Breitung, Jörg; Schmeling, Maik

Working Paper

Quantifying survey expectations: What's wrong with the probability approach?

Diskussionsbeitrag, No. 485

Provided in Cooperation with:
School of Economics and Management, University of Hannover

Suggested Citation: Breitung, Jörg; Schmeling, Maik (2011) : Quantifying survey expectations: What's wrong with the probability approach?, Diskussionsbeitrag, No. 485, Leibniz Universität Hannover, Wirtschaftswissenschaftliche Fakultät, Hannover

Quantifying survey expectations: What's wrong with the probability approach?*

Jörg Breitung‡ **Maik Schmeling****

December 2, 2011

*We would like to thank the Center for European Economic Research (ZEW) in Mannheim for providing the data used in this study and Roy Batchelor, Andreas Schrimpf and Jan-Egbert Sturm for helpful comments. Schmeling gratefully acknowledges financial support by the German Science Foundation (DFG).

‡Department of Economics, University of Bonn, Adenauerallee 24-42, D-53113 Bonn, Germany, Phone: +49 (0) 228 73 9201, Fax: +49 (0) 228 73 9189, e-mail: breitung@uni-bonn.de.

**Corresponding author. Department of Economics, Leibniz Universität Hannover, Königsworther Platz 1, D-30167 Hannover, Germany, Phone: +49 (0) 511 7238213, Fax: +49 (0) 511 7234796 , e-mail: schmeling@gif.uni-hannover.de.

Quantifying survey expectations: What's wrong with the probability approach?

Abstract

We study a matched sample of individual stock market forecasts consisting of both qualitative and quantitative forecasts. This allows us to test for the quality of forecast quantification methods by comparing quantified qualitative forecasts with actual quantitative forecasts. Focusing mainly on the widely used quantification framework advocated by Carlson and Parkin (1975), the so-called "probability approach", we find that quantified expectations derived from the probability approach display a surprisingly weak correlation with reported quantitative stock return forecasts. We trace the reason for this low correlation to the importance of asymmetric and time-varying thresholds, whereas individual heterogeneity across forecasters seems to play a minor role. Hence, our results suggest that qualitative survey data may not be a very useful device to obtain quantitative forecasts and we suggest ways to remedy this problem when designing qualitative surveys.

JEL-Classification: C53, D84, G17
Keywords: Quantification, Stock Market Expectations, Probability Approach

1 Introduction

The concept of (rational) expectations is an essential element of modern macroeconomics and finance. Consequently, a large amount of research has been dedicated to investigating whether expectations are formed rationally or not (e.g. Elliott et al., 2005), how well expectations actually forecast future economic activity and financial market outcomes (e.g. Nolte and Pohlmeier, 2007), whether expectations are stabilizing or not (Frankel and Froot, 1990), why forecasters form divergent beliefs (Patton and Timmermann, 2010) or how these belief dispersions impact financial markets (e.g. Beber et al, 2010, Anderson et al., 2009).

However, actually *measuring* expectations is a difficult task and there are many proposals in the literature on how to deal with this problem. One particular way of measuring expectations is the use of survey data where individuals are directly asked for their expectations regarding some financial or macroeconomic variable. Often, these surveys are conducted in a qualitative way, i.e. respondents do not issue precise point forecasts but rather vote on an ordinal scale such as "up", "unchanged", or "down" (see e.g. Pesaran and Weale, 2006, for an overview of survey designs). In this situation, a common procedure is to apply some quantification algorithm to convert qualitative survey expectations into quantitative forecasts. While many shortcomings of available quantification methods have been discussed (see Nardo, 2003, for an overview), it is still common to apply variants of these quantification techniques when confronted with qualitative data (e.g. Doepke et al., 2008, Mankiw et al., 2003, Menkhoff et al., 2009).

In this paper, we first investigate how well standard quantification methods achieve their goal to deliver a reliable proxy for actual, quantitative expectations. We mainly focus on the workhorse of quantification methods, i.e. the so-called "probability approach" as advocated by Carlson and Parkin (1975, henceforth CP) which is still used widely today. We find that standard quantification methods lead to a poor fit between quantified expectations and actual quantitative expectations. In a second step, we examine the reason of this poor performance and investigate the relative importance of different possible explanations. We find that (i) assumptions about the distribution of return expectations in the CP procedure do not seem to matter, (ii) that the assumption of time-constant

threshold values (which forecasters use to convert their true expectations into qualitative expectations) explains most of the poor performance of the CP procedure, and (iii) that forecaster-specific heterogeneity matters only to a limited extent. Hence, our results provide guidance on how to design a survey and on how to convert qualitative expectations into quantitative expectations which we discuss in more detail below. We are, to the best of our knowledge, the first to analyze a sample of stock return expectations that contains both qualitative and quantitative forecasts and to provide *direct* evidence on the weakness (as well as the reasons for this weakness) of popular quantification procedures in such a setting.

A special feature of our empirical study is that we have available a unique matched sample of individual stock market forecasts consisting of both qualitative *and* quantitative forecasts issued by the same forecasters (see, e.g. Defris and Williams (1979), Batchelor (1986), Maag (2010), Lui, Mitchell, and Weale, 2011a, and Lui, Mitchell, and Weale, 2011b) for matched samples dealing with inflation expectations or business surveys). This allows us to test for the quality of forecast quantification methods by comparing the outcome of a quantification procedure based on qualitative forecasts with actual quantitative forecasts. More specifically, we ask the following question: how close are quantified expectations to actual quantitative expectations? This validity check is arguably the most direct way to evaluate the CP procedure.

In contrast to earlier papers we investigate expectations of financial experts on stock prices. Analyzing these data has several interesting features compared to earlier studies which are mostly based on consumer's expectations. First, the survey participants are professional forecasters who are supposed to form unbiased and rational expectations. Rationality of forecasts is crucial to calibrate the level of expected returns in standard quantification procedures. Second, the target variable (DAX30 stock price index) is precisely defined and well known among the survey participants, whereas in consumer surveys respondent's perceived prices may deviate substantially from the official consumer price index. Finally, stock returns are quite different from inflation rates. In particular, stock returns are much more volatile and less persistence than inflation rates. Furthermore, the expected component of returns (i.e., the risk premium) is less important for future returns than expected inflation is for future inflation outcomes. Hence, results of our study shed

light on the performance of the CP procedure in a quite different forecasting context than most earlier studies and thus form an interesting out of sample test of these results.

In our empirical analysis, we first employ standard quantification methods to convert qualitative expectations into quantified expectations. Next, we show that these quantified expectations are only mildly correlated with actual quantitative expectations (roughly 45%). Furthermore, the level of quantitative and quantified expectations differs substantially over extended periods of time. To investigate the source of this low correlation, we first provide parametric and and non-parametric estimates of key model parameters of quantification procedures, namely the indifference limits (or "threshold values") which forecaster use to convert quantitative expectations into qualitative expectations. We find that these threshold values are strongly time-varying, a feature not captured by standard quantification methods, and that allowing for this time-variation in thresholds increases the correlation between quantified and quantitative expectations to levels well above 90%. Hence, taking time-variation in indifference limits into account seems to be the most important feature both for designing surveys and for the end-users of these surveys, at least when dealing with expectations of a highly volatile series such as stock returns.

Finally, we also examine the extent to which forecaster-specific heterogeneity matters for the poor performance of the CP procedure based on a novel testing procedure. We find significant evidence of individual heterogeneity among forecasters but show that most of this heterogeneity washes out at the aggregate level and, hence, does not matter much for the low correlation between quantified and quantitative expectations. To further scrutinize this issue, we conduct some stylized Monte Carlo simulations which are able to reproduce our main findings when allowing for time-varying thresholds. It should be noted, however, that we are mainly interested in the performance of the CP procedure at the aggregate level, i.e., the way it is usually employed in practice and by other researchers, but not at the individual level of forecasters. Our results reflect this focus and do not deal in detail with the importance of heterogeneity at the level of individual forecasters (see, e.g., Mitchell, Smith, and Waele, 2002, 2005 for analyses based on a panel of *individual*

forecasters).[1]

The remainder of the paper is organized as follows: We discuss related literature in the next Section, describe our data in Section 3, briefly discuss some quantification methods in Section 4, and provide estimation details for these methods in Section 5. We estimate time-varying indifference limits in Section 6, investigate the importance of forecaster-specific heterogeneity in section 7, and back up our main findings by means of Monte Carlo simulations in Section 8. Section 9 concludes.

2 Related literature

Our general conclusion that quantification procedures need to be interpreted with great care, is not new. For example, several papers found some form of parameter instability (e.g. Batchelor and Orr, 1988, Dasgupta and Lahiri, 1992), deviations from the normal distribution and asymmetries (e.g. Berk, 1999, Maag, 2010), and conclude that quantified qualitative survey expectations do not necessarily forecast the target variable better than simple time-series models (e.g. Claveria et al., 2007, Nolte and Pohlmeier, 2007, Breitung 2008), or directly suggested improvements and extensions of the CP approach to alleviate other problems of the method (e.g. Fishe and Lahiri, 1982, Löffler, 1999, Mitchell et al., 2007, or Müller, 2010). Regarding the latter task, Mitchell, Smith, and Weale (2002, 2005) show how to use panel data on qualitative forecasts to circumvent problems associated with the CP procedure which focuses on aggregate expectations.

In contrast to these papers, however, our data allow us to directly assess the reliability of quantification approaches rather than analyzing merely the forecast performance or statistical properties of quantified expectations. It is important to note that common quantification procedures do not pass this direct test (at least in our specific sample) which casts some doubt on the common practice to simply infer quantitative forecasts from qualitative survey data.

[1] More specifically, we do not assume that the researcher has available data on individual quantitative, retrospective data that can be used to adapt the CP methodology to exploit individual expectations. We rather focus on the situation where the researcher has to quantify expectations based solely on aggregate balance statistics which is likely to be more common in applied work.

