Forecaster overconfidence and market survey performance

This draft: March 25, 2013

ABSTRACT

We document using the ZEW panel of stock market forecasters that weak forecasters tend to be overconfident in the sense that they provide extreme forecasts and their confidence intervals are too narrow relative to their rate of success. Moderate filters based on short rolling windows are somewhat successful in improvements in predictability vs. a forecast based on the historical equity premium.

Richard Deaves,* McMaster University (Canada)

Jin Lei, McMaster University (Canada)

Michael Schröder, ZEW & Frankfurt School of Finance and Management (Germany)

***Corresponding author: deavesr@mcmaster.ca / 905-529-7070 ext. 23976. The other authors' emails are leij3@mcmaster.ca and schroeder@zew.de.**

1. Introduction

Abundant research has documented the pitfalls of overconfidence in financial decision-making. For example, investors so affected are likely to trade too much (e.g., Barber and Odean (2000)) and under-diversify (Goetzmann and Kumar (2008)), while susceptible managers are prone to excessive M&A activity (Malmendier and Tate (2008)) and market entry (Camerer and Lovallo (1999)). Daniel Kahneman, in his recent bestseller *Thinking, Fast and Slow* (2011), argues that professional forecasters are often bested by simple algorithms because they "try to be too clever, think outside the box, and consider complex combinations of features in making their predictions (p. 224)." This is another way of saying that they are overconfident: they believe they know more than they actually do.

While forecast disagreement can occur because of heterogeneity in information, information-updating frequency and model choice (Capistran and Timmerman (2009a)), behavioral bias might also contribute. The purpose of this paper is to explore the impact of overconfidence on forecasting stock market returns in the context of surveys of professional forecasters. The questions we ask ourselves are these. Does overconfidence weaken forecast accuracy? And, given that there is heterogeneity in performance in part induced by heterogeneity in overconfidence, is there a payoff to filtering out weaker forecasters to improve survey accuracy?

Excess market returns have proved to be notoriously difficult to predict out of sample. While there is an extensive literature documenting return predictability within sample using such fundamental variables as dividend yields, interest rates and term spreads, as pointed out by Goyal and Welch (2008), this has not translated into out-of-sample performance as (typically) measured by out-of-sample R^2 (OS-R^2) relative to a naïve

benchmark such as the historical average equity premium.[1] Nevertheless Rapach, Strauss and Zhou (2010) have shown that a combination forecast methodology whereby several predictive variables are optimally combined can lead to a modicum of out-of-sample success. The same holds in Ferreira and Santa-Clara (2011) where the components of stock market returns are predicted separately. Nevertheless predictability is modest, in the former case being less than 4% (using quarterly data) and in the latter case less than 2% (using monthly).

While it is logical to expect that panels of professional forecasters, not only with such predictive variables at their disposal but also armed with experiential judgment, should easily be able to outperform naïve benchmarks, the Kahneman perspective encourages skepticism in this regard. Take the ZEW survey in Germany, which since February 2003 has solicited point forecasts for the DAX.[2] While the mean forecast of the excess market return coming from this survey produces OS-R^2 of 6.19% (p-value=0.073) for March 2003-June 2010, success is concentrated in the first year as OS-R^2 = 1.09% (p-value=0.239) during February 2004-June 2010.[3]

Some forecasters are weaker than others and these may skew the consensus. We conjecture that weak forecasters may be weak in part because they are more overconfident than other forecasters. One possibility is that, relying too much on intuition, they have a tendency to make extreme forecasts. Denrell and Fang (2010) document that those who have made a very accurate recent prediction – since markets are volatile this often implies an extreme prediction – are likely to be inferior forecasters going forward. Indeed our data indicate that survey respondents with higher forecast standard deviations have higher mean squared prediction errors (MSPEs).

[1] See Neely, Rapach, Tu and Zhou (2010) for many references on return predictability.

[2] **T**he DAX is an index composed of the 30 largest and most important German companies traded on the German Stock Exchange in Frankfurt.

[3] The ZEW survey actually requests six-month DAX forecasts. The reported OS-R^2s are based on imputed one-month forecasts (as described below) so (given this imputation) the February 2003 survey solicits forecasts for March 2003.

Overconfidence can also manifest itself in the tendency to be too sure of one's views, leading to overly narrow confidence intervals.[4] This tendency is echoed in the model of Daniel, Hirshleifer and Subrahmanyam (1998), where overconfident investors put too much stock in private information and exert pressure on prices in the direction of their information, with the result that if such investors dominate markets overreaction and eventual reversal in security prices can ensue. We further document that forecasters whose confidence intervals are wide enough to contain the eventual DAX realization more often than other forecasters are better forecasters in the sense that they have lower MSPEs. This is not tautological because better forecasters actually have *narrower* confidence bounds.

Next consensus forecast improvement is considered. We show that filtering out from the survey inferior forecasters can lead to modest but statistically significant improvements in accuracy. For example, if we drop the 30% of forecasters whose *prior* MSPEs over the preceding three forecasts was highest, OS-R^2 reaches 4.18% which is significant at 2%. It is not obvious that this should be so since one might expect that inferior forecasts would be as likely to be too high (relative to the realization) as too low. Evidently, some error clustering is occurring, consistent with what has been found for analysts (Hirshleifer and Teoh (2003)).

In what follows, we begin by providing appropriate background on the ZEW DAX survey. In section 3 we explore the characteristics of successful forecasters and the contributing role of overconfidence. In the penultimate section, we document that filtering out weaker forecasters can lead to meaningful out-of-sample predictability. Finally, in section 5, we discuss our findings and sum up.

[4] Deaves, Lüders and Schröder (2010) have previously documented that the ZEW forecasters are overconfident in this sense. Ben-David, Graham and Harvey (2013) have performed a similar exercise using a U.S. panel of market forecasts.

2. ZEW survey

The *ZEW Finanzmarkttest* is a monthly survey of over 300 private sector forecasters in Germany. From 1991 to the present it has solicited predicted directional changes (rise/fall/unchanged) in a series of key macroeconomic and financial market variables for the key industrialized economies as of six months in the future.[5] Starting in February 2003, ZEW survey respondents were also asked to provide quantitative forecasts and confidence intervals for the DAX. Specifically, point estimates for the DAX six months in the future, as well as lower and upper bounds forming 90% confidence intervals began to be solicited. These are the forecasts that we investigate here.[6] The cleaned dataset has over 20,000 forecaster-survey observations, with a survey minimum/mean/maximum of 135/228/269.

