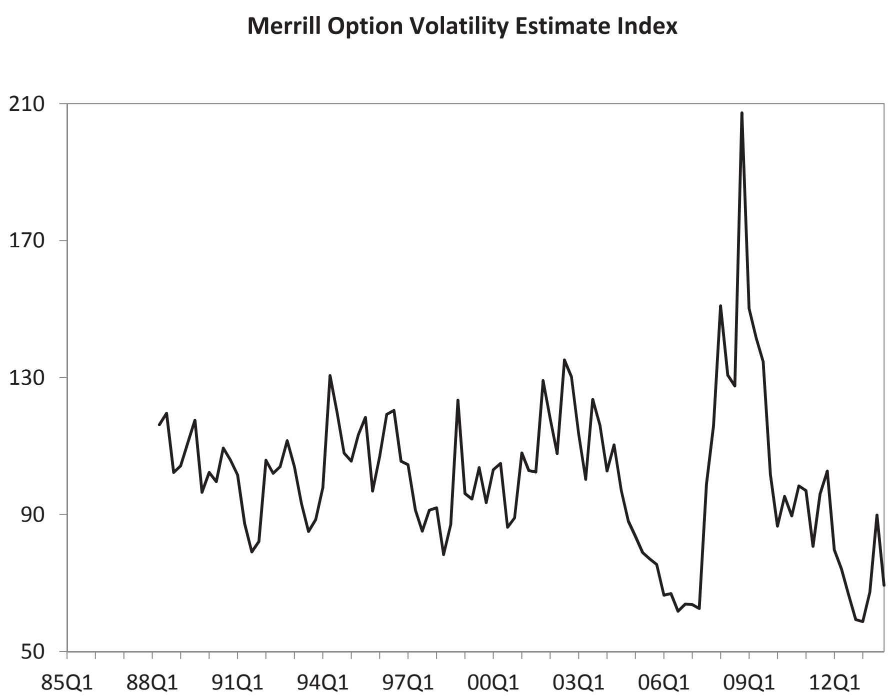


Figure 3: Root Mean Squared Errors of Iterative Forecasts: 