Do experts' SKU forecasts improve after feedback?

Rianne Legerstee[1]

Philip Hans Franses

Econometric Institute

Erasmus University Rotterdam

Econometric Institute Report 2011-31

September 22, 2011

Abstract

We analyze the behavior of experts who quote forecasts for monthly SKU-level sales data where we compare data before and after the moment that experts received different kinds of feedback on their behavior. We have data for 21 experts located in as many countries who make SKU-level forecasts for a variety of pharmaceutical products for October 2006 to September 2007. We study the behavior of the experts by comparing their forecasts with those from an automated statistical program, and we report the forecast accuracy over these 12 months. In September 2007 these experts were given feedback on their behavior and they received a training at the headquarters' office, where specific attention was given to the ins and outs of the statistical program. Next, we study the behavior of the experts for the 3 months after the training session, that is, October 2007 to December 2007. Our main conclusion is that in the second period the experts' forecasts deviated less from the statistical forecasts and that their accuracy improved substantially.

Keywords: model forecasts; expert forecasts; judgmental adjustment; outcome feedback; performance feedback; cognitive process feedback; task properties feedback

Authors notes:

We are very grateful to participants of the 31st Annual International Symposium on Forecasting in Prague on June 26-29, 2011 for their helpful comments.

[1] Corresponding author. Econometric Institute, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, Netherlands, e-mail: legerstee@ese.eur.nl

1 Introduction

Much empirical and experimental research is dedicated to the analysis of forecasts from experts who receive statistical model forecasts and then can quote their own forecasts by possibly adjusting the model forecasts. The main focus in this research is usually on forecast quality, that is, do the experts improve or deteriorate forecast accuracy? Theoretically, the experts should be able to improve the statistical model forecasts (see for example Goodwin, 2000) and in some instances they were found to do so (see for example Blattberg and Hoch, 1990; Mathews and Diamantopoulos, 1992; Fildes et al., 2009). Other empirical evidence, however, suggests that too often the experts increased the forecast error (see for example Franses and Legerstee, 2010) and hence, more research is needed to understand what it is that the managers do and how this relates to forecast accuracy (see for example Sanders, 1992; Fildes and Goodwin, 2007; Fildes et al., 2009; Franses and Legerstee, 2009, 2010; Legerstee et al., 2011).

The literature provides many recommendations to experts when they create their forecasts. Examples of such instructions range from making lesser adjustments (Fildes and Goodwin, 2007; Franses and Legerstee, 2009), to making smaller-sized adjustments (Franses and Legerstee, 2010) or, in contrast, making larger-sized adjustments (Fildes et al., 2009; Trapero et al., 2010) and making lesser upward adjustments (Franses and Legerstee, 2009; Fildes et al., 2009). We believe that these instructions are all rather vague, hard to measure and to quantify and sometimes even contradictory. It is therefore questionable to what extent experts are able to improve their forecasts on the basis of those recommendations. For example, would telling the experts to adjust less often upwards result in improved forecast accuracy? On the other hand, quite some research exists on possible forecast improvement as a result of different kinds of feedback on judgmental forecasts, see Lawrence et al. (2006). However, these outcomes are usually based on laboratory experiments and do not include actual forecasters.

In this paper we aim to contribute to the literature by presenting the results of a natural experiment in which the actual experts, who are responsible for final SKU-level sales forecasts, receive information on the model used to create the statistical model forecasts and receive performance and cognitive process feedback. We have the statistical model forecasts, the final forecasts and the realized values of SKU-level sales for the period *before* and *after* the experts received that extra information and feedback. The feedback was based on the discussion and summary statistics presented in Franses and Legerstee (2009, 2010). By collecting the same kind of data for the period after the feedback we now have a unique opportunity to assess the impact of the feedback and model information.

In the next section we discuss the literature on feedback where we focus on forecasts by experts. After that we give the setting of our research and describe the data and the novelty of our research. In the third section we describe the results, where we first present the total results and after that the results per expert. This last part also shows how variations in adjustments based on the feedback result in variations in forecast improvements. The final section concludes with a summary of the main findings and also discusses some limitations that provide challenges for further research.

2 Literature on feedback

Most of the literature on feedback dates back to the eighties and nineties and is based on laboratory experiments, see Lawrence et al. (2006). These authors provide an excellent overview of the literature on judgemental forecasting up to 2006, in which they separately discuss feedback and judgmental adjustments. The experiments consider the effects of different kinds of feedback on the accuracy of forecasts provided by different kinds of students. Focus of study are the forecast accuracy of point forecasts, probability forecasts and judgmental prediction interval forecasts before and after feedback. In our empirical work below we deal with point forecasts and hence we focus on the effect of feedback on such forecasts.