To avoid the overlapping data problem inherent in the fact that forecasts are made monthly for six-month-ahead DAX levels, we here follow the methodology of Deaves, Lüders and Schröder (2010), where one-month point forecasts and 90% confidence intervals are imputed from six-month. It is assumed that forecasters believe that the growth rate in the DAX will be constant over the next six months. More specifically, letting L6, F6 and U6 be the six-month interval lower bound, forecast point estimate and interval upper bound respectively, the one-month forecast point estimate (F1) is calculated as*:*

$$(1) \quad F1 = \left(\frac{F6}{DAX0}\right)^{1/6} * DAX0$$

where *DAX0* is the current level of the DAX. On the assumption of *i.i.d.* DAX one-month returns, the standard deviation of one-month returns is $1/\sqrt{6}$ times the six-month

[5] Most of these individuals work for a commercial bank, investment bank, insurance company or investment department of a large German company. For example, participants are asked to predict the inflation rate, long-term and short-term interest rates, economic activity, and stock market levels for these countries.
[6] The final survey in our dataset is May 2010.

standard deviation. Confidence intervals are chosen to reflect what is believed to be the correct number of standard deviations on each side of the point estimate, as follows:

$$(2)\quad U1 = F1 * \left(\frac{U6}{F6}\right)^{\frac{1}{\sqrt{6}}}$$

$$(3)\quad L1 = F1 * \left(\frac{L6}{F6}\right)^{\frac{1}{\sqrt{6}}}$$

Respondents typically are given several weeks to make their forecasts, with first solicitation occurring usually near the end of the preceding month. For example, for the September 2004 survey the first received response was on August 28, and the last on September 14. For these reasons, equations (1)-(3) require adjustment. Since they are not told to do otherwise, logically respondents would be making their forecasts for *exactly* six months in the future. If we use these equations without adjustment, respondents' imputed one-month forecasts (and intervals) would be for different DAX dates and thus would not be comparable. The way to obviate this problem is to use a respondent-specific imputation that doesn't generate a one-month ahead forecast (and interval) but rather yields a one-month-ahead-of-the-end-of-forecast-month forecast (and interval), as follows:

$$(1a)\quad F1a = (\frac{F6}{DAX0})^{(30+d)/180} * DAX0$$

$$(2a)\quad U1a = F1a * \left(\frac{U6}{F6}\right)^{\sqrt{(30+d)/180}}$$

$$(3a)\quad L1a = F1a * \left(\frac{L6}{F6}\right)^{\sqrt{(30+d)/180}}$$

where *d* is the number of days from forecast receipt to the end of the forecast month. Averaging subsets of *these* imputed forecasts provides the ZEW consensus forecasts that are investigated here.

3. Characteristics of successful forecasters

In this section we explore the characteristics of successful forecasters, where forecast success is calculated using MSPE. Certain of the variables considered are logical *ex ante* markers of superior performance, while others are potentially linked to overconfidence. Beginning with the former, as described in section 2, forecasts are made at different times. Those made later, when more information is likely to be available, would be expected to be better forecasts. Cross-sectionally, individuals tend to have different survey response habits, with some tending to forecast early and others doing so towards the end of the survey month. STALENESS_MEAN (i.e., the average number of days prior to the end of the survey month the forecaster in question submits her forecast) captures this. The expectation is that those contributing early and thus having higher STALENESS_MEAN will tend to have higher MSPE.

Second, forecasters submit not only point forecasts (which are used to assess MSPE) but also 90% confidence intervals surrounding their point forecasts. Logically those who feel they have a better sense of where the DAX is going should submit narrower confidence intervals. Thus average (scaled) confidence interval width (CONF_INT_MEAN), defined as (U6-L6)/DAX0, provides information on confidence. Importantly, this is not the same as overconfidence, which requires a comparison of perceived and revealed ability. The expectation is that those with lower CONF_INT_MEAN will tend to have lower MSPE. Of course it is possible that their confidence is entirely unfounded, in which case there will be no impact.

Third, frequent submission is likely to be a signal of attention. On the other hand, consistent with the inattention model of Peng and Xiong (2006), those participating sporadically are signaling inattention and perhaps a reduced ability to see where markets are moving. We define EXPERIENCE as the overall number of forecasts submitted during the sample, with the expectation that higher EXPERIENCE is

associated with lower MSPE. Diminishing returns seem likely: logically going from 10 forecasts to 20 is a stronger incremental signal of interest than going from 50 to 60 since everyone responding 50 times or more is exhibiting commitment. For these reasons we perform not only regressions with EXPERIENCE but also those including a squared term (EXPERIENCE _2), with the expectation that the coefficient on the latter should be positive to reflect convexity vs. MSPE.

The tendency to produce extreme forecasts thereby relying to a great extent on one's own intuition points in the direction of overconfidence. Consistent with Denrell and Fang (2010), the expectation is that those whose forecasts tend to be most variable will be weaker forecasters. Such a relationship is far from obvious, since, given the volatility that exists in stock indexes, a "perfect foresight" forecaster will have extremely variable forecasts. In our regressions the explanatory variable is the natural log of the standard deviation of point forecasts (SD_LOG), but our results are robust to other specifications. It is expected that SD_LOG and MSPE are positively related. Table 1 reviews our expectations.

Table 2 reveals whether the data conform to expectations. Its four panels differ in the minimum number of forecasts that a forecaster must submit in order to remain in the sample, with minima ranging from n=5 to n=30. While each panel displays three regressions, initially we focus on the first two, with the first positing a linear relationship for EXPERIENCE, and the second by including a squared term allowing for diminishing returns.[7] Turning to regression (2) in Panel B (where forecasters are only included if they have made at least 10 forecasts over the full sample and non-linearity in EXPERIENCE is allowed for), we see the coefficients line up exactly as anticipated, with all variables being of the anticipated sign and statistically significant at 1% or very close

[7] We estimate a weighted least squares model, where the weights are the number of observations that are used to calculate the mean squared prediction errors (MSPEs), so that larger weights designate more accurately measured observations. White's robust *t*-statistics are presented in parentheses.

to it. Regression (1) from the same panel is comparable, with a reduced significance level for EXPERIENCE because linearity is imposed.

The other panels can be thought of as robustness checks. STALENESS_MEAN, CONF_INT_MEAN, and the overconfidence marker SD_LOG are extremely robust, with all other coefficients indicating significance in the anticipated direction at 10% or better. As for EXPERIENCE, both the unsquared and squared terms become insignificant for n=30, which should perhaps not be surprising because given non-linearity most of the meaningful impact of EXPERIENCE comes for more moderate experience levels.

As a further robustness check, we re-estimate regression (2) by replacing CONF_INT_MEAN with average relative imputed individual volatility, or RELATIVE_IMPUTED_IND_VOL_MEAN. The latter variable begins with IMPUTED_IND_VOL, namely the conversion of respondents' confidence intervals into individual volatility estimates by using the Davidson and Cooper (1976) method to recover respondent-specific probability distributions under normality:[8]

$$(4) \quad \text{IMPUTED\_IND\_VOL} = \frac{(U1a - L1a)}{3.2 * DAX0}$$

This variable is calculated for each forecaster in every survey month. We then standardize relative to all forecasters participating in the same survey month. Finally, we calculate for all forecasters the average across all months for which there was participation. Regression (3) appears in the third column. Consistent with regression (2), survey respondents with higher average relative imputed individual volatilities have higher MSPEs.

[8] See Pearson and Tukey (1965), Moder and Rodgers (1968), and Ben-David, Graham, and Harvey (2013). Equation (4) is based on the fact that respondents' confident intervals are 90%.