2000q1 to 2008q3

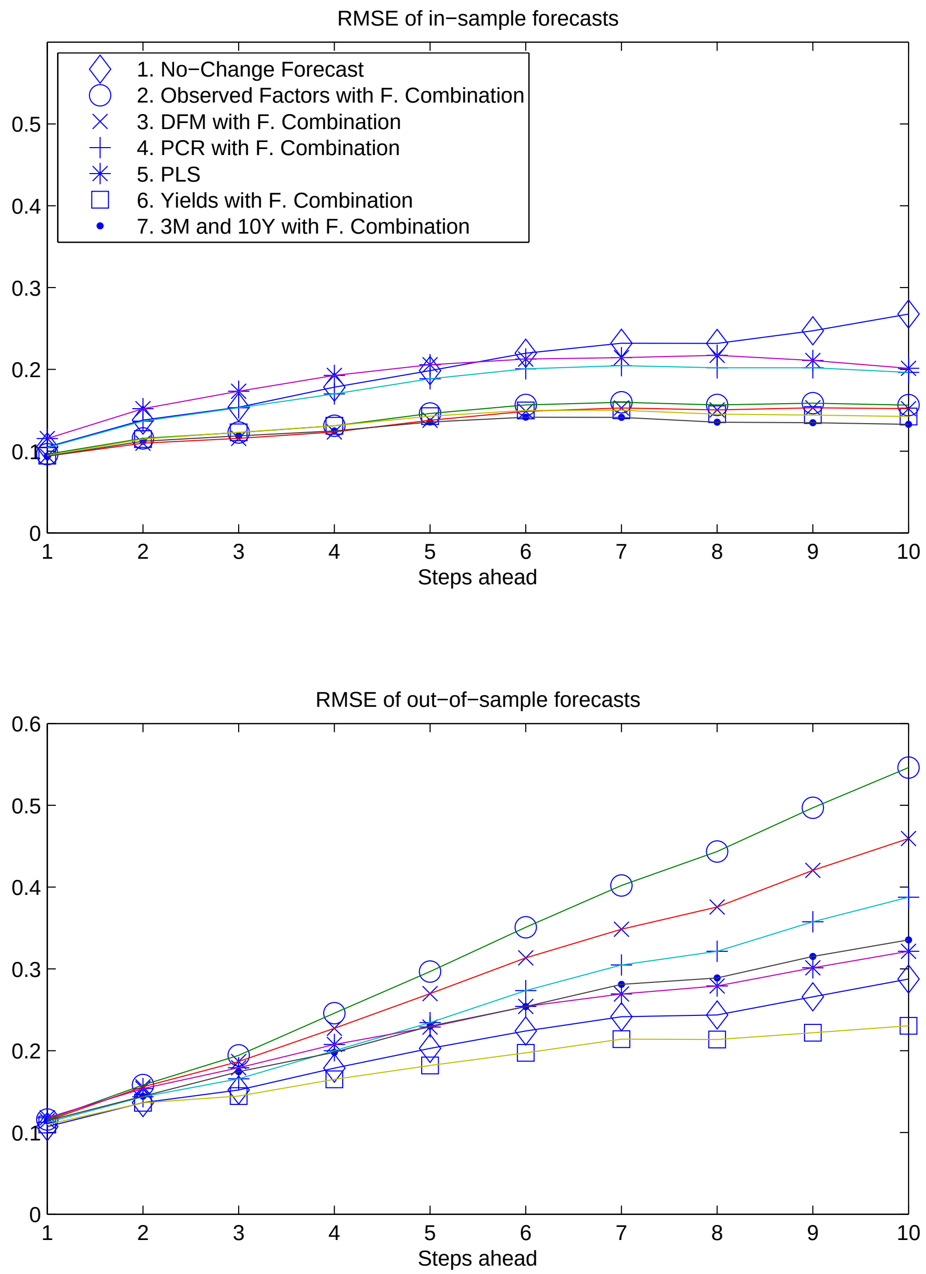
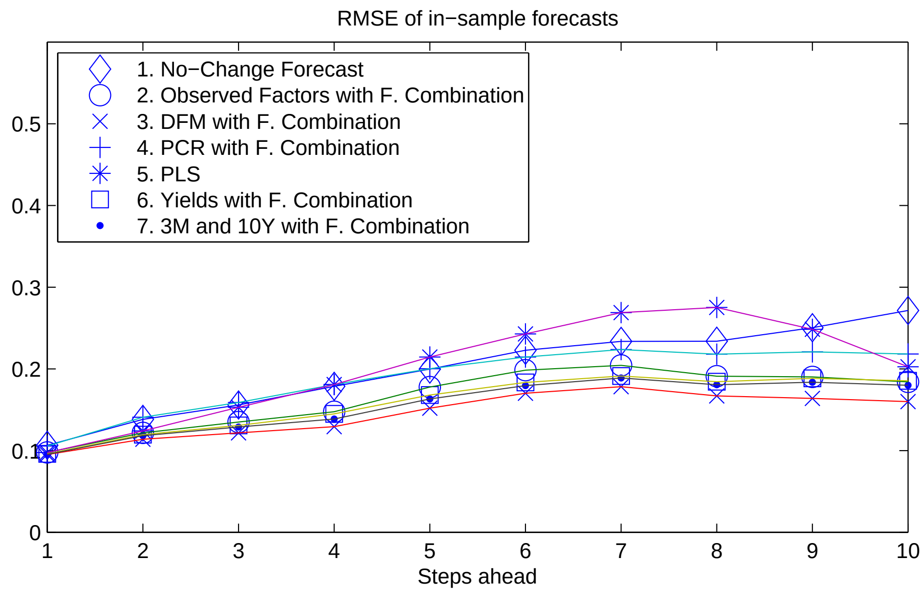