Although the labels and descriptions of the different kinds of feedback vary a little bit across the literature, it seems that a general distinction can be made between outcome feedback, performance feedback and cognitive process feedback. The first simply provides the forecaster with the realized values of the variable for which forecasts were generated. This type of feedback is most common in practice as forecasters are usually able to observe the actual data for the past few periods for which they created forecasts. However, as other types of feedback often show how to improve the forecast accuracy, outcome feedback is typically found to be the least effective, see also Lawrence et al. (2006). As stated in Goodwin and Fildes (1999, p. 41) and Lawrence et al. (2006, p. 507), forecasters seem to be unable to filter the noise component from the realized values of the variable to be predicted and to assess systematic inadequacy in their forecasts.

The second type of feedback is performance feedback and it provides the forecaster with information on forecast accuracy with statistics such as the root mean squared prediction error. Remus et al. (1996) did not find evidence in their laboratory experiment that performance feedback improves forecasting practices as compared to outcome feedback. In contrast, pertaining to judgmental interval predictions and pertaining to probability forecasts, Bolger and Önkal-Atay (2004) and Stone and Opel (2000) do find that performance feedback improves the forecasts. Furthermore, in the *Principles of Forecasting Handbook* (Armstrong, 2001; Armstrong and Pagell, 2003) two of the principles, identified by 40 international studies to increase forecast quality, state that forecasting methods should be compared on their past performance and feedback on forecasts should be sought (see also www.forecastingprinciples.com). Interestingly, Fildes et al. (2009) and Gönül et al. (2009) find in their surveys that these principles are not often followed. In fact, of the respondents only 75% and 35.5%, respectively, indicated to use performance feedback. Gönül et al. (2009) further investigate the reasons for adjusting externally acquired financial and economic forecasts. Getting performance feedback on the external forecasts is shown to result in more adjustments and in lesser reliance on other factors to determine whether to adjust (such as information on the source of the forecasts).

Another study on performance feedback worth noting is Athanasopoulos and Hyndman (2011). To our knowledge, this recent study on feedback is the only study that is not based on a laboratory experiment or a survey, as it is based on an online forecasting competition. This study shows that performance feedback significantly improves forecasting accuracy, although the setting is a bit different from most actual situations. After submission of the forecasts, the forecasters get performance feedback based on a random unknown portion of the forecasts and are able to resubmit a new set of forecasts. This is different from most laboratory settings and real-world situations in which a forecaster gives a forecast for time t, receives feedback on it at $t + 1$ and can then give a forecast for time $t + 1$.

The third kind of feedback is cognitive process feedback and it gives the forecaster information on his own forecasting practices. Such information can include how the forecaster reacts to certain cues or the behavior needed to improve the forecasts. Remus et al. (1996) found no evidence that cognitive process feedback might be helpful (in addition to task properties feedback, see below for a description) and this is a confirmation of the results found by Balzer et al. (1992) pertaining to probability forecasts. Lim et al. (2005) showed that the effectiveness of this type of feedback might be improved by the way the feedback is presented, that is multimedia messages might be more effective than textual messages.

A separate kind of feedback is task properties feedback, which is sometimes also called environmental feedback. It involves providing the forecaster with statistical information on the variable to be forecasted. It can encompass data characteristics or statistical model forecasts. Note that it might be argued that this is not genuine feedback as it is provided before the judgemental forecast is given and it is not feedback on the performance of the judgmental forecaster, see Björkman (1972). This task properties feedback has received most attention in research on feedback on judgmental forecasting, see Remus et al. (1996), Sanders (1992), Welch et al. (1998) and Goodwin and Fildes (1999). In all cases it is found to improve forecast accuracy and in general it is found to be the most effective form of feedback (Lawrence et al., 2006).

Forecasters usually receive a statistical model forecast before stating their own judgmental forecasts in the case of judgmental adjustments, and as such this can be viewed as task properties feedback. Goodwin and Fildes (1999) investigate if providing a statistical model forecast improves forecast accuracy and also if providing additional information on the statistical forecasts helps to further improve the judgmental forecasts. Both seems to be the case, although two remarks can be made. First, the statistical model forecasts appear not to be used efficiently, as is confirmed in Franses and Legerstee (2010). Second, providing information for trend-seasonal series did not improve the forecasting results, possibly due to problems of the subjects to comprehend this information.