The miscalibration-based variant of overconfidence, which exists when x% confidence intervals (subject to sampling error) contain fewer than x% correct answers, can be directly calculated from the data. Using the first two years of the ZEW forecasts, Deaves, Lüders and Schröder (2010) found that the average forecaster in this dataset was egregiously overconfident in this sense, but, consistent with learning, they adjusted their confidence interval widths depending on past success. Here we take a different perspective. If overconfidence gets in the way of judicious forecasting, then we would expect more overconfident forecasters to have higher MSPEs. Letting HIT_PERCENTAGE be defined as the percentage of the time one's (imputed) one-month confidence interval contains the eventual value of the DAX, with lower values indicating higher overconfidence, according to this argument HIT_PERCENTAGE should be negatively related to MSPE.

While on the surface it might appear viable to introduce HIT_PERCENTAGE as an additional explanatory variable in the MSPE regressions, there is a problem in doing so. Once we control for the average confidence width (CONF_INT_MEAN), HIT_PERCENTAGE will *by construction* be negatively related to MSPE. This is because holding constant interval width a successful forecaster will almost certainly have more "hits" than an unsuccessful one. Matters are quite different however if we relate HIT_PERCENTAGE to MSPE *without* controlling for CONF_INT_MEAN. It is helpful to roughly partition overconfidence as follows:

$$(5)\quad OVERCONFIDENCE = KNOWLEDGE\ PERCEPTION - ACTUAL\ KNOWLEDGE$$

Overconfidence exists when one's perception of knowledge (i.e., one's confidence) exceeds one's actual knowledge. More precisely, an increase in *KNOWLEDGE PERCEPTION* (in the present context, confidence interval shrinkage) reflects *ceteris paribus* higher overconfidence, while an increase in *ACTUAL KNOWLEDGE* (in the present context, lower MSPE) reflects *ceteris paribus* lower overconfidence. Since the

regression results show that confidence interval width and MSPE are positively related (i.e., low-MSPE forecasters not only have high levels of knowledge but also high perceptions of their knowledge), the relationship between overconfidence (HIT_PERCENTAGE) and MSPE is an open question. We conjecture a negative relationship between overconfidence and forecast performance, which is logical if the tendency to be overly certain of one's view induces one to economize on effort. To test this conjecture, we employ decile analysis.

For the same four cross-sectional samples as in Table 2, we form deciles based on MSPEs, with Decile 1 containing the lowest-MSPE forecasters and Decile 10 the highest. If overconfident forecasters tend to make weak forecasts, then this would imply that Decile 10 will have a lower HIT_PERCENTAGE than Decile 1. There is suggestive evidence to this effect. In all four cases, Decile 1 has a higher average HIT_PERCENTAGE than does Decile 10. Except for the case of n=30, Decile 1 has the highest (or close to the highest) average HIT_PERCENTAGE, while in *all* cases Decile 10 has the lowest HIT_PERCENTAGE. For n=20, the Decile 1 vs. Decile 10 difference is statistically significant at 10%.

4. Filtering the ZEW survey

There are compelling reasons to pool forecasts (Timmerman (2009)). For example, if different forecasts use non-matching sources of information, efficient information aggregation may result. And diverse forecasting techniques may be affected differently by structural breaks. While in theory weighting individual forecasts is appealing, a simple equal-weighted approach often dominates because of parameter estimation error. Moreover, more subtle techniques such as least squares estimation of weights are difficult to operationalize with an unbalanced panel such as the one studied here (Capistran and Timmerman (2009b)). Trimming or filtering out poor forecasters (or

models) who mostly contribute noise has been shown to improve forecast combinations (e.g., Aiolfi and Favero (2005)).[9]

Here we consider the mean ZEW DAX forecast either with or without filtering based on prior performance.[10] The purpose is to investigate whether elimination of some of the weaker forecasters improves forecast combination accuracy. In order to generate out-of-sample forecasts it is important that filtering be based on known information. Specifically we eliminate the x% of forecasters whose *prior* MSPEs fall in the bottom x% of all forecasters participating in a given month. We consider increments of 10% (10-90%) along with 95%, 99% and "All but best." The latter means that only the forecaster with the lowest prior MSPE is kept.[11]

When utilizing past information, the two choices are a recursive or rolling window.[12] In the former case, all previous data are conditioned on while in the latter a constant-length window is maintained. The advantage of the former is that all information is used, but the disadvantage is some of this information might be so stale that it is best ignored. For example, suppose there are two ways to forecast the DAX, one primarily technical and the other primarily fundamental, with some forecasters employing the first approach and others the second.[13] Further suppose that the return generating function for the DAX is regime-dependent. Under the first regime, a technical approach would generate better forecasts, while under the second regime a fundamental approach would outperform. The problem with using a recursive approach is that it is less sensitive to the current regime since it could well be the case that a forecaster looks good because her technique performed well early in the sample when one regime was

[9] Though unexplored here, further improvement may also arise by combining survey data with time series models (Pesaran and Weale (2006)).
[10] All results presented here are little affected by using the median instead of the mean.
[11] For the 99% filter, typically two forecasters remain, though with ties the number can reach seven.
[12] Note that we say "window" we mean the number of forecasts that we look back at to assess performance *prior* to the forecast in question. Thus this forecast is *not* included in the window.
[13] Dick and Menkhoff (2012) use this categorization in investigating ZEW exchange rate forecasts.

in place but her recent performance has been weaker now that a second regime is in effect. By varying the length of the rolling window one can get a sense of the optimal amount of past data to condition on. In truth, however, such a comparison is going to have an in-sample flavor, as there is no guarantee that this optimal window length will continue to be optimal going forward.

Figure 2 displays both OS-R^2s and corresponding p-values for one-, two- and three-year recursive windows. To clarify, in the (say) two-year case, for possible inclusion in the consensus respondents are ranked based on MSPE over the first 24 surveys and if they are in the lowest x% they remain in the sample for the 25th survey. Moving forward one period, to form the 26th survey consensus, the holdout sample is based on the first 25 forecasts, and so on. Note that to be considered for inclusion we impose the screen that at least 10 forecasts must have been made by a forecaster during the holdout window (i.e., prior to the forecast to be evaluated). It can be observed that while filtering improves matters somewhat the OS-R^2 is never significant even at 10%.[14] Evidently, there is little obvious value added in using a recursive approach.[15]

In Figure 3 the same one-, two- and three-year windows as in Figure 2 are utilized, this time though using a rolling methodology. Again, we employ the screen that at least 10 forecasts over the rolling window must have been made. The first evaluated forecast is done in an identical fashion to the recursive approach, but moving forward the window size is kept constant, implying that early observations are ignored in forecast evaluation. Again, in all cases at least 10 observations over the preceding one, two or three years are required in order to be considered for inclusion. A rolling one-year approach reveals some improvement vs. no filtering with OS-R^2s for 30-50% filters ranging from 2.66-3.38% with p-values at 10% or better. The superiority of a one-year

[14] As it were, there are two filters. The first, which to avoid confusion we call a screen, requires a sufficiently long track record so that past performance can be assessed, and the second drops people based on poor past performance.