Defris and Williams (1979), Batchelor (1986), and Maag (2010) also investigated matched samples of qualitative and quantitative forecasts. The first two studies found a rather poor performance of the CP procedure. While Batchelor (1986) found a modest correlaton of 0.61 between the CP measure of consumer's inflation expectations, the short term movements (measured changes in expectations) reveal no statistical significant relationship to the reported quantitative figures. More recently Maag (2010) compares the outcome of the 5-category probability method of Batchelor and Orr (1988) to reported quantitative inflation expectations from the Swedish Consumer tendency survey. Applying various empirical tests he found that "the actual response scheme is neither symmetric nor homogenous across individuals", as assumed by the (generalized) CP methodology. Furthermore, "quantitative beliefs are not normally distribued and cannot be reconciled with a noncentral t distribution either."

A growing body of work has also investigated the relationship between qualitative and quantitative data at the individual level (e.g., Lui, Mitchell, Weale, 2011a, Lui, Mitchell, Weale, 2011b) where indifference limits are directly estimated via ordered choice models. Hence, since threshold parameters are directly estimated from the data, these approaches allow for tests of the degree of threshold heterogeneity across forecasters.

3 The data set and some descriptive results

Our sample is based on a monthly survey conducted by the Centre for European Economic Research (ZEW), one of the largest economic research institutes in Germany. Each month, approximately 350 professional forecasters from large German banks, institutional investors, or treasury departments of large corporations are asked for their forecasts of a variety of macro and financial variables over the next six months. The forecasts can be assigned to the categories "up", "unchanged", "down", or no "no opinion", so participants issue qualitative forecasts. The number of participants in the survey is quite large (compared e.g. to the Survey of Professional Forecasters) and the average number of participants is 223 forecasters (excluding those that answer "no opinion").[2]

Among other things, the survey covers forecasts for the German stock market and specif-

[2]A discussion of these survey data can also be found in Pesaran and Weale (2006).

ically asks for forecasts of DAX30 returns six months ahead, which have been collected on a monthly basis since December 1991. Part of our sample data are based on this qualitative survey.

In addition to this qualitative survey, the ZEW has started to collect quantitative expectations for the DAX as well. More specifically, starting in February 2003, the same forecasters that issue qualitative forecasts were also asked to issue point forecasts for the DAX for the same forecast horizon of six months. Since we will need to look at expected returns instead of point forecasts in the following, we convert the point forecasts into return forecasts based on current DAX index level on the day the forecast is issued.

Our total sample extends from February 2003 to October 2008. Hence, the sample is not excessively long but still includes both the bull market from 2003 to 2007 as well as a large chunk of the recent market crash. Descriptive statistics for the average quantitative return expectations are presented graphically in Figure 1. Shown are the cross-sectional mean, standard deviation, skewness, and kurtosis of individual forecasters' return expectations for each of the 69 months in our sample. It can be seen that all four moments are changing substantially over time and they are far from being constant. It is also apparent that the mean and standard deviation of return expectations tend to move in the same direction. While return expectations clearly show some persistence due to the overlapping forecast horizons, it is clearly implausible to assume that returns (and expectations thereof) are nonstationary. We test for this using a number of tests (ADF, DF-GLS, Phillips-Perron tests) and find significant evidence against a unit root for each test just as expected. Finally, it is worth noting that the average of return forecasts in our sample is roughly 3.4% which translated into an annualized return expectation of about 6.9%, a figure that seems reasonable for the German stock market.

FIGURE 1 ABOUT HERE

For reasons that will become apparent below it is also interesting to investigate whether return expectations are normally distributed. Although results for the skewness and kurtosis in Figure 1 do not support the assumption of normally distributed expectations,

we also directly test for this using a standard Jarque-Bera test. The p-values of this test for each cross-section, i.e. each month, in our sample are shown in Figure 2. It is obvious from this graph that the test usually rejects normality of return expectations. In fact, the test rejects normality at the 5% level in about 82% of all months in our sample.

FIGURE 2 ABOUT HERE

4 Quantification methods

The probability framework was proposed by Anderson (1951) and Theil (1952) and refined by Carlson and Parkin (1975). Since the latter authors' approach is most common in the literature, we mainly focus on their method. The CP procedure is based on on a set of assumptions that can be summarized as follows:

Assumption CP: *(a) The expectations of respondent i at period t is independently and identically distributed as $y_{it}^e \sim \mathcal{N}(\mu_t, \sigma_t^2)$. (b) The respondents report $r_{it} = 1$ (increase) if $y_{it}^e > \delta^+$, $r_{it} = -1$ (decrease) if $y_{it}^e < \delta^-$ and $r_{it} = 0$ (no change) if $\delta^- \leq y_{it}^e \leq \delta^+$. (c) The number of respondents N_t in period t is sufficiently large such that $n_t^+/N_t \approx p_t^+$ and $n_t^-/N_t \approx p_t^-$, where n_t^+ (n_t^-) denotes the number of respondents in t reporting an increase (decrease), $p_t^+ = Prob(y_{it}^e > \delta^+)$ and $p_t^- = Prob(y_{it}^e < \delta^-)$.*

Let $\Phi(z)$ denote the c.d.f. of the standard normal distribution and define the inverse normal scores as

$$q_t^+ = \Phi^{-1}(1 - p_t^+) = (\delta^+ - \mu_t)/\sigma_t \quad (1)$$

$$q_t^- = \Phi^{-1}(p_t^-) = (\delta^- - \mu_t)/\sigma_t \ . \quad (2)$$

Solving for μ_t yields

$$\mu_t = \delta^+ z_{1t} - \delta^- z_{2t} \quad (3)$$

where $z_{1t} = q_t^-/(q_t^- - q_t^+)$ and $z_{2t} = q_t^+/(q_t^- - q_t^+)$. Assuming rational expectations of the survey respondents we have $\mu_t = E(y_t|I_t)$, where I_t is the information set available at period t.

The original CP approach employs a constant information set $I_t = \{1\}$ and symmetric indifference limits $\delta^+ = -\delta^- \equiv \delta$. This gives rise to the instrumental variable estimator of the indifference limit

$$\widehat{\delta}_{cp} = \frac{\sum_{t=1}^{T} y_t}{\sum_{t=1}^{T} z_t}, \tag{4}$$

where $z_t = z_{1t} + z_{2t}$.

Various modifications of this setup have been proposed in the literature (see Nardo, 2003, for a survey of the literature) and we briefly discuss some of these modified procedures below. One extension of the original framework is to include z_t in the information set (cf. Bachelor, 1982 and Breitung, 2008) yielding the least-squares (LS) estimator

$$\widehat{\delta}_{ls} = \frac{\sum_{t=1}^{T} y_t z_t}{\sum_{t=1}^{T} z_t^2} \tag{5}$$

As a third variant of the probability approach, the regression estimator suggested by Berk (1999) includes z_{1t} and z_{2t} separately in the information set I_t. This allows estimating the (asymmetric) limits δ^+ and δ^- from the OLS regression

$$y_t = \delta^+ z_{1t} - \delta^- z_{2t} + u_t. \tag{6}$$

Such a regression may provide a better fit since there is little a priori reason to assume that thresholds should be symmetric. We denote this estimator as "ALS" in the following.

As an alternative to the probability approach, Pesaran (1984, 1985) proposed a method known as the *regression approach*. Assume that conditional on reporting $r_{it} = 1$ ("up") the expectation of y_{it}^e is α, whereas $E(y_{it}^e | r_{it} = -1) = \beta$ and $E(y_{it}^e | r_{it} = 0) = 0$. It follows that for large N_t

$$\mu_t \approx \alpha p_t^+ + \beta p_t^- \tag{7}$$

and, therefore, quantitative reference values of the qualitative expectations can be obtained as the fitted values of the regression equation

$$y_t = \alpha p_t^+ + \beta p_t^- + e_t$$

The widely-used balance statistic (BS) of Anderson (1952) results as a special case by setting $\alpha = -\beta = 1$. As argued by Pesaran (1987) and Breitung (2008), the regression

approach results as a special case of the CP methodology by assuming that the survey expectations are uniformly distributed.

5 Empirical results based on time invariant thresholds

We now present empirical results of the quantification procedures described above and highlight some interesting similarities and differences. Furthermore, we provide some preliminary findings concerning the relative performance of alternative quantification methods.

The left panel of Table 1 shows results for regressions of the form (4)-(7). We employ six-months ahead DAX returns as the dependent variable in our regressions. Using the CP-method and future actual returns yields a threshold estimate of 2.96% which is not significantly different from zero. We also present results for the symmetric LS variant of the CP method obtained from the regression $y_t = \delta z_t + u_t$ with z_t entering the forecasters' information set yielding $\widehat{\delta}_{ls} = (\sum y_t z_t) / (\sum z_t^2)$. The threshold estimate is now statistically significant with a value of 3.8%. The asymmetric LS method (ALS) obtained from regression (6) yields insignificant parameter estimates $\widehat{\delta}^{+} = -0.58$ and $\widehat{\delta}^{-} = -18.70$ which differ strongly from the estimates of the other variants above. One possible explanation for this result is that the regressors z_{1t} and z_{2t} are highly multicollinear (with a variance inflation factor of 3.78).

Table 1 about here

As a first indication of the performance of the quantification procedure, Figure 3 compares the quantified survey expectations $\widehat{y}_t$ obtained from the CP estimator $\widehat{y}_t = \widehat{\delta}_{cp} z_t$ and the average of the reported quantitative expectations.[3] The correlation between quantified expectations $\widehat{y}_t$ (either computed by the CP or LS approach) and $\widehat{\mu}_t$ is about 0.45 and, thus, rather low. Also, there are large deviations between quantitative and quantified expectations for sustained periods of time (e.g. $\widehat{\mu} = 3\%$ versus 12% derived from the

[3]Using the LS estimator instead of the CP estimator yields an almost identical figure.

CP method around 2005). In sum, our preliminary results already provide a first indication that the probability approach may not yield reliable measures of the underlying quantitative expectations.