In the next sections we study how various types of feedback can lead to different forecasts, where we study actual experts and actual feedback in a natural experiment. Although the results from previous research are sometimes contradictory, we expect that feedback in general results in more accurate judgmental forecasts. How the experts are expected to change their behavior in order to achieve higher accuracy is discussed in the next section.

3 Setting

The natural experiment that we present is based on SKU-level sales data from a large pharmaceutical company. The data concern forecasts for monthly sales of pharmaceutical products in many countries and for various horizons. Final forecasts EF are delivered by experts who first receive statistical model forecasts MF created using (a version of) ForecastPro. The performance of the experts is assessed by their forecasting accuracy and part of their bonus depends on it. In Franses and Legerstee (2009, 2010) the behavior and effectiveness (in terms of accuracy) of the experts is analyzed. The analysis was performed using about two years of monthly data, covering September 2004 to September 2006.

The main conclusion from these studies is that the managers responsible for creating the final forecasts deviate too much from the statistical model forecasts. It is found that the difference between EF and MF is predictable, while it should not be, and that MF receives too small a weight in the final EF forecasts. The experts make frequent adjustments and these tend to be upwards. As a result the expert forecasts are either equally accurate as the model forecasts or much less accurate. When $EF - MF$ increases, that is, when the size of the upward adjustment becomes larger, it is found that forecast performance is deteriorated.

In August-September of 2007 the managers (experts) responsible for forecasting received feedback by way of a presentation at the headquarters' office. They received three kinds of feedback. First of all they received cognitive process feedback, as statistics were presented to the managers on their behavior in adjusting the model forecasts. Second, they received performance feedback, in the form of accuracy measures of their forecasts. Finally, they received more information and explanation on the statistical models used to create the forecasts. So, although they already received task properties feedback in the form of the statistical model forecasts, this type of feedback is extended by the extra information given at the headquarters' office. We have benchmark observations for the period in which the managers received outcome and simple task properties feedback. We also have new forecasts for the period after the presentation, in which the experts received cognitive process feedback, performance feedback and additional task properties feedback. We are now interested in studying the behavior and performance of the experts *before* and *after* the feedback session.

We use a data set that contains forecasts created in September 2006 to December 2007. In 2008 the pharmaceutical company was acquired by another company and many managers who were responsible for the forecasts left. So, data after December 2007 cannot be used for our purposes. We restrict our focus to 1-step-ahead forecasts and only the observations for products for which forecasts and realizations are available for all 16 ($t = 1, .., 16$) months are retained. We compare the data for September 2006 to September 2007 (first sample, 8411 observations) with the data for October 2007 to December 2007 (second sample, 1941 observations). The final forecasts are created by 21 managers located in as many countries.

We first address the behavior of experts. We consider judgemental adjustment, defined as

$$Adj_{i,t} = EF_{i,t} - MF_{i,t}, \tag{1}$$

and relative adjustment, defined as

$$AdjR_{i,t} = (EF_{i,t} - MF_{i,t})/MF_{i,t}, \tag{2}$$

where $EF_{i,t}$ is the SKU-level expert forecast created in month t for month $t+1$ and for product i. We present in the next section various statistics of these (relative) forecast adjustments for the periods before and after the feedback session and we test if any differences between these statistics are significant.

The second issue is whether any changes in behavior lead to changes of forecast accuracy. For that purpose we use the difference between absolute forecast errors of the expert forecasts and absolute forecast errors of the model forecasts, defined by

$$Err_{i,t} = |R_{i,t+1} - EF_{i,t}| - |R_{i,t+1} - MF_{i,t}|, \tag{3}$$

where $R_{i,t+1}$ is the realization of SKU-level sales of product i in month $t+1$ corresponding to the forecasts created the month before. Furthermore, we look at the relative difference in absolute forecast error, that is,

$$ErrR_{i,t} = (|R_{i,t+1} - EF_{i,t}| - |R_{i,t+1} - MF_{i,t}|)/R_{i,t+1}. \tag{4}$$

For this variable we also present various statistics and test results to see if before and after the feedback session performance has changed.

To compare the statistics in both samples we use the common large-sample test as described in Wackerly et al. (2002a).

4 Results

In this next section we first analyze the statistics and test results as they are computed for all the experts together. After that we consider the same statistics and test results but then computed per expert to see if there are significant differences and to see how possible changes in behavior of the experts influences forecast accuracy.