[15] Note that even the 0% filter is based on the "minimum of 10" restriction.

vs. two- and three-year windows suggests that it is best to limit the window length so that forecasting success in the more distant past is ignored.

Figure 4 investigates how narrow the window should be in order to maximize combination forecast improvement. Four approaches are displayed. The first (Min_10_for_12) repeats the rolling one-year window used in Figure 3 as a point of departure. The other three filters employ rolling windows of six months (Min_5_for_6), three months (Min_2_for_3) and one month (Min_1_for_1). It is also necessary to specify a minimum number of prior forecasts in the rolling window (again noting that the window does not include the forecast under consideration). For six months/three months/one month, the minimum is five/two/one. To interpret the Min_1_for_1 case, included forecasters must participate in two consecutive surveys, the one whose success is being examined as well as the one immediately preceding (where past success is based on how close the latter forecast was to the eventual DAX). Related to Figure 4 is Figure 5. Figure 5 utilizes the same four approaches, but now the unfiltered mean forecast is the benchmark against which we compare filtered mean forecasts (which is why we begin at 10%).

Beginning with Min_1_for_1, the highest OS-R^2 observed in Figure 4 (6.75%, p-value=0.063) is *without* filtering. Thus, exclusion of forecasters is not helpful: in fact it worsens matters, and for filters of 70% or more it is very much counterproductive. This is should not be surprising since a track record of a single previous forecast (beyond the one under examination) is naturally rife with noise, and is clearly subject to the Denrell and Fang (2010) extreme-forecast success critique. Nevertheless it should be noted that there is a marginal gain from attention due to the fact that only those forecasters participating twice in a row are considered. The reference point in this regard is an OS-R^2 of 6.19% (p-value = 0.073), which applies to the case when we only assess the mean forecast without any past history requirement.

As for the other two (new) cases in Figure 4, filtering improves matters for both the rather short 6-month and 3-month rolling windows. For example, for the very narrow three-month window (where we insist that a forecaster was active for the majority (i.e., 2 of 3) of prior forecasts), the OS-R^2s range from 3.35-4.18% for 10-50% filters. These values are statistically significant at the 5% level when compared to the historical mean. To ascertain the success of filtering, refer to Figure 5. Broadly speaking, filtering out inferior forecasters is somewhat helpful, with a moderate amount of filtering producing the best results. Again, for the Min_2_for_3 case, the OS-R^2 (vs. no filtering) at a 10% filter is 1.45% with a p-value of 0.090.[16]

5. Discussion and concluding remarks

The ability to forecast market returns is critical for many decision-makers. It matters for market timing, asset allocation, pension fund deficit calculation and corporate planning. While it is recognized that returns have at best a modest predictable component, any improvements that can be garnered over such naïve models as the short rate plus the average realized equity premium are without doubt worth pursuing. Panels of expert forecasters are a ready source of informed opinion, but it is not clear how to make the best use of panel data.

We have considered how overconfidence impacts forecast performance. Overconfidence as proxied by the tendency to make extreme forecasts leads to poor performance. Further, controlling for the fact that good forecasters have some knowledge of their skill which causes them to generate more narrow confidence intervals, it is still true that overconfidence as proxied by the hit ratio (i.e., percentage of

[16] For the Min_5_for_6 case, the OS-R^2 (vs. no filtering) at a 20% filter is 2.11% with a p-value of 0.087. For the Min_10_for_12 case, the OS-R^2 (vs. no filtering) at a 10% filter is 1.50% with a p-value of 0.078. For brevity, we do not provide the "vs. 0% filter" analogous (to Figures 2 and 3) charts. In a nutshell 10% filtering is effective (at 10% or close to it) for the three recursive approaches. On the other hand, filtering does not pay off for the 24-month and 36-month rolling windows.

the time that an interval contains the eventual realization) is associated with poor performance. It is beneficial to have information on the sources of forecast weakness because if one has such information but the forecaster under the microscope has an insufficient track record one can still make educated guesses about future performance.

Given forecaster heterogeneity it is logical to explore whether filtering out weak forecasters based on prior MSPE is a viable strategy. Short rolling windows, which delicately balance ignoring relevant information and noise reduction, work best. This suggests that structural change renders early forecasts uninformative. The trick is to find the filter that equates the marginal benefit of the first to the marginal cost of the second. The ZEW data indicate that this is achieved at moderate (10-20%) filters.

REFERENCES

Aiolfi, M., and C. A. Favero, 2005, Model uncertainty, think modelling and the pedictability of stock returns, *Journal of Forecasting* 24, 233-54.

Barber, B., and T. Odean, 2000, Trading is hazardous to your wealth: The common stock investment performance of individual investors, *Journal of Finance* 55, 773-806.

Ben-David, I., J. R. Graham, and C. R. Harvey, 2013, Managerial miscalibration, *Quarterly Journal of Economics*, Forthcoming.

Camerer, C. F., and D. Lovallo, 1999, Overconfidence and excess entry: An experimental approach, *American Economic Review* 89, 306-18.

Capistran, C., and A. Timmerman, 2009a, Disagreement and biases in inflation expectations, *Journal of Money, Credit and Banking* 41, 365-96.

Capistran, C., and A. Timmerman, 2009b, Forecast combination with entry and exit of experts, *Journal of Business and Economic Statistics* 27, 428-40.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam, 1998, Investor psychology and security market under- and overreactions, *Journal of Finance* 53, 1839-85.

Davidson, L. B., and D. O. Cooper, 1976, A simple way of developing a probability distribution of present value, *Journal of Petroleum Technology*, September, 1069-1078.

Deaves, R., E. Lüders, and M. Schröder, 2010, The dynamics of overconfidence: Evidence from stock market forecasters, *Journal of Economic Behavior and Organization* 75, 402-12.

Denrell, J., and C. Fang, 2010, Predicting the next best thing: Success as a signal of poor judgment, *Management Science* 56, 1653-67.

Dick, C. D., and L. Menkhoff, 2012, Exchange rate expectations of chartists and fundamentalists, Working paper.

Ferreira, M. A., and P. Santa-Clara, 2011, Forecasting stock market returns: The sum of the parts is more than the whole, *Journal of Financial Economics* 100, 514-37.

Goetzmann, W. N., and A. Kumar, 2008, Equity portfolio diversification, *Review of Finance* 12, 433-63.

Goyal, A., and I. Welch, 2008, A comprehensive look at the empirical performance of equity premium prediction, *Review of Financial Studies* 21, 1455-508.

Hirshleifer, D., and S. H. Teoh, 2003, Herd behavior and cascading in capital markets: A review and synthesis, *European Financial Management* 9, 25-66.

Kahneman, D., 2011. *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York).

Malmendier, U., and G. Tate, 2008, Who makes acquisitions? CEO overconfidence and the market's reaction, *Journal of Financial Economics*.

Moder, J. J. and Rodgers, E. G., 1968, Judgment estimates of the moments of PERT Type distributions, *Management Science* 15, B76-B83.

Neely, C. J., D. E. Rapach, J. Tu, and G. Zhou, 2010, Out-of-sample equity premium prediction: Economic fundamentals vs. moving-average rules, Working paper.