FIGURE 3 ABOUT HERE

This result is rather uncomforting and already suggests that care has to be taken when interpreting quantified survey expectations as proxies for actual expectations. Hence, we present results from a simple, but nevertheless instructive, diagnostic test. The CP procedure (and many variants thereof) identify the threshold parameters by assuming unbiased expectations conditional on some information set. Since we have quantitative expectations at hand, we can circumvent this assumption and directly identify the threshold parameters by regressing actual quantitative expectations ($\widehat{\mu}_t$) on z_t (or the share of up and down votes in the regression approach). This seems interesting since forecaster's rationality is not warranted and even if forecasts were unbiased, there is no reason to believe that estimates based on this assumption are very reliable in small samples.

The right panel of Table 1 reports results for this regression setup. Using the CP and LS methods with average survey expectations $\widehat{\mu}_t$ yields results that appear much more reasonable and in line with economic intuition. Both the CP and the LS method yield a highly significant threshold value of roughly 2% (the two estimates only differ after the third digit), whereas the ALS method yields values of about 1.8% for δ^+ and about -3.3% for δ^-. These estimates seem much more reliable than those obtained from actual future returns and their (absolute) magnitude seems intuitively reasonable. Finally, Pesaran's (1984, 1985) regression-based approach (column "PRA" in Table 1) yields estimates which are not directly interpretable in terms of threshold values of some indifference interval, but we generally find similar results as above.

It is important to note that the R^2s of the regressions with quantitative expectations ($\widehat{\mu}_t$) as dependent variables are surprisingly low (only 0.2 in three out of four regressions). If Assumption CP is fulfilled the expectation is a linear combination of z_{1t} and z_{2t} (cf. (3)) and, therefore, we expect an R^2 close to one in a regression of $\widehat{\mu}_t$ on z_{1t} and z_{2t}.

This suggests that probability-based approaches and Pesaran's regression approach do not track the (usually unobserved) quantitative expectations very well. The regression approach (or the balance statistic) yield qualitatively similar results and, therefore, we do not present the respective results from the subsequent steps of our analysis.

For robustness, we have also employed a sample that stops directly before the recent financial crisis in June 2007. We find that these even strengthens our main results. The threshold parameters implied by the CP method (and their variants) become quite large while the same is not true for the threshold estimates based on quantitative expectations. For example, the standard CP method leads to a threshold estimate of 6.4% and the LS method leads to 5.8%, both of which are quite high relative to typical stock market movements during that time period. In contrast, regressing quantitative expectations on z_t (as in the right panel of Table 1) delivers almost identical threshold estimates of e.g. 1.81% for the CP method and 1.88% for the LS method. Hence, our main result is robust to an in- or exclusion of the crisis and, if anything, rather strengthens our argument.

6 Time-varying thresholds

Two of the main assumptions of the CP methodology are that (i) returns are normally distributed and (ii) that the indifference thresholds are constant over time. We have already shown above (Figure 2) that return expectations are not normal but given the extensive evidence in the literature that non-normality does not matter too much (e.g. Mitchell, 2002), we do not expect non-normality to be an important driver of our results (we will validate this conjecture below). However, it seems reasonable to assume that forecasters have heterogeneous threshold values and that these thresholds vary over time, for example in response to changing market volatility. In this section we investigate whether these threshold levels are indeed time invariant (as assumed in the CP approach) and we consider two different methods for estimating the thresholds δ_t^+ and δ_t^- for each time period separately.

6.1 Parametric estimation

Let

$$\widehat{\mu}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} y_{it}^e \tag{8}$$

$$\text{and } \widehat{\sigma}_t^2 = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_{it}^e - \widehat{\mu}_t)^2 \tag{9}$$

denote the first two sample moments of the expectations y_{it}^e at period t. Under Assumption CP we have

$$q_t^- = \Phi^{-1}(p_t^-) \simeq (\delta_t^- - \widehat{\mu}_t)/\widehat{\sigma}_t \tag{10}$$

$$q_t^+ = \Phi^{-1}(1-p_t^+) \simeq (\delta_t^+ - \widehat{\mu}_t)/\widehat{\sigma}_t \ . \tag{11}$$

For $N_t \to \infty$ these approximations become identities provided that the conditions for the weak law of large numbers are satisfied. From these relationships we obtain the following estimators for the indifference limits at period t:

$$\widehat{\delta}_t^+ = \widehat{\sigma}_t q_t^+ + \widehat{\mu}_t \tag{12}$$

$$\widehat{\delta}_t^- = \widehat{\sigma}_t q_t^- + \widehat{\mu}_t \ . \tag{13}$$

The resulting estimators may be used to investigate whether the assumption of fixed limits (Assumption CP (b)) is valid.

Figure 4, Panels (a) and (b), presents the estimators $\widehat{\delta}_t^+$ and $\widehat{\delta}_t^-$ along with simulated 95% confidence intervals. The confidence intervals are obtained from 10,000 estimates of the threshold parameters, where the individual expectations are simulated by normally distributed random variables with expectation $\widehat{\mu}_t$ and variance $\widehat{\sigma}_t^2$. For reference the symmetric estimates $\widehat{\delta}_{ls}$ estimated by the CP and LS methods are also included as horizontal lines. The results clearly indicate that the assumption of fixed indifference level is violated.

FIGURE 4 ABOUT HERE

6.2 Nonparametric estimation

Since the descriptive analysis of Section 2 clearly indicates that the normality assumption is violated for the vast majority of time periods, we also apply a nonparametric estimation procedure, where y^e_{it} is allowed to have an arbitrary distribution function $F_t(z)$ that may be different from the normal distribution. Note that

$$p^-_t = Prob(y^e_{it} < \delta^-_t) \tag{14}$$

$$= F_t(\delta^-_t) \tag{15}$$

and, therefore, $\delta^-_t = F_t^{-1}(p^-_t)$. The inverse of the empirical distribution function $\widehat{F}_t(\cdot)$ can be obtained from the ranks of y^e_{it}. To this end we form the ranks of y^e_{it}, yielding the ordered set $\{y^e_{(1),t}, \ldots, y^e_{(N_t),t}\}$, where $y^e_{(1),t}$ is the smallest value of the set $\{y^e_{1t}, \ldots, y^e_{N_t t}\}$ and $y^e_{(N_t),t}$ is the largest value. Next, we select the value of $y^e_{(i),t}$ $(i = 1, \ldots, N_t)$ such that its rank is equal (or closest) to $N_t \cdot p^-_t$. This value is the nonparametric estimator of the lower indifference limit and it is labeled as $\widetilde{\delta}^-_t$. In a similar manner the upper limit results as the value of $y^e_{(i),t}$ for which $1 - (i/N_t)$ comes closest to p^+_t.

To compute the standard errors for the nonparametric estimator, a bootstrap procedure is employed. Let $\mathcal{Y}^e_{it}$ denote the bootstrap analog of y^e_{it} obtained by drawing (with replacement) from the sample $\{y^e_{1t}, \ldots, y^e_{N_t t}\}$. From the new sample $\{\mathcal{Y}^e_{1t}, \ldots, \mathcal{Y}^e_{N_t t}\}$ we estimate the lower and upper limits by using the nonparametric approach obtaining the threshold estimates $\widetilde{\delta}^-_t(\mathcal{Y}^e_{it})$ and $\widetilde{\delta}^+_t(\mathcal{Y}^e_{it})$. The confidence intervals for the nonparametric estimators of the indifference limits can be estimated as the $\alpha/2$ and $1 - \alpha/2$ quantiles of the simulated distribution of $\widetilde{\delta}^-_t(\mathcal{Y}^e_t)$ and $\widetilde{\delta}^+_t(\mathcal{Y}^e_t)$, where α denotes the probability level of the confidence interval.[4]

The resulting estimates are presented in Figure 5 along with 95%-confidence intervals. It turns out that the nonparametric estimates of the indifference limits correspond fairly well to the parametric estimates presented in Table 4 and, therefore, the violation of the normality assumption does not have an important effect on the estimated threshold as argued above. Therefore, the results do not change much if the parametric estimates of

[4]Note that this bootstrap treats each cross-section separately and, hence, delivers correct *point-wise* standard errors for each month. However, care must be taken when looking at the whole path of standard errors where time-dependence of forecasts would have to be taken into account.

the limits are replaced by nonparametric estimates.

Next we investigate whether the use of the time-varying nonparametric estimates of the thresholds is able to improve the fit of the quantitative expectations. To this end we compute quantified expectations according to empirical analog of eq. (3)

$$\widetilde{\mu}_t^* \;=\; \widetilde{\delta}_t^{+} z_{1t} - \widetilde{\delta}_t^{-} z_{2t}$$

where the fixed thresholds are replaced by the estimated time varying thresholds.

If the qualitative data is generated by the probability model characterized by Assumption CP but with time varying thresholds, we expect that $\widetilde{\mu}_t^*$ is highly correlated with the sample means of the expectations $\widehat{\mu}_t$. In fact the substantially improved correlation of 0.95 (instead of 0.45) suggests that the lack of fit of the original CP method can be explained (at least to a large extent) by imposing time-constant thresholds.

FIGURE 5 ABOUT HERE

6.3 Explaining time varying thresholds

Since our empirical results point to a substantial time-variation of the indifference limits we further investigate the temporal variation by estimating variants of the following dynamic regressions:

$$\widetilde{\delta}_t^{+} \;=\; a_0 + a_1\, \widetilde{\delta}_{t-1}^{+} + \beta' x_t + e_t \;, \tag{16}$$

where $\widetilde{\delta}_t^{+}$ denotes the nonparametric estimate of the upper threshold in period t and x_t represents additional explanatory variables such as past returns or standard deviations. A similar regression is performed for $\widetilde{\delta}_t^{-}$. Based on the theory of signal extraction Bachelor and Orr (1988) specify $x_t = \sigma_t^2$ and estimate the parameters using a data set on inflation expectations as $\widehat{a}_1 = 0.16$, $\widehat{a}_1 = 0.27$ and $\widehat{\beta} = 0.11$.