Figure 4: Root Mean Squared Errors of Direct Forecasts: 2000q1 to 2008q3



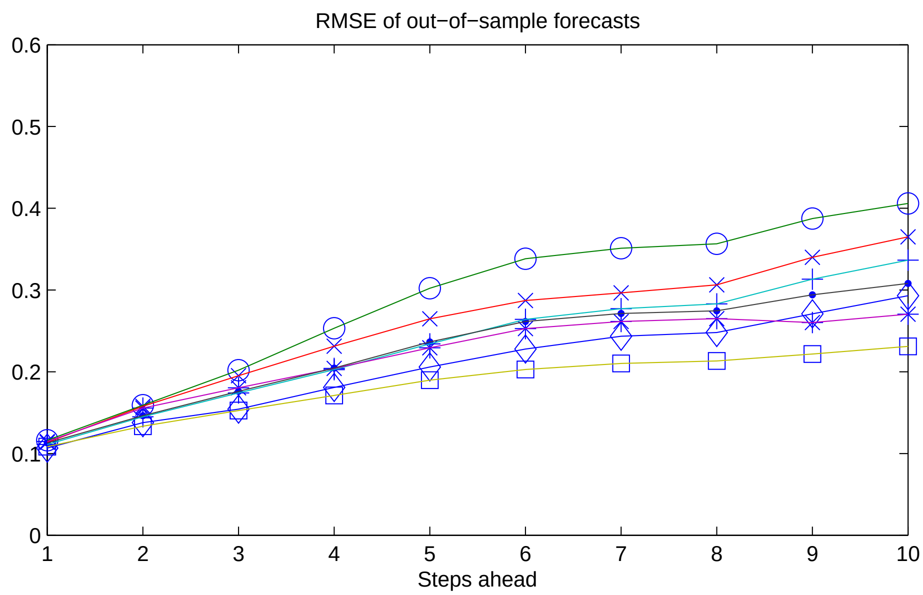


Figure 5: Root Mean Squared Errors of Iterative Forecasts – Alternative Specification (changes on changes): 2000q1 to 2008q3