Pearson, E. S. and Tukey, J. W., 1965, Approximate means and standard deviations based on distances between percentage points of frequency curves, *Biometrik*a 52, 533-546.

Peng, L., and W. Xiong, 2006, Investor inattention, overconfidence and category learning, *Journal of Financial Economics* 80, 563-602.

Pesaran, M. H., and M. Weale, 2006, Survey expectations, in G. Elliott, C. W. J. Granger, and A. Timmerman, eds.: *Handbook of Economic Forecasting, Volume 1* (Elsevier B. V.).

Rapach, D. E., J. K. Strauss, and G. Zhou, 2010, Out-of-sample equity prediction: Combination forecasts and links to the real economy, *Review of Financial Studies* 23, 821-62.

Timmerman, A., 2006, Forecast combinations, in G. Elliott, C. W. J. Granger, and A. Timmerman, eds.: *Handbook of Economic Forecasting, Volume 1* (Elsevier B. V.).

TABLE 1: Sign expectations of determinants of MSPE

Independent Variables	Expected sign
STALENESS _MEAN	+
CONF_INT_MEAN	+
EXPERIENCE	-
EXPERIENCE_2	+
SD_LOG	+

TABLE 2: Cross-sectional MSPE regressions

Panel A: At least 5 survey responses			
Independent Variables	(1)	(2)	(3)
STALENESS_MEAN	0.00007454***	0.00008235***	0.00007572***
	(4.245)	(4.626)	(4.475)
CONF_INT_MEAN	0.00310119***	0.00288965***	
	(2.968)	(2.915)	
SD_LOG	0.00104774**	0.00131581***	0.00129062***
	(2.182)	(2.807)	(2.777)
EXPERIENCE	-0.00001031***	-0.00007368***	-0.00007658***
	(-2.595)	(-4.074)	(-4.188)
EXPERIENCE_2		0.00000062***	0.00000064***
		(3.914)	(4.032)
RELATIVE_IMPUTED_IND_VOL_MEAN			0.00027703**
			(2.388)
Constant	-0.00525175*	-0.00596834*	-0.00508886
	(-1.651)	(-1.920)	(-1.643)
Observations	381	381	381
Adj. R-squared	0.077	0.110	0.112

Panel B: At least 10 survey responses			
Independent Variables	(1)	(2)	(3)
STALENESS_MEAN	0.00006214***	0.00006898***	0.00006383***
	(4.200)	(4.708)	(4.426)
CONF_INT_MEAN	0.00247965***	0.00229919**	
	(2.653)	(2.572)	
SD_LOG	0.00132709***	0.00161398***	0.00157628***
	(3.511)	(3.968)	(3.888)
EXPERIENCE	-0.00000922**	-0.00007053***	-0.00007314***
	(-2.168)	(-3.662)	(-3.773)
EXPERIENCE_2		0.00000058***	0.00000060***
		(3.662)	(3.783)
RELATIVE_IMPUTED_IND_VOL_MEAN			0.00020601**
			(2.233)
Constant	-0.00696927***	-0.00776490***	-0.00693472***
	(-2.698)	(-2.971)	(-2.729)
Observations	347	347	347
Adj. R-squared	0.086	0.119	0.118

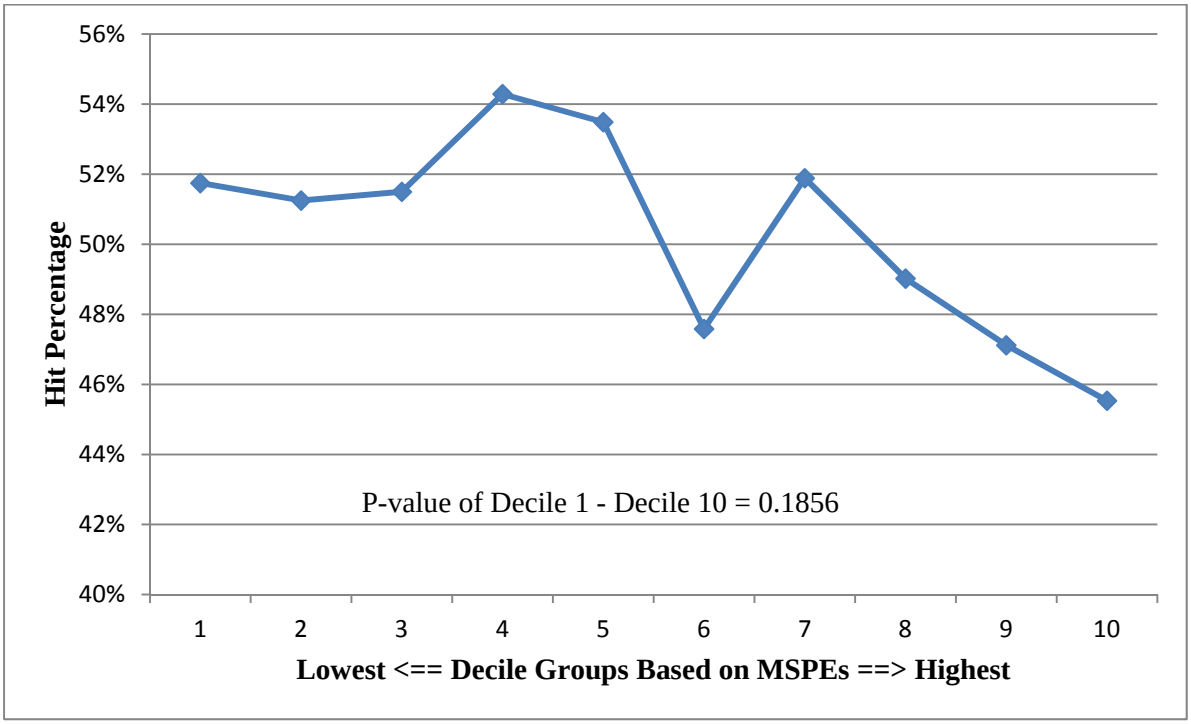
Panel C: At least 20 survey responses			
Independent Variables	(1)	(2)	(3)
STALENESS_MEAN	0.00006164***	0.00007195***	0.00006810***
	(4.096)	(4.699)	(4.507)
CONF_INT_MEAN	0.00182910**	0.00172883**	
	(2.062)	(1.994)	
SD_LOG	0.00157821***	0.00179645***	0.00179793***
	(3.395)	(3.676)	(3.683)
EXPERIENCE	-0.00000825*	-0.00009699***	-0.00009865***
	(-1.709)	(-3.225)	(-3.261)
EXPERIENCE_2		0.00000079***	0.00000080***
		(3.296)	(3.336)
RELATIVE_IMPUTED_IND_VOL_MEAN			0.00015628*
			(1.864)
Constant	-0.00872970***	-0.00825901***	-0.00785522**
	(-2.737)	(-2.606)	(-2.522)
Observations	296	296	296
Adj. R-squared	0.086	0.130	0.131