Our estimation results based on different specifications are presented in Table 2. The standard errors of the coefficients are computed by using robust (HAC) standard errors

as suggested by Newey and West (1987). The first column presents the estimation results for the specification proposed by Batchelor and Orr (1988). For positive thresholds, our estimates of the constant and the lagged thresholds are somewhat larger but the estimate for the lagged standard deviation correspond well to the estimate reported by Batchelor and Orr (1988). Including lagged returns of the last month or the last 6 months (the time interval that is used to compute the dependent variable) yields a substantial improvement of the fit. Somewhat surprisingly, past returns are negatively correlated with the thresholds. This may suggest that following an increase in stock prices the survey participants tend to match their current view on future stock prices with the positive market conditions by lowering the upper and lower thresholds. On the other hand, an increase in lagged return volatility (measured as lagged volatility over one month or six months based on daily data) only has a marginal effect on the indifference limits. Similarly, the time dependent volatility (as measured by the conditional variances of a GARCH(1,1) specification for the returns) do not contribute much for explaining temporal fluctuations of the indifference thresholds. In any case our results indicate that the assumption of fixed indifference limits that underlies the probability approach is grossly violated.

Table 2 about here

Overall, our analysis suggests that incorporating additional information when modeling the threshold levels maybe a fruitful exercise. This seems especially relevant, since we find that *economic* variables (especially lagged returns in our setup) add quite a bit of explanatory power. Hence, simple time-series models for thresholds (as suggested by the AR(1) component in our setup) or time-varying parameter models[5] do not necessarily provide the best fit.

[5]See e.g. Nardo (2003) for an overview of these models and Claveria et al. (2007) for a recent application. Also see Seitz (1988) or, more recently, Henzel and Wollmershäuser (2005) for approaches based on time-varying thresholds.

7 Individual specific heterogeneity

So far we have focused on temporal heterogeneity. In this section we investigate the effects of individual specific heterogeneity.[6] To this end we assume that the indifference thresholds and the mean of the expectations vary across individuals such that

$$\begin{aligned} \delta_i^+ &= \delta^+ + \nu_i^+ \\ \delta_i^- &= \delta^- + \nu_i^- \\ \mu_{it} &= \mu_t + \eta_i \end{aligned}$$

where ν_i^+, ν_i^- and η_i are random disturbances with expectations equal to zero. The probability of a negative response is given by

$$p_t^- = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{\delta^- - \mu_t}{\sigma_t} + \frac{\epsilon_{1i}}{\sigma_t}\right), \tag{17}$$

where $\epsilon_{1i} = \nu_i^- + \eta_i$. If $E(\epsilon_{1i}^2) = \sigma_{\epsilon_1}^2$ is small relative to σ_t^2 we can apply a Taylor series expansion yielding

$$\begin{aligned} p_t^- &= \Phi\left(\frac{\delta^- - \mu_t}{\sigma_t}\right) + \phi\left(\frac{\delta^- - \mu_t}{\sigma_t}\right)\frac{\frac{1}{n}\sum_{i=1}^{n}\epsilon_{1i}}{\sigma_t} + O_p(\sigma_{\epsilon_1}^2/\sigma_t^2) \qquad (18) \\ &= \Phi\left(\frac{\delta^- - \mu_t}{\sigma_t}\right) + O_p(n^{-1/2}) + O_p(\sigma_{\epsilon_1}^2/\sigma_t^2), \end{aligned}$$

where $\phi(\cdot)$ denotes the density function of the standard normal distribution.[7] This suggest that the distortions due to mild individual specific heterogeneity are negligible if n gets large. A similar reasoning applies to p_t^+. Accordingly, introducing temporal heterogeneity seems sufficient to explain the low correlation between the CP measure and the quantitative expectations. We also investigate this issue in more detail below by means of a simulation experiment.

It is nevertheless interesting to assess the individual specific heterogeneity in the expectations of the respondents. To illustrate our empirical strategy, assume that under

[6] There are several reasons why forecasters might be heterogeneous. One reason considered in the more recent literature is that forecasters might employ different loss functions when forming their forecasts (see, e.g., Elliott et al., 2005). Das et al. (1999) and Lui, Mitchell and Weale (2011a) have investigated this issue on the basis of qualitative forecasts. For the analysis in this paper, we abstract from the specific source of heterogeneity but simply assume that forecasters have the same loss function when forming qualitative and quantitative expectations.

[7] The analysis based on a Taylor series expansion is inspired by and similar to the analysis in Mitchell, Smith and Weale (2002).

the null hypothesis all parameters of the probability model comprised in the vector $\theta_t = (\mu_t, \sigma_t, \delta_t^-, \delta_t^+)'$ are known. The qualitative indicator $r_{it} \in \{1, 0, -1\}$ can be represented as

$$\begin{aligned} r_{it} &= E(r_{it}) + e_{it} \\ &= P(r_{it} = 1) - P(r_{it} = -1) + e_{it} \\ &= \left[1 - \Phi\left(\frac{\delta_t^+ - \mu_t}{\sigma_t}\right)\right] - \Phi\left(\frac{\delta_t^- - \mu_t}{\sigma_t}\right) + e_{it} \\ &= G(\theta_t) + e_{it}. \end{aligned} \tag{19}$$

If the probability model is correctly specified we have $E(r_{it}) = G(\theta_t)$ or $E(e_{it}) = 0$ for all i and t. On the other hand, if some parameters of the model are individual specific, then $E(e_{it}) \neq 0$. This suggest to compare the mean of the observed responses of individual i

$$\overline{r}_i = \frac{1}{T}\sum_{t=1}^{T} r_{it}$$

with the expected value based on the probability model $\overline{G}(\widehat{\theta}) = T^{-1}\sum_{t=1}^{T} G(\widehat{\theta}_t)$, where the parameters in the vector θ_t are replaced by their sample counterparts, i.e., we employ the (cross-sectional) average quantitative return expectations $\widehat{\mu}_t$, standard deviations $\widehat{\sigma}_t$, and non-parametric threshold estimates $\widehat{\delta}_t^+$, $\widehat{\delta}_t^-$ from our analysis above. If $\overline{r}_i > \overline{G}(\widehat{\theta})$, then the respondent is relative optimistic, whereas a pessimistic respondent is indicated by observing $\overline{r}_i < \overline{G}(\widehat{\theta})$.

First we apply a simple t-statistic to test the hypothesis that the model is *overall* correctly specified.[8] We observe

$$\left(\frac{1}{n}\sum_{i=1}^{n}\overline{r}_i\right) - \overline{G}(\widehat{\theta}) \;=\; 0.0764$$

with associated t-statistic of 0.065. We therefore tentatively conclude that the (homogenous) probability model is able to reproduce the overall mean of the qualitative indicator. For individual probabilities, however, the Taylor expansion according to (18) yields

$$P(r_{it} = -1) = \Phi\left(\frac{\delta^- - \mu_t}{\sigma_t}\right) + \phi\left(\frac{\delta^- - \mu_t}{\sigma_t}\right)\frac{\epsilon_{1i}}{\sigma_t} + O_p(\sigma_{\epsilon_1}^2/\sigma_t^2) \tag{20}$$

$$P(r_{it} = 1) = 1 - \Phi\left(\frac{\delta^+ - \mu_t}{\sigma_t}\right) - \phi\left(\frac{\delta^+ - \mu_t}{\sigma_t}\right)\frac{\epsilon_{2i}}{\sigma_t} + O_p(\sigma_{\epsilon_2}^2/\sigma_t^2), \tag{21}$$

[8]For this test we assume that responses are independent across individuals and that the estimation error in $\overline{G}(\widehat{\theta})$ is negligible.

where $\epsilon_{2i} = \nu_i^+ + \eta_i$. Therefore, the difference between $\overline{r}_i$ and $\overline{G}(\theta)$ results as

$$E(\overline{r}_i|\epsilon_i) - G(\theta) \;\approx\; \varphi_1(\theta)\epsilon_{1i} + \varphi_2(\theta)\epsilon_{2i} \tag{22}$$

where

$$\varphi_1(\theta) \;=\; -\sum_{t=1}^{T} \frac{\phi_t^-(\theta_t)}{\sigma_t} \quad \text{and} \quad \varphi_2(\theta) \;=\; -\sum_{t=1}^{T} \frac{\phi_t^+(\theta_t)}{\sigma_t}$$

and $\phi_t^a(\theta_t) = \phi[(\delta_t^a - \mu_t)/\sigma_t]$ with $a \in \{+, -\}$. Accordingly, the difference between $\overline{r}_i$ and $\overline{G}(\theta)$ are (approximately) proportional to the individual effects ϵ_{1i} and ϵ_{2i}.

Assuming that the probability model is well specified, we have $\widehat{\theta} = (\hat{\mu}_t, \hat{\sigma}_t, \hat{\delta}_t^-, \hat{\delta}_t^+)' = \theta + O_p(n^{-1/2})$ and $G_t(\widehat{\theta}) = G(\theta) + O_p(n^{-1/2})$. Thus, the t-statistic for $E(\epsilon_{1i}) = E(\epsilon_{2i}) = 0$ results as

$$\begin{aligned} \overline{m}_i(\widehat{\theta}) \;&=\; \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \widehat{e}_{it} \\ &=\; \left(\frac{1}{\sqrt{T}} \sum_{t=1}^{T} e_{it} \right) + O_p\left(\sqrt{\frac{T}{n}} \right). \end{aligned} \tag{23}$$

If $T/n \to 0$ it follows from the central limit theorem that

$$\overline{m}_i(\widehat{\theta}) \;\xrightarrow{d}\; \mathcal{N}(0, s_i^2) \tag{24}$$

where

$$s_i^2 = \lim_{N,T\to\infty} E\left[\frac{1}{T} \left(\sum_{t=1}^{T} e_{it} \right)^2 \right].$$

Note that due to the overlapping horizon of the expectations, the expectation errors are autocorrelated up to the fifth lag. We therefore estimate the variance s_i^2 by using the HAC standard errors as suggested by Newey and West (1987).