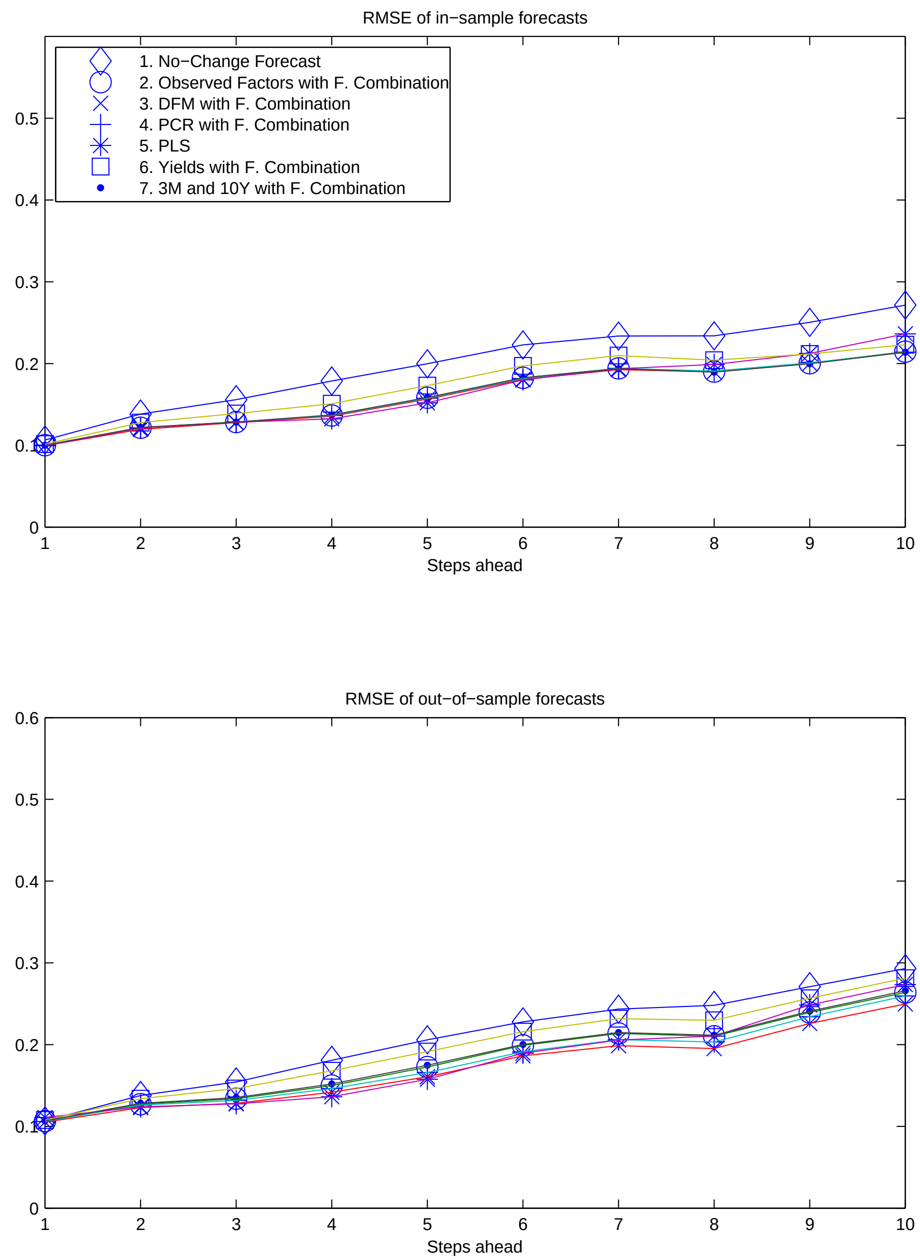


Figure 6: Root Mean Squared Errors of Iterative Forecasts: 2000q1 to 2012q3

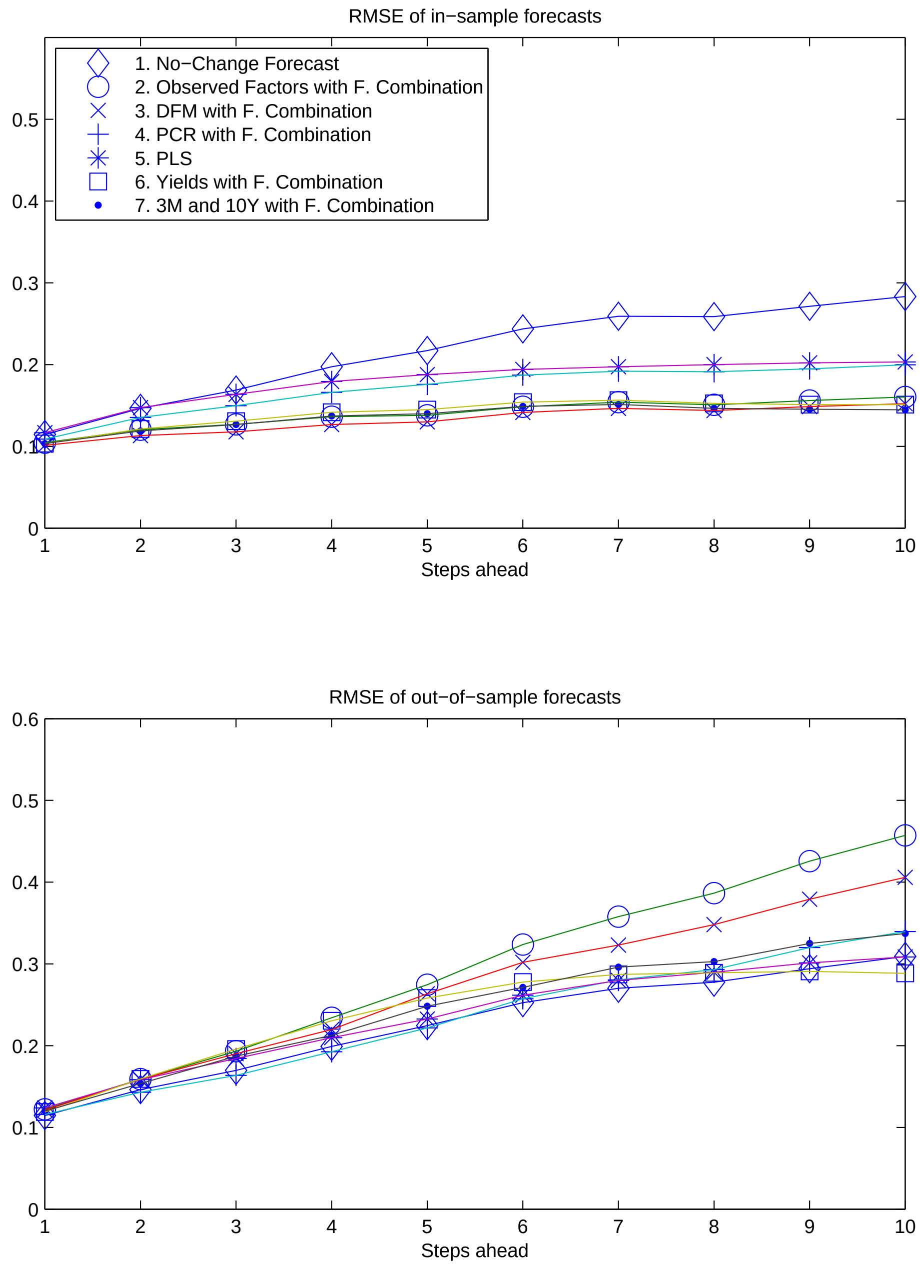