Panel D: At least 30 survey responses			
Independent Variables	(1)	(2)	(3)
STALENESS_MEAN	0.00006109***	0.00006448***	0.00006066***
	(4.195)	(4.317)	(4.133)
CONF_INT_MEAN	0.00174215*	0.00165906*	
	(1.965)	(1.916)	
SD_LOG	0.00158171***	0.00168869***	0.00169950***
	(3.358)	(3.395)	(3.425)
EXPERIENCE	0.00000136	-0.00003945	-0.00003977
	(0.292)	(-0.972)	(-0.983)
EXPERIENCE_2		0.00000034	0.00000034
		(1.085)	(1.099)
RELATIVE_IMPUTED_IND_VOL_MEAN			0.00016419*
			(1.963)
Constant	-0.00940600***	-0.00908627***	-0.00880228***
	(-2.909)	(-2.811)	(-2.773)
Observations	264	264	264
Adj. R-squared	0.114	0.117	0.121

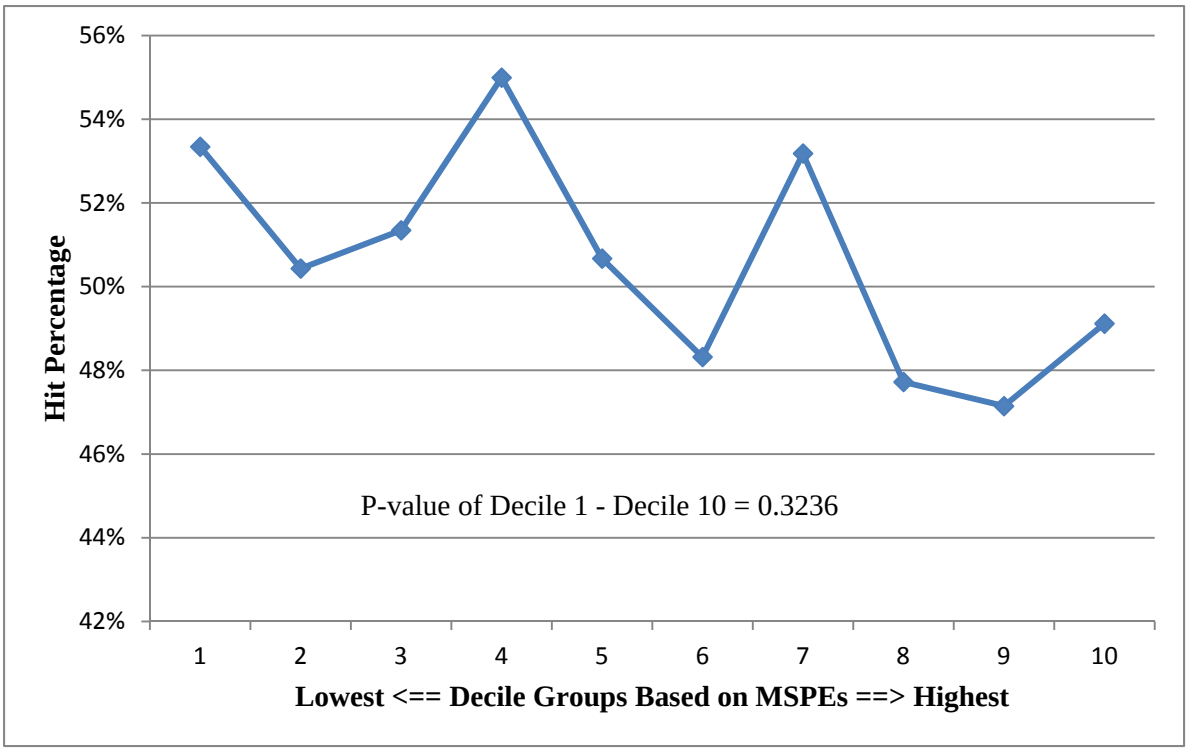
Notes: We estimate a weighted least squares model, where the weights are the number of observations that are used to calculate the mean squared prediction errors (MSPEs), so that larger weights designate more accurately measured observations. Robust *t*-statistics are presented in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10.