Applying this testing strategy to those respondents that participate in at least 30 time periods (yielding $n = 257$) we obtain 44.4 percent of the test statistics $t_i(\widehat{\theta}) = \overline{m}_i(\widehat{\theta})/\widehat{s}_i$ $(i = 1, \ldots, 257)$ larger than 1.96, whereas 21.0 percent of the test statistics are smaller than -1.96. That is, in only 34.6 percent of the cases the test statistics accept the null hypothesis that the residuals have a zero mean. Under the null hypothesis the test statistics are asymptotically i.i.d. $\mathcal{N}(0, 1)$. For large N the rejection frequency under the null hypothesis is (approximately) normally distributed with $\mathcal{N}(0.05, 0.05 \cdot 0.95/n)$ yielding an 0.05 upper critical value for the rejection frequency of 0.077. Since the actual

rejection rate of 0.654 is much higher than the upper limit of 0.077 we can safely reject the null hypothesis of individual homogeneity.

8 Simulation experiments

Finally, we present results of two simulation experiments designed to investigate the reasons for the poor performance of the CP methodology. The first experiments examines whether time-varying thresholds alone can reproduce the poor performance of the CP procedure documented in Table 1. We find that introducing time-varying thresholds go a long way towards reproducing our empirical findings above. The second experiment additionally considers forecaster-specific heterogeneity in thresholds (and return expectations) to discriminate between the importance of pure time-variation in thresholds and heterogeneity between forecasters. We find that adding forecaster-specific heterogeneity leads to an even closer fit between simulated results and our empirical results based on actual data. However, the improvements are fairly minor and suggest that time-varying thresholds capture the lion's share of the poor performance of the CP procedure.

8.1 Time variation in thresholds

To examine the effect of time-varying thresholds on the performance of the CP procedure, we simulate quantitative expectations for each forecaster distributed as $\mathcal{Y}_{it}^e \sim \mathcal{N}(\widehat{\mu}_t, \widehat{\sigma}_t^2)$, where $\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ are sample average and variance of the reported survey expectations. We transform these quantitative expectations into qualitative expectations ($\widetilde{r}_{i,t}$) by comparing the artificial quantitative expectation to the upper and lower threshold values

$$\widetilde{r}_{it} = \begin{cases} 1, & \text{if} \quad \mathcal{Y}_{it}^e > \delta_t^+ \\ 0, & \text{if} \quad \delta_t^- \leq \mathcal{Y}_{it}^e \leq \delta_t^+ \\ -1, & \text{if} \quad \mathcal{Y}_{it}^e < \delta_t^- \ . \end{cases} \tag{25}$$

for $i = 1, \ldots, N$. From the simulated qualitative expectations $\widetilde{r}_{it}$ we calculate the time-series for the regressor $\widetilde{z}_t$ as described in Section 3. Running a regression of $\widetilde{\mu}_t = N^{-1} \sum_{i=1}^N \mathcal{Y}_{it}^e$ on $\widetilde{z}_t$ yields the estimated threshold parameter $\widehat{\delta}_{ls}$ and the regression R^2. In all our simulations, we set the number of forecasters equal to 200 and $T = 69$ months.[9]

[9]Increasing the number of cross-sections has no material effect on our results.

We employ five different simulation setups. Setup I sets a symmetric and constant threshold equal to the time-series mean of our parametric threshold estimates from above. To generate data with symmetric thresholds we set $\widehat{\delta}_{sym} = T^{-1} \sum_{t=1}^{T} \frac{1}{2}(\widetilde{\delta}_t^+ - \widetilde{\delta}_t^-)$. The time constant parameter σ^2 is fixed at the sample mean of $\widehat{\sigma}_t^2$. Setup II also employs constant thresholds and a constant standard deviation but we allow for asymmetric thresholds by setting the upper and lower threshold equal to the sample means of the respective threshold time-series $\widehat{\delta}_t^+$ and $\widehat{\delta}_t^-$. Setup III relaxes the assumption of constant standard deviations and we employ the time series of estimated cross-sectional standard deviations $\widehat{\sigma}_t$ obtained from the original data set. Setup IV uses time-constant standard deviation (again computed as square root of the sample mean of $\widehat{\sigma}_t^2$) but now we allow the thresholds to vary over time. We employ the parametric estimates $\widehat{\delta}^+$ and $\widehat{\delta}^-$ to simulate expectations in this setup. Finally, Setup V allows for time-varying thresholds (as in setup IV) and time-varying standard deviations (as in setup III).

Simulation results based on 25,000 repetitions for each setup are shown in Table 3. Panel A shows percentiles of the simulated distribution of estimated threshold parameters whereas Panel B reports the same for regression R^2s. Considering the median values first, we find that the most restrictive setup I yields an estimate for δ of 2.78 and a median R^2 of 0.91. This can be compared to our estimate for the actual survey data with $\widehat{\delta}_{ls} = 1.96$ and an R^2 of 0.20 (see Table 1). Thus, we find that both the threshold parameter and the R^2 are much to high. In fact, the estimated threshold parameter is significantly different from 1.96 on a 5%-level and an R^2 of 0.20 as in the actual data has a chance of being observed of less than 0.5%. Thus, the constant and symmetric threshold model with constant standard deviations is clearly not supported by our data.

Setup II (asymmetric and constant thresholds, constant standard deviation) yields threshold estimates which are closer to their in-sample counterparts in Table 1 but still yields very large values for the regression R^2 which are not observed in the real-world data. Similar results are obtained for setup III. However, setups IV and V, which both allow for time-varying thresholds, lead to a distribution of threshold parameter estimates and R^2 which are well in line with our empirical findings reported above. The median threshold parameters are close to 2 and the median R^2 are 0.23 and 0.25, respectively. These values are very close to our empirical findings of $\widehat{\delta}_{ls} = 1.96$ and an R^2 of 0.20 in Table 1. There-

fore we conclude that time varying thresholds alone are able to explain most of the poor observed fit between quantitative and quantified expectations documented in Section 3.

8.2 Individual heterogeneity in thresholds

In a second simulation experiment, we add forecaster-specific heterogeneity by allowing for noise in individual thresholds. More specifically, we employ the same five simulation setups as above, but add normally distributed error terms to the thresholds, so that $\widehat{\delta}_{i,sym} = \widehat{\delta}_{sym} + \nu_i$ (Setup I), $\widehat{\delta}_i^+ = \widehat{\delta}^+ + \nu_i^+$ and $\widehat{\delta}_i^- = \widehat{\delta}^- + \nu_i^-$ (Setup II and III), or $\widehat{\delta}_{i,t}^+ = \widehat{\delta}_t^+ + \nu_{i,t}^+$ and $\widehat{\delta}_{i,t}^- = \widehat{\delta}_t^- + \nu_{i,t}^-$ (Setup IV and V). All error terms have mean zero and we consider standard deviations of $0\%, 10\%, 20\%, ..., 100\%$ of the standard deviation of expected returns. The latter specification just serves to benchmark the "amount" of forecaster heterogeneity to some observable variable. Note that a relative standard deviation of 0% basically reproduces the experiment in the preceding section whereas a relative standard deviation of 100% implies that thresholds across forecasters are as dispersed as return expectations.[10]

Results for these simulations are shown in Table 4 where we report CP threshold estimates in Panel A, and regression R^2s in Panel B as in Table 3 above. However, to conserve space, we only report medians across the 25,000 simulations. As can be inferred from these results, allowing for an increasing heterogeneity of forecasters tends to decrease the size of simulated threshold estimates (Panel A) and R^2s and drives them closer to the benchmark of $\delta_{ls} = 1.96$, $R^2 = 0.20$ (Table 1). However, allowing for forecaster-specific thresholds alone does not seem to suffice to reproduce our empirical findings from Section 3 above. As in our simulations above, we are only able to reproduce the low observed R^2 when allowing for time-varying thresholds as in Setup IV and V. For these two setups, we do find, however, that introducing forecaster-specific heterogeneity allows for a more or less perfect replication of our benchmark results in Table 1. For example, a relative standard deviation of 90% leads to a median threshold estimate of 1.95 and median R^2 of 20% in Setup V which is extremely close to the values we find for our actual data in Table

[10] Other papers have also investigated threshold heterogeneity across forecasters, see e.g., Lui, Mitchell, Weale (2011a) and Lui, Mitchell, Weale (2011b) which directly estimate thresholds from a panel of individual forecasters.

1. Hence, these results corroborate our claim above that time-variation in thresholds captures the lion's share of the observed poor performance of the CP procedure whereas forecaster-specific heterogeneity plays a minor role.

Table 4 about here

9 Conclusions

In this paper we analyze a unique sample of individual stock market forecasts consisting of both qualitative and quantitative forecasts issued by the same forecasters. The probability approach for quantifying qualitative survey data suggested by Carlson and Parkin (1975) implies a strong and stable correlation among qualitative and quantitative expectations. In contrast we observe that in our data set the reported quantitative figures are only weakly correlated with the respective qualitative forecasts. We investigate several potential explanations for this surprising result. Although we find evidence for severe violations of the normality assumption in most of the time periods, this does not seem to have an important effect on the relationship between quantitative and qualitative expectations. On the other hand, we find that temporal and individual heterogeneity of the indifference thresholds are able to explain the break-down in the correlation between quantitative and qualitative forecasts.

Provided that qualitative expectations are only weakly correlated with quantitative expectations, the classical probability approach renders measures of market expectations unreliable. Since we find that temporal variation in indifference thresholds is a major source of the low correlation between qualitative and quantitative expectations, we propose to specify explicitly the indifference limits in the questionnaire. For example, one may ask the respondents whether they think that stock prices will rise/fall more than ± 5 percent during the next 6 months. Such a detailed definition of qualitative expectations seems practical and will eventually improve the reliability of quantified survey expectations.