4.1 All experts

Behavior

In Table 1 we present statistics and test results for expert adjustments Adj (see equation (1)) and relative expert adjustments $AdjR$ (see equation (2)). The first observation that is noticeable is that there is a large and significant difference between average adjustments before the experts received feedback and after that session. As we also immediately see from the table that there is not a significant difference (p-value is 0.478) between average absolute adjustments, we might conclude that the experts make more negative adjustments in the second sample and that this causes the difference in average adjustments.

[INSERT TABLE 1 ABOUT HERE]

We also see that the standard deviation of adjustments and absolute adjustments is much smaller in the second sample. The averages of relative adjustments and absolute relative adjustments also get significantly smaller. This indicates that the forecast adjustments have smaller variation in the second sample than in the first and that the adjustments in the second sample are on average relatively smaller than the adjustments in the first sample.[2]

[2] We used the variance test as described in Wackerly et al. (2002b) to test if the difference between the variances is significant for the variables Adj, $AdjR$, Err and $ErrR$. Test results showed highly significant differences for all four variables. However, this test requires that the variable for which the variance is being tested is normally distributed and this is never the case. Therefore, test results are not reliable and omitted.

[INSERT TABLE 2 ABOUT HERE]

To get more insights, we also consider the fraction of zero-adjustments, that is, how often is the model forecast unadjusted anyway? Table 2 shows that this fraction is less than 0.003 in the first sample, while in the second sample this fraction increased significantly to 0.016. Thus the feedback the managers received made them to adjust less often, although the model forecasts are still adjusted very frequently. Second, we see a significant decline in positive adjustments of 3.5% (from 0.571 to 0536) and a little bit smaller but also significant increase in negative adjustments of 2.2% (see Table 2 again). Hence, there is indeed a shift from positive adjustments to no adjustments and negative adjustments. Considering the fact that before the feedback at the headquarters' office the percentage of positive adjustments was around 57.1 and the percentage of negative adjustments was around 42.6, this results in more balance between positive and negative adjustments in the second sample, although there is still a clear difference between the two.

Is the decline in mean of adjustments, absolute adjustments, relative adjustments and absolute relative adjustments completely due to the increase in the number of no adjustments or are the adjustments that are made in the second sample also smaller than before? To answer that question we calculate these four statistics while leaving out the zero-adjustment observations and we test if they differ significantly across the two samples. In the second panel of Table 1 we see that the mean of adjustments, relative adjustments and absolute relative adjustments still decline significantly or almost significantly (largest p-value is 0.1) once we leave out the zero-adjustment observations. Hence, the adjustments that are made after the feedback are relatively smaller than before, but not in an absolute sense as the mean of absolute adjustments increases slightly.

If we take a closer look at the distribution of $AdjR$ before and after the feedback we see that the differences exist mainly in the relatively small adjustments (Table 2 fifth row from below). However, we also see a significant decline of approximately

1% in the amount of extremely large positive relative adjustments (larger than 100% of the size of the model forecast). Furthermore, note that the number of large forecast adjustments (larger than 25% of the size of the model forecast, but smaller than 100% of the size of the model forecast) did not change significantly, both for negative and positive adjustments.

From Tables 1 and 2 we can conclude that the experts truly incorporated the feedback as they changed their forecasting behavior. They adjust less often and the adjustments that they still do are relatively smaller on average and there is more balance between positive and negative adjustments. However, one may feel that there is still room for improvement as adjustments still happen more often upward than downward and also around 25% of the forecasts still are associated with large or extremely large adjustments (larger than 25% of the size of the model forecast).

Forecast accuracy

In Table 3 we present statistics and test results for the differences in absolute forecast error Err (see equation (3)) and the relative differences in absolute forecast error $ErrR$ (see equation (4)). The first row shows a promising result. Where the experts perform worse than the model forecasts, after feedback the forecast accuracy increases substantially (p-value of 0.066). We furthermore see a decrease in the standard deviation of the differences in absolute forecast error. So not only is the difference on average lower, there is also less variation in the differences.

[INSERT TABLE 3 ABOUT HERE]

The average relative difference $ErrR$ also decreases, from EF being 1.8% of R less accurate than MF, to EF being 0.2% of R less accurate than MF, see the bottom two rows of Table 3. However, this improvement is not significant. Thus the forecast improvement as measured by the mean of Err is mainly achieved by improvements that are small relative to R.

[INSERT TABLE 4 ABOUT HERE]

In Table 4 we observe that the fraction of positive Err and $ErrR$, which is the fraction of forecasts where the managers deteriorate forecast accuracy, is lower in the second sample as compared to the first, with a p-value of 0.069. We also see that the fraction of negative Err and $ErrR$ increases after the feedback, but that this increase is not significant. As might be expected from the significant increase in no adjustments, we also find a significant increase in the number of forecasts with no difference in forecast accuracy between EF and MF.