FIGURE 1: Hit_Percentages for MSPE deciles

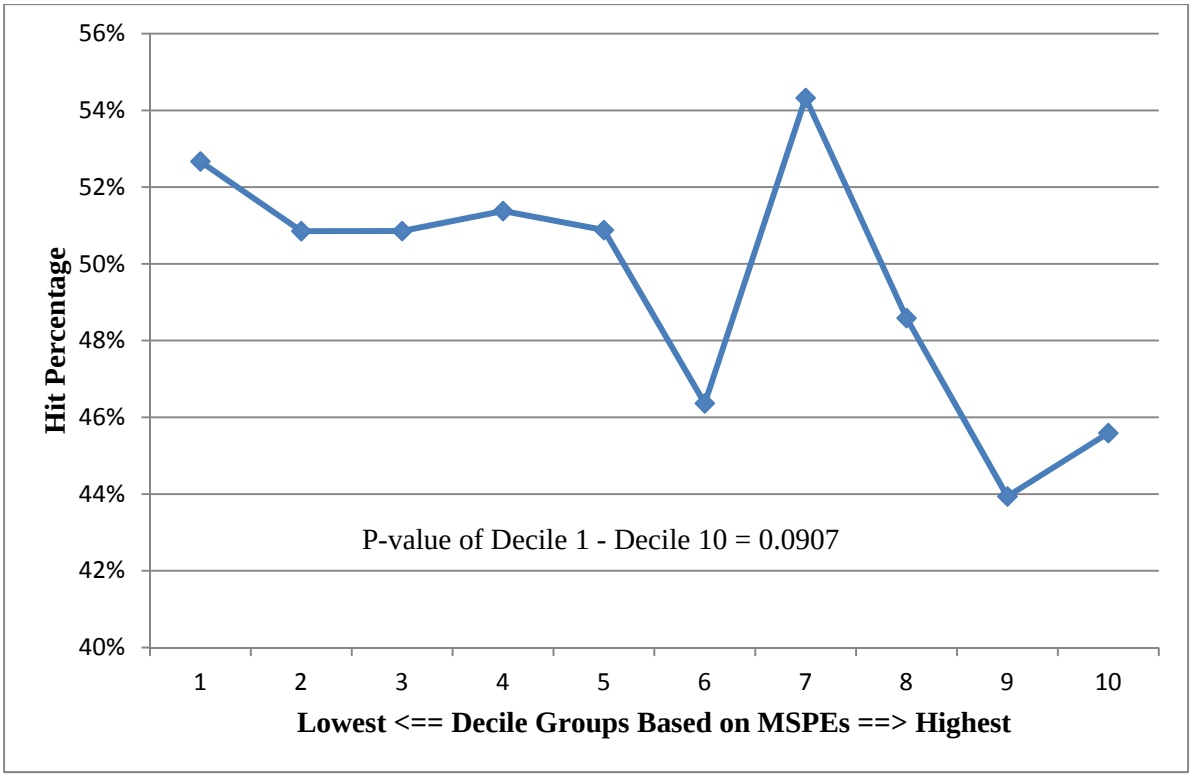
Panel A: At least 5 survey responses



Panel B: At least 10 survey responses



Panel C: At least 20 survey responses



Panel D: At least 30 survey responses

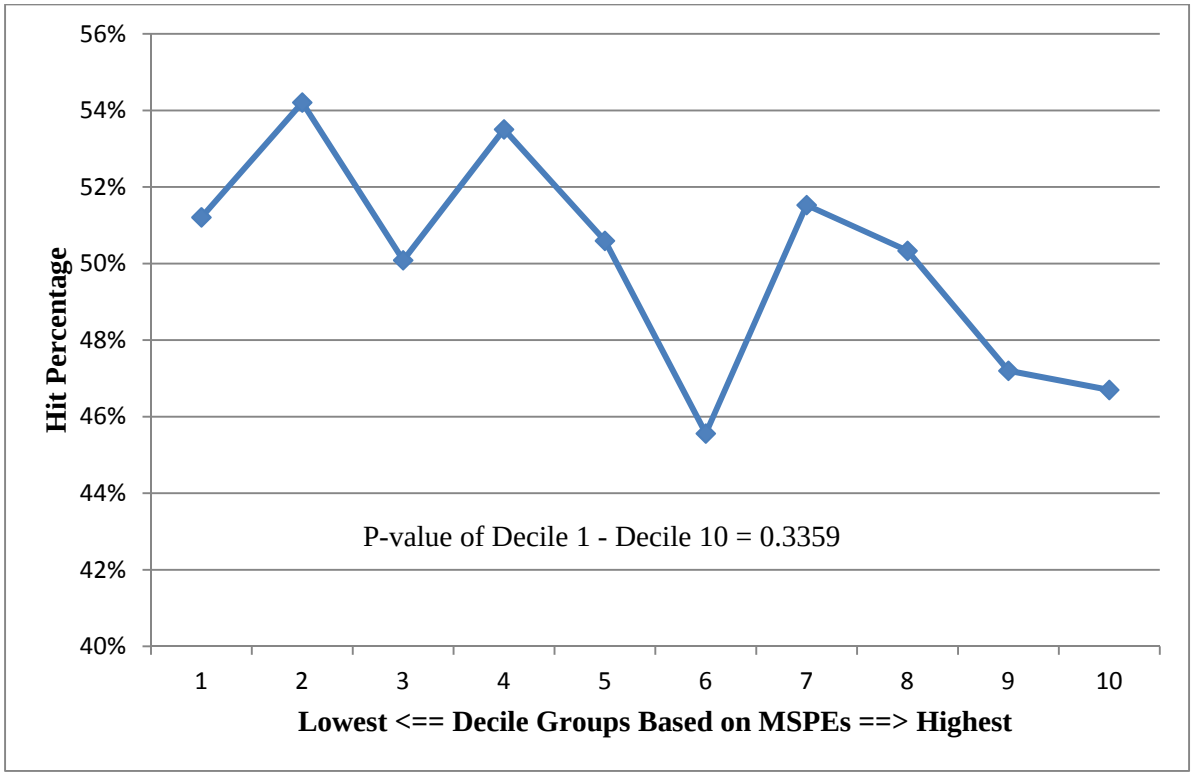


FIGURE 2: OS-R^2s and p-values for 1-year to 3-year recursive screens