Furthermore, we show that there is substantial cross-sectional heterogeneity across forecasters which also interferes with standard quantification procedures (although to a lesser degree than time-varying thresholds in our stock market setting). Yet, many published survey results do not provide information about the dispersion in forecasts so that it seems impossible to deal with this heterogeneity in applied research. We thus suggest to publish more information and not just the share of optimistic and pessimistic forecasters. Alternatively, an even better solution would be to publish data on the full panel of forecasters which would allow for improved quantification methods (Mitchell, Smith, Weale, 2002, 2005).

References

Anderson, O. (1952), The Business Test of the IFO-Institute for Economic Research, Munich, and its Theoretical Model, Review of the International Statistical Institute, 20, 1–17.

Anderson, E., E. Ghysels, and J. Juergens (2009), The Impact of Risk and Uncertainty on Expected Returns, Journal of Financial Economics, 94, 233–263.

Batchelor, R. (1982), Expectations, Output and Inflation: The European Experience, European Economic Review, 17, 1-25.

Batchelor, R. (1986), Quantitative v. Qualitative Measures of Inflation Expectations, Oxford Bulletin of Economics and Statistics, 48, 99 - 120.

Batchelor, R. and A. Orr (1988), Inflation Expectations Revisited, Economica, 55, 317-331.

Beber, A., A. Buraschi, and F. Breedon (2010), Difference in Beliefs and Currency Risk Premia, Journal of Financial Economics, forthcoming.

Berk, J. (1999), Measuring Inflation Expectations: A Survey Data Approach, Applied Economics, 31, 1467-1480.

Breitung, J. (2008), Assessing the Rationality of Survey Expectations: The Probability Approach, Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik), 228, 6, 630–643.

Carlson, J. and M. Parkin (1975), Inflation Expectations, Economica, 42, 123-138.

Claveria, O., E. Pons, and R. Ramos (2007), Business and Consumer Expectations and Macroeconomic Forecasts, International Journal of Forecasting, 23, 47–69.

Das, M., J. Dominitz, and A. van Soest (1999), Comparing Predictions and Outcomes: Theory and Application to Income Changes, Journal of the American Statistical Association, 94, 75–85.

Dasgupta, S. and Lahiri, K. (1992), A Comparative Study of Alternative Methods of Quantifying Qualitative Survey Responses using NAPM Data, Journal of Business and Economic Statistics, 10, 391–400.

Defris, L. and R. Williams (1979), Quantitative versus Qualitative Measures of Price Expectations, Economics Letters, 2, 169–173.

Doepke, J., J. Dovern, U. Fritsche, and J. Slacalek (2008), The Dynamics of European Inflation Expectations, The B.E. Journal of Macroeconomics, 8, Article 12.

Elliott, G., I. Komunjer, and A. Timmermann (2005), Estimation and Testing of Forecast Rationality under Flexible Loss, Review of Economic Studies, 72, 1107–1125.

Fishe, R. and Lahiri, K. (1981), On the Estimation of Inflationary Expectations from Qualitative Responses, Journal of Econometrics, 16, 89-102.

Frankel, J. and K. Froot (1990), Chartists, Fundamentalists, and Trading in the Foreign Exchange Market, American Economic Review (P & P), 80, 181-185.

Henzel, S. and T. Wollmershäuser (2005), An Alternative to the Carlson-Parkin Method for the Quantification of Qualitative Inflation Expectations: Evidence from the Ifo World Economic Survey, Ifo Working Papers.

Löffler, G. (1999), Refining the Carlson-Parkin Method, Economics Letters 64, 167–171.

Lui, S., Mitchell, J., Weale, M. (2011a), The Utility of Expectational Data: Firm-Level Evidence Using Matched Qualitative-Quantitative UK Surveys, International Journal of Forecasting, 27, 1128–1146.

Lui, S., Mitchell, J., Weale, M. (2011b), Qualitative Business Surveys: Signal or Noise?, Journal of the Royal Statistical Society: Series A, 174, 327–348.

Maag, T. (2010), On the Accuracy of the Probability Method for Quantifying Beliefs about Inflation, KOF Working Papers No. 230, KOF Zurich.

Mankiw, G., Ricardo Reis, and J. Wolfers (2004), Disagreement about Inflation Expectations, NBER Macroeconomics Annual 2003, 209–248.

Menkhoff, L., R. Rebitzky, and M. Schröder (2009), Heterogeneity in Exchange Rate Expectations: Evidence on the Chartist-Fundamentalist Approach, Journal of Economic Behavior and Organization 70, 241–252.

Mitchell, J. (2002), The Use of Non-Normal Distributions in Quantifying Qualitative Survey Data on Expectations, Economics Letters, 76, 101–107.

Mitchell, J., K. Mouratidis, and M. Weale (2007), Uncertainty in UK Manufacturing: Evidence from Qualitative Survey Data, Economics Letters, 94, 242–252.

Mitchell, J., Smith, R., Weale, M. (2002), Quantification of Qualitative Firm-level Survey Data, Economic Journal, 112, 117–135.

Mitchell, J., Smith, R., Weale, M. (2005), Forecasting Manufacturing Output Growth Using Firm-Level Survey Data, The Manchester School, 73, 479–499.

Müller, C. (2010), You CAN Carlson-Parkin, Economics Letters, 108, 33–35.

Nardo, M. (2003), The Quantification of Qualitative Survey Data: a Critical Assessment, Journal of Economic Surveys, 17, 645-668.

Newey, W.K. and K.D. West (1987), A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, Econometrica, 55, 703–708.

Nolte, I. and W. Pohlmeier (2007), Using Forecasts of Forecasters to Forecast, International Journal of Forecasting, 23, 15–28.

Patton, A. and A. Timmermann (2010), Why do Forecasters Disagree? Lessons from the Term Structure of Cross-Sectional Dispersion, Journal of Monetary Economics, 57, 803–820.

Pesaran, M.H. (1984), Expectations formation and macroeconomic modelling, in P. Magrange and P. Muet, ed., Contemporary Macroeconomic Modelling, Blackwell, Oxford, 27-53.

Pesaran, M.H. (1985), Formation of Inflation Expectations in British Manufacturing Industries, Economic Journal, 95, 948–975.

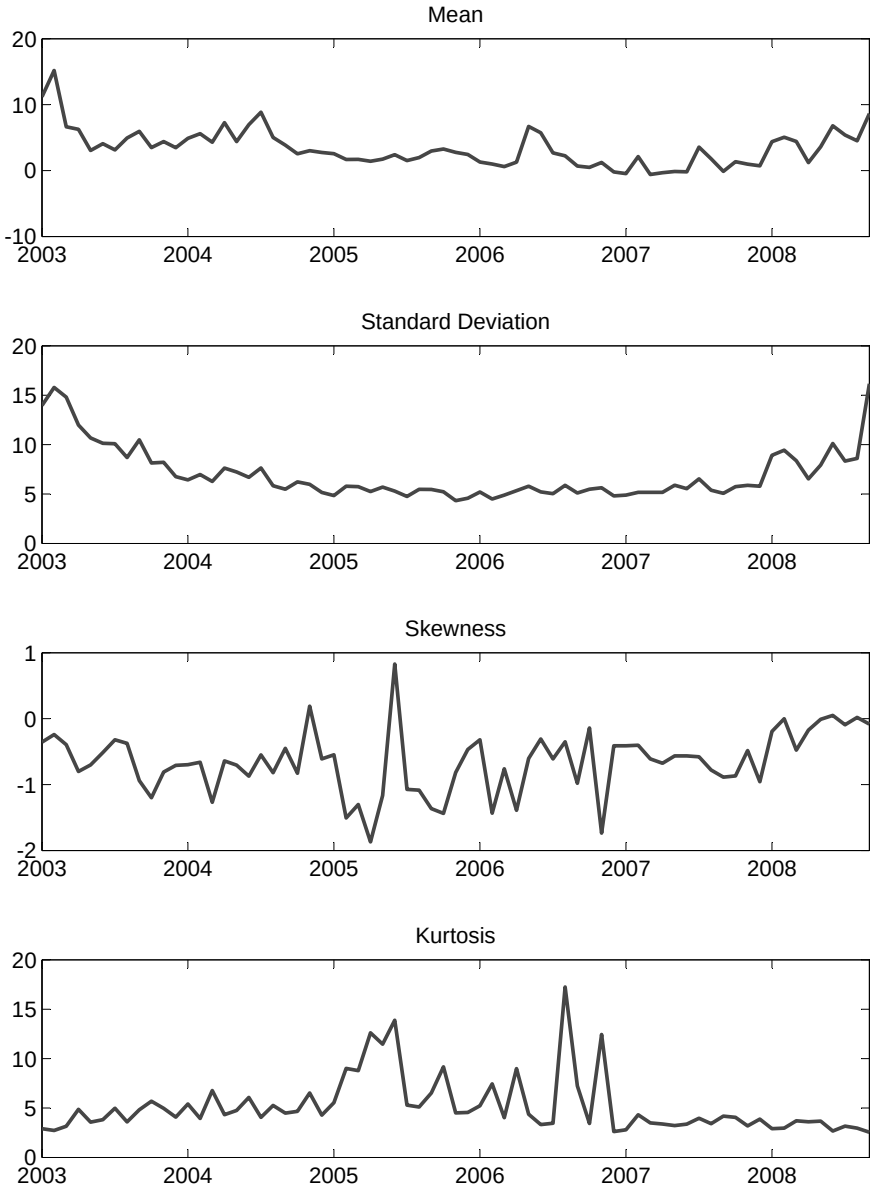
Pesaran, M.H. (1987), The Limits to Rational Expectations, Oxford: Basil Blackwell.

Pesaran, H.M. and M. Weale (2006), Survey Expectations, in: G. Elliott, C.W.J. Granger and A. Timmermann (eds.) Handbook of Economic Forecasting, Vol. 1, North Holland, 715–776.

Seitz, H. (1988), The Estimation of Inflation Forecasts from Business Survey Data, Applied Economics, 20, 427–438.