Figure 7: Three-month and Ten-year yields in the 2013 DFA Stress Test Scenarios

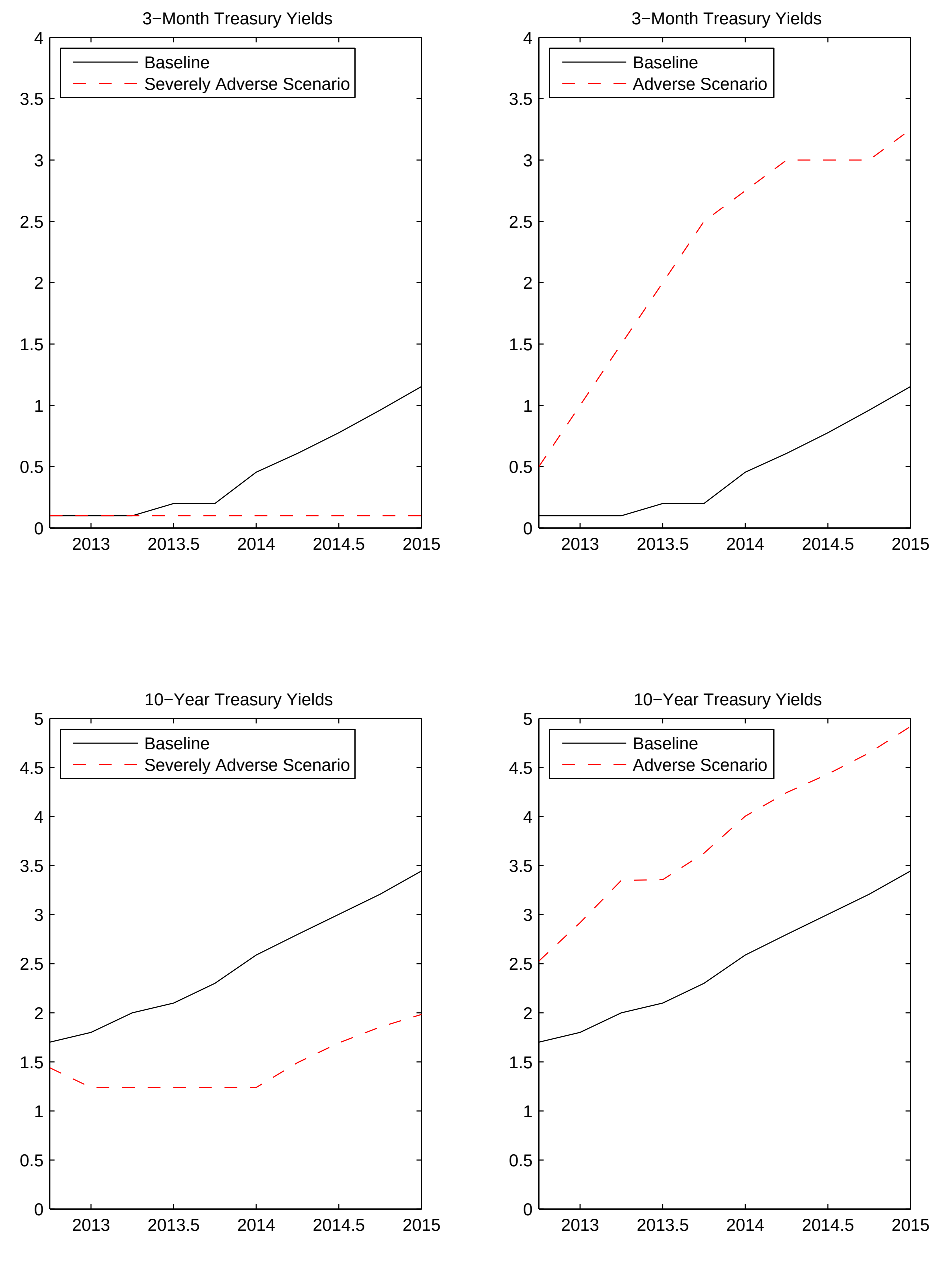


Figure 8: Forecast for NIMs Conditional on the 2013 DFA Stress Test Scenarios