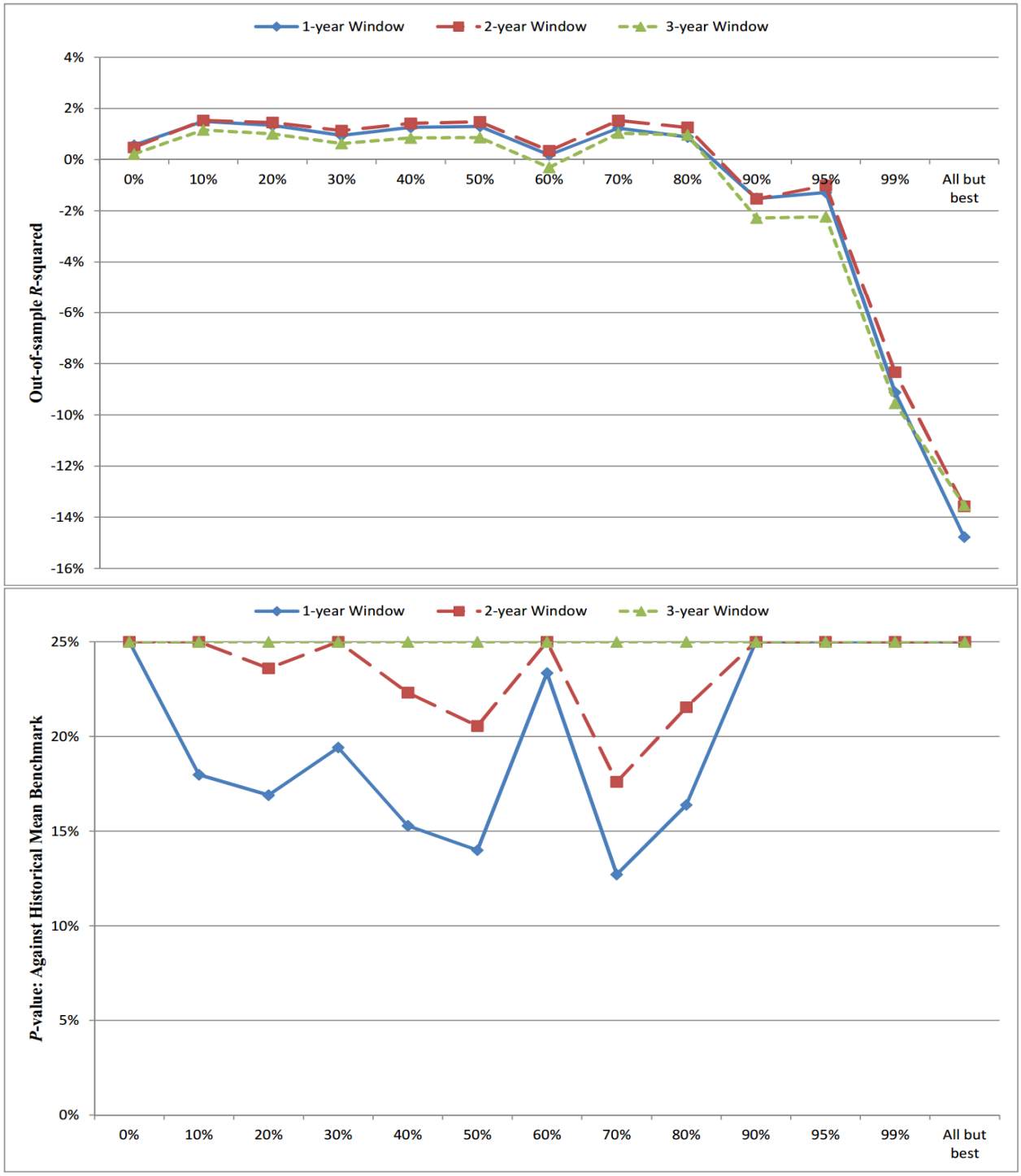


FIGURE 3: OS-R^2s and p-values for 1-year to 3-year rolling screens

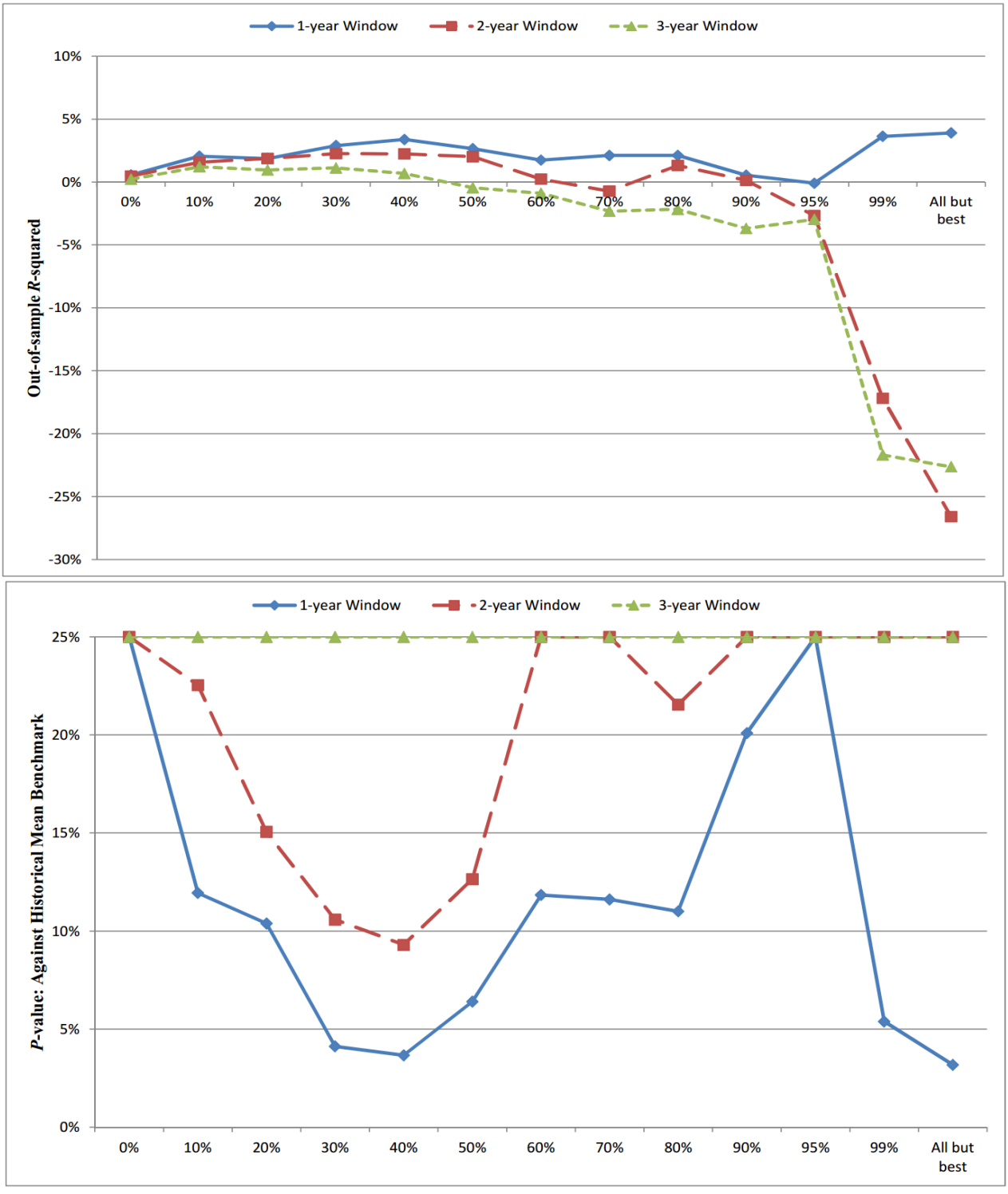


FIGURE 4: OS-R^2s and p-values for short rolling screens

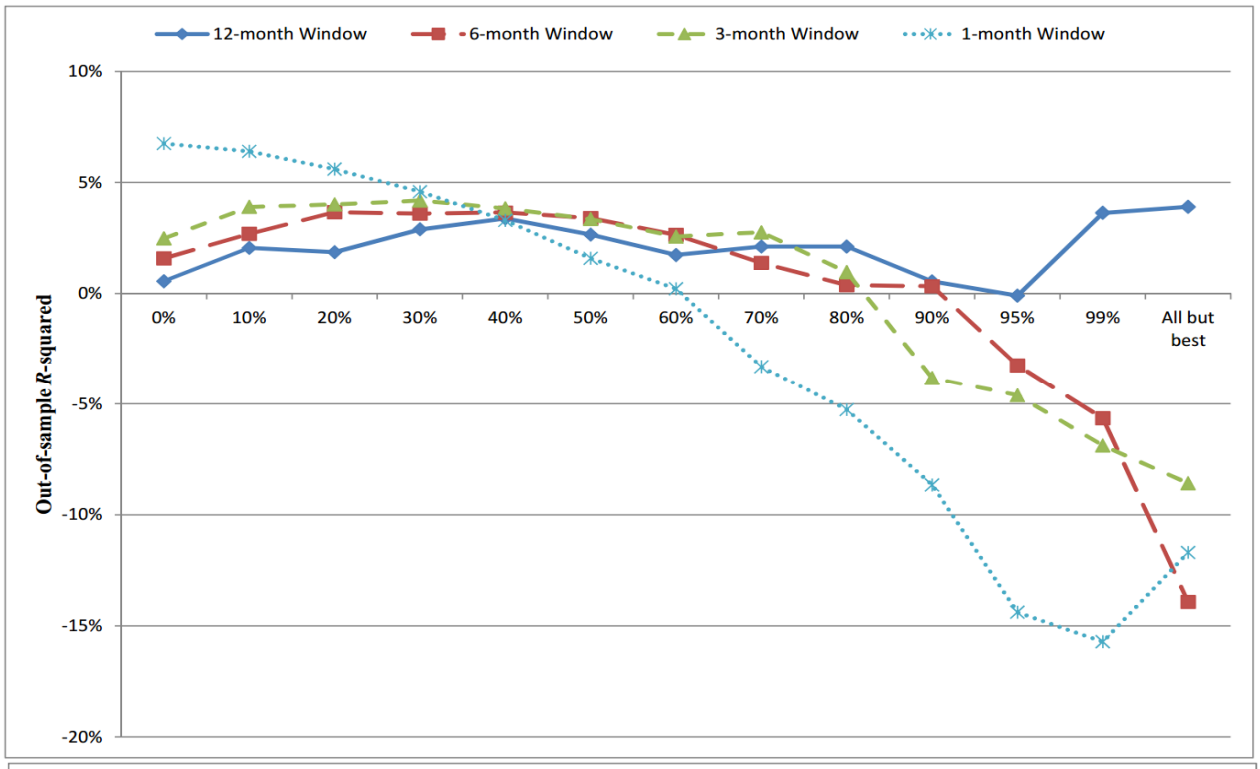


FIGURE 5: OS-R^2s and p-values for short rolling screens (Against 0% Filter Benchmark)