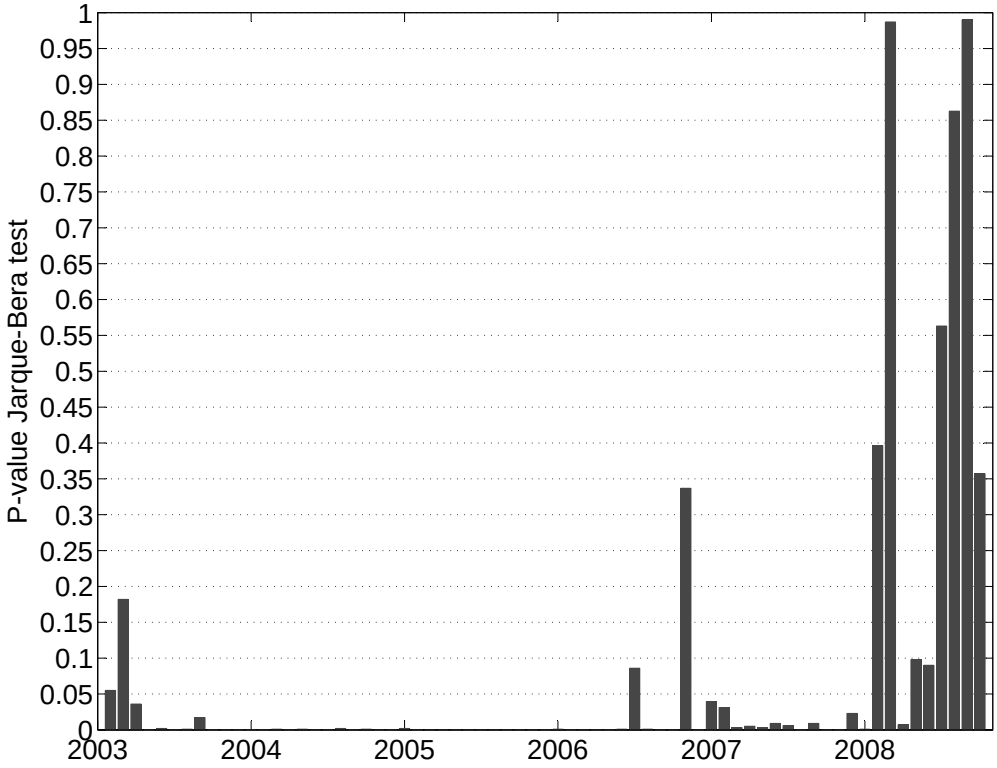
Theil, H. (1952), On the Time Shape of Economic Microvariables and the Munich Business Test, Revue de lInstitute International de Statistique 20.

Figure 1: Cross-sectional moments of return expectations over time



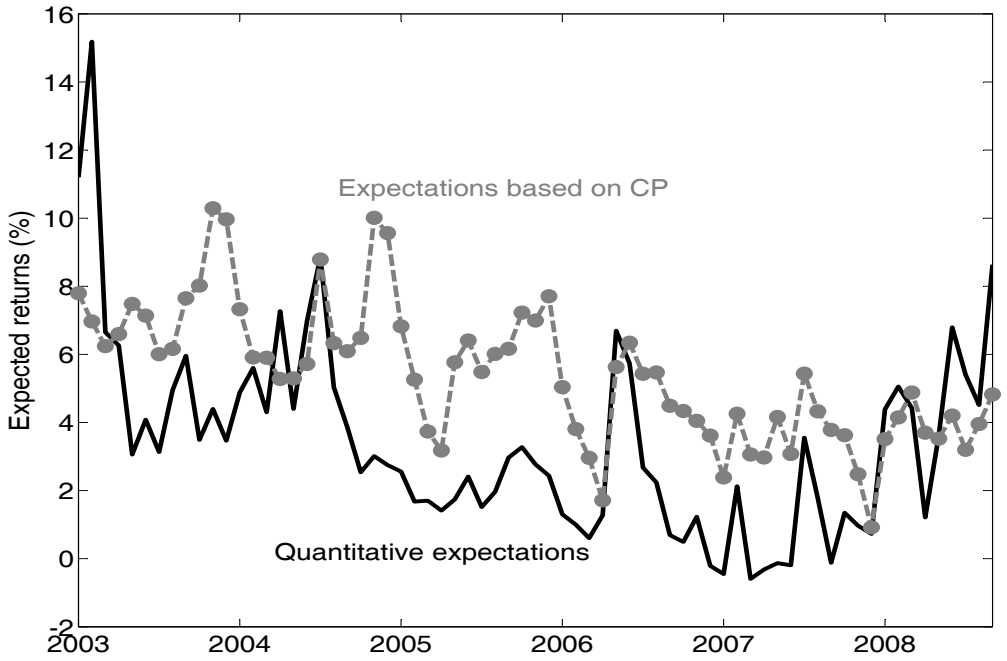
Notes: This figure shows time-series plots of the cross-sectional mean, standard deviation, skewness, and kurtosis of individual return expectations. The sample period is February 2003 – October 2008.

Figure 2: Tests for normality



Notes: This figure shows time-series plots of Jarque-Bera tests for normality of return expectations for each cross-section of forecasts in our sample. The sample period is February 2003 – October 2008.

Figure 3: Quantitative expectations versus quantified expectations



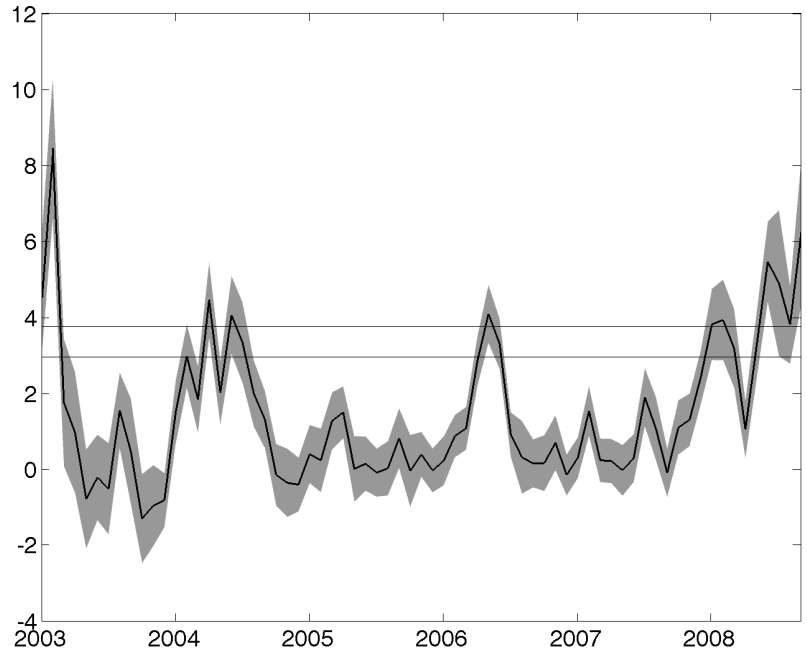
Notes: This figure depicts the quantified return expectations based on the method of Carlson and Parkin (solid line) and average quantitative expectations from the survey of point forecasts (dashed line, circles). Expected returns are in percent.

Table 1: Parameter estimates

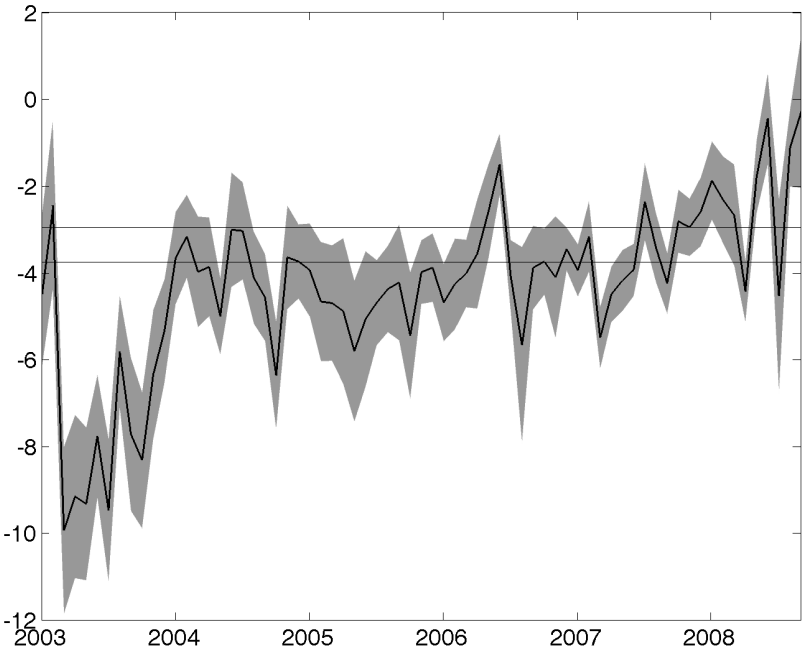
	Dependent: y_t					Dependent: $\widehat{\mu}_t$			
	CP	LS	ALS	PRA		CP	LS	ALS	PRA
δ	2.96	3.76			δ	1.96	1.96		
	[1.41]	[2.13]				[6.85]	[6.81]		
δ^+			-0.58		δ^+			1.78	
			[-0.10]					[3.37]	
δ^-			-18.70		δ^-			-3.34	
			[-1.24]					[-1.81]	
α				0.29	α				0.07
				[4.63]					[5.87]
β				-0.86	β				-0.07
				[-2.70]					[-1.37]
$\overline{R}^2$	0.09	0.09	0.11	0.23	$\overline{R}^2$	0.20	0.20	0.29	0.20

Notes: This table reports parameter estimates from different quantification procedures. The left panel of the table employs future returns (y_t) as dependent variable whereas the right panel of the table employs actual quantitative expectations ($\widehat{\mu}_t$). CP is the estimator originally suggested by Carlson and Parkin (1975) where the threshold is estimated from $y_t = \delta z_t + u_t$ using only a vector of ones as instruments (see (4)). The LS estimator is also based on on the regression $y_t = \delta z_t + u_t$ but uses z_t as instrument. The asymmetric LS estimator (ALS) of Berk (1999) is based on the LS regression (6), and the PRA denotes the regression approach of Pesaran (1985).

Figure 4: Parametric threshold parameters with point-wise 95% confidence intervals



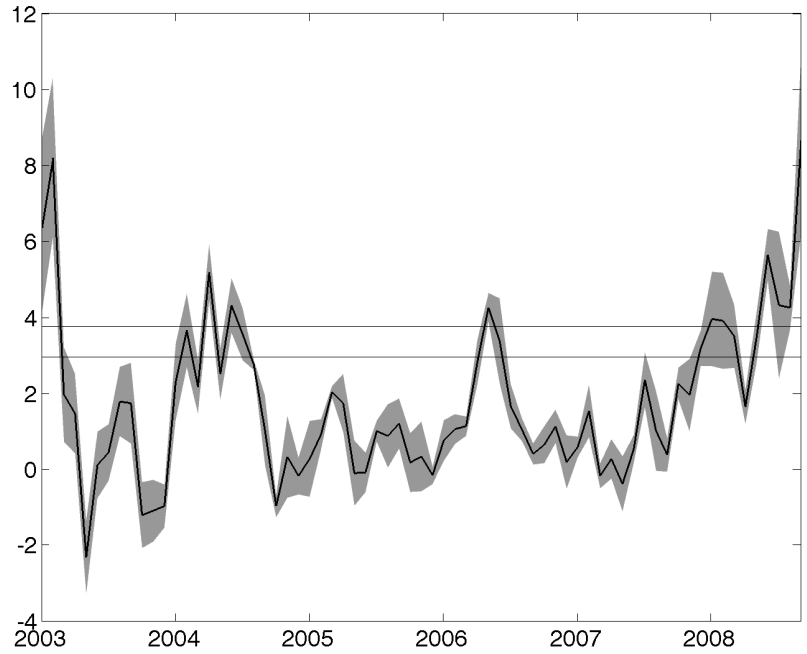
(a) Parametric δ^+



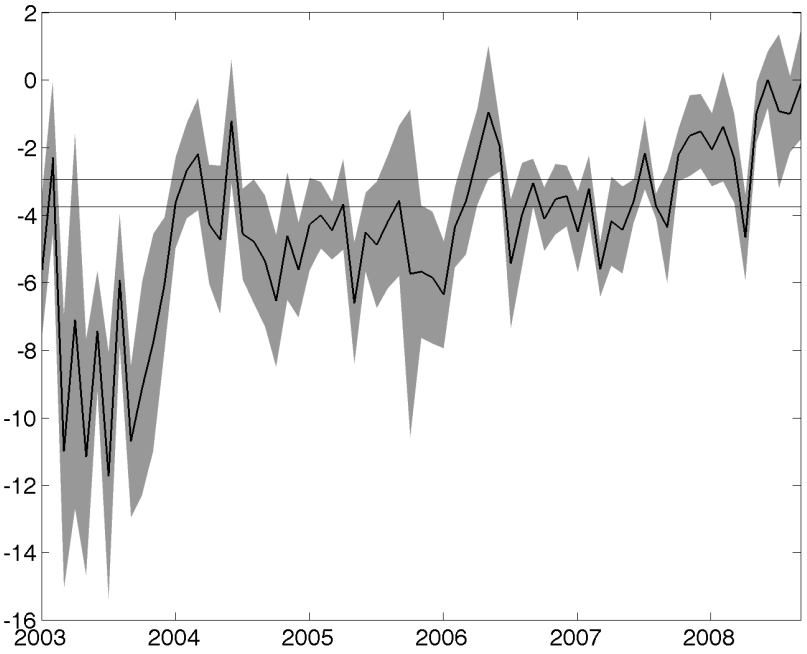
(b) Parametric δ^-

Notes: The figure presents parametric upper thresholds ($\widehat{\delta}_t^+$, Panel (a)) and lower thresholds ($\widehat{\delta}_t^-$, Panel (b)) along with 95% point-wise confidence intervals. Time-invariant thresholds from the CP and LS method are also shown as horizontal lines. Confidence intervals are based on a bootstrap with 10,000 replications.

Figure 5: Non-parametric threshold parameters with point-wise 95% confidence intervals



(a) Non-parametric δ^+



(b) Non-parametric δ^-

Notes: The plot shows non-parametric upper thresholds ($\widetilde{\delta}_t^+$) and lower thresholds ($\widetilde{\delta}_t^-$) along with 95% point-wise confidence intervals. Time-invariant thresholds from the CP and LS method are also shown as horizontal lines. Confidence intervals are based on a bootstrap with 10,000 replications.

Table 2: Explaining thresholds

	dependent variable: $\widetilde{\delta}_t^+$				dependent variable: $\widetilde{\delta}_t^-$			
const.	0.64	1.08	1.03	0.99	-1.83	-1.93	-1.93	-2.01
	[2.26]	[6.63]	[4.47]	[4.58]	[-3.56]	[-4.61]	[-4.75]	[-4.38]
$\widetilde{\delta}_{t-1}$	0.55	0.57	0.51	0.49	0.55	0.42	0.42	0.02
	[5.95]	[8.45]	[8.51]	[7.42]	[5.96]	[4.44]	[4.31]	[0.11]
σ_{t-1}	0.11	0.03	0.00	0.00	-0.02	-0.09	-0.13	-0.05
	[1.54]	[1.65]	[0.13]	[-0.13]	[-0.29]	[-1.69]	[-2.26]	[-0.88]
$\sigma_{t-1:t-6}$			0.04	0.04			0.02	0.05
			[3.64]	[2.86]			[1.14]	[2.73]
$\widehat{\sigma}_t^{GARCH}$				0.00				-0.06
				[1.24]				[-3.42]
r_{t-1}		-0.15	-0.12	-0.13		-0.16	-0.13	-0.08
		[-11.04]	[-7.37]	[-6.76]		[-3.62]	[-3.18]	[-2.66]
$r_{t-1:t-6}$			-0.03	-0.03			-0.03	-0.07
			[-4.06]	[-3.73]			[-2.00]	[-3.68]
$adj.R^2$	0.45	0.76	0.78	0.78	0.29	0.43	0.43	0.59

Notes: This table shows regressions results of (nonparametrically) estimated threshold parameters ($\widetilde{\delta}_t^+$, $\widetilde{\delta}_t^-$) on various determinants. $\widetilde{\delta}_{t-1}$ denotes the lagged threshold parameter depending on the dependent variable, r_{t-1} ($r_{t-1:t-6}$) denotes the lagged market return over the previous month (previous six months), and σ_{t-1} ($\sigma_{t-1:t-6}$) denotes the lagged standard deviations of market volatility over the previous month (previous six months). $\widehat{\sigma}_t^{GARCH}$ denotes the fitted volatility forecast from a GARCH(1,1)-model for monthly DAX returns. t-statistics in squared brackets are based on Newey-West HAC standard errors.

Table 3: Simulation results: Time-varying thresholds

Panel A: Threshold parameter estimates							
Setup	0.5%	2.5%	10%	50%	90%	97.5%	99.5%
I	1.31	2.05	2.44	2.78	2.95	3.03	3.09
II	0.00	1.29	1.85	2.15	2.27	2.31	2.36
III	1.70	1.89	2.01	2.15	2.26	2.31	2.35
IV	1.84	1.90	1.95	2.05	2.13	2.18	2.22
V	1.87	1.92	1.97	2.06	2.14	2.18	2.21
Panel B: R^2 from regressions of μ on z							
Setup	0.5%	2.5%	10%	50%	90%	97.5%	99.5%
I	0.20	0.66	0.84	0.91	0.94	0.95	0.96
II	-1.02	0.44	0.78	0.85	0.90	0.91	0.92
III	0.70	0.77	0.81	0.86	0.90	0.91	0.92
IV	0.08	0.13	0.16	0.23	0.30	0.33	0.36
V	0.10	0.14	0.18	0.25	0.32	0.36	0.39

Notes: This table shows results of the simulation exercises where we generate simulated qualitative expectations from observed quantitative expectations. Panel A reports various percentiles of the distribution of estimated thresholds δ. Panel B reports R^2s from these regression. We run 1,000 simulations for five different simulation setups (I – V) which are described in the text.

Table 4: Simulation results: Individual heterogeneity

	Relative standard deviation										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Setup	Panel A: Threshold parameter estimate (median)										
I	2.78	2.78	2.78	2.77	2.76	2.74	2.72	2.70	2.67	2.64	2.60
II	2.15	2.15	2.15	2.15	2.15	2.14	2.12	2.10	2.08	2.05	2.03
III	2.15	2.15	2.15	2.14	2.12	2.09	2.07	2.03	1.99	1.95	1.90
IV	2.05	2.05	2.04	2.04	2.03	2.01	1.99	1.97	1.94	1.91	1.87
V	2.06	2.06	2.06	2.05	2.04	2.03	2.02	2.00	1.97	1.95	1.92
Setup	Panel B: R^2 from regression of μ on z (median)										
I	0.91	0.91	0.91	0.91	0.90	0.90	0.89	0.88	0.87	0.85	0.84
II	0.85	0.85	0.85	0.85	0.84	0.84	0.83	0.82	0.81	0.80	0.78
III	0.86	0.86	0.86	0.85	0.84	0.83	0.81	0.79	0.76	0.74	0.70
IV	0.23	0.23	0.23	0.23	0.22	0.21	0.20	0.19	0.17	0.15	0.13
V	0.25	0.25	0.25	0.25	0.24	0.24	0.23	0.22	0.21	0.20	0.18

Notes: This table shows results of extended simulation exercises where we follow the simulation design underlying Table 3 but additionally introduce forecaster-specific heterogeneity in thresholds by means of normally distributed random errors (for each forecaster). The standard deviation of these forecaster-specific errors ranges from 0% to 100% of the standard deviation of expected returns (i.e., the "relative standard deviation" ranges from 0.0 to 1.0 as indicated in the first two rows). We run 25,000 repetitions in each simulation setup and for each relative standard deviation and report medians of the 25,000 runs.