In the remainder of Table 4 we see that the change in distribution of $ErrR$ largely follows the change in distribution of $AdjR$ in Table 2. The number of large deteriorations increases slightly (0.010 to 0.013) and the number of large improvements decreases slightly, possibly as a result of the slightly more large adjustments. The number of small deteriorations decreases significantly, as does the number of small positive adjustments. Finally, the number of forecasts with no difference and with small improvements in forecast accuracy increases, although the last one not significantly.

In sum, we can conclude that there is a small but significant improvement in forecast accuracy of EF over MF after feedback, and it seems to be related to the way the adjustments changed. It seems that the managers have partly changed their behavior as predicted by previous research and as a result of new feedback. Also, the changes have resulted in the expected improvements in forecast accuracy. There does seem to be room for further improvement though.

In the next subsection we analyze the behavior and forecast accuracy across the 21 managers, to see if there exist large differences across the managers and whether these differences result in different forecast accuracy.

4.2 21 experts

In our data set there are 21 managers producing final forecasts, where each manager is responsible for the forecasts in a specific country. In this section we focus on relative forecast adjustment (equation (2)) and relative difference in absolute forecast error (equation (4)) for each expert separately. We only look at $AdjR$ and $ErrR$, because the forecasts and sales figures substantially differ in size across the countries, so a comparison of Adj and Err is hard in this case.

Changes in behavior and accuracy

First we look at the distribution of some statistics concerning $AdjR$ and $ErrR$, see Table 5. In the first row of this table we see for the first sample that the mean of $AdjR$ per manager ranges between 0.008 and 0.921, so it is always positive. In the second sample this mean ranges between -0.078 and 0.314, so both minimum and maximum are lower than in the first sample, and hence the distribution has shifted. If we consider the 21 differences (one for each manager) between the mean of the second sample and the mean of the first sample, we see that these differences range between -0.812 and 0.096 with 14 of these differences being negative (see last column). Clearly, two-thirds of the managers decreased their average relative adjustments as a result of the received feedback.

[INSERT TABLE 5 ABOUT HERE]

The standard deviation of the mean of $AdjR$ has also decreased quite significantly, see the second row of Table 5. The maximum standard deviation changed from 9.309 to 0.934 and 15 of the 21 managers decreased the variation of their relative adjustments.

For absolute $AdjR$ we see the same patterns, see rows 3 and 4 of Table 5. Although the minimum of the mean and the minimum of the standard deviation of this variable have hardly changed, the maximum of the mean decreased from 1.060 to 0.422 and the maximum of the standard deviation decreased from 9.304 to 0.879, respectively. Furthermore, 14 of the managers decreased their average relative adjustments in absolute sense and 16 of the managers decreased the variation of absolute relative adjustments. Hence, the size of the adjustments is on average lower for 67% of the managers and is less extreme for 76% of the managers.

For the number of zero-adjustments and the number of positive adjustments we do not observe many changes. There are 8 managers who always adjusted before the feedback session at the headquarters' office and who always adjusted after that meeting. Only 4 managers increased the number of no adjustments relative to the number of forecasts. As the increase in no adjustments was significant over the complete group of forecasts (so for all managers together, see previous subsection), we might already suspect that those four managers substantially increased the number of no adjustments. We see indeed that the maximum of the fractions of no adjustments increased from 0.031 to 0.301.

What is most obvious from the fractions of positive adjustments is that the variation of these fractions increased. The minimum decreased (0.307), the maximum increased (0.769). Only 11 of the managers brought the fraction closer to the value of 0.5.

Both the minimum and maximum of the mean of relative differences in forecast accuracy decreased, as we would expect to see (the lower the $ErrR$, the more accurate is the manager as compared to the model). A little bit over 60% of the managers improved their forecasts as compared to the model forecast and relative to the size of the realization. Furthermore, both the minimum and maximum standard deviation of $ErrR$ decreased and 12 managers reduced the variation in $ErrR$. Although we would have expected the same for the fraction of $ErrR$ that is positive, in contrast we see an increase in the maximum of the fractions of $ErrR$ that is larger than zero (bottom row of Table 5). The minimum does decrease however and also 12 managers were able to decrease the fraction.

Relation between behavior and accuracy

We now turn to analyze whether the changes in adjustments relate to any changes in forecast accuracy. Do lesser adjustments and lesser positive adjustments result in better expert forecasts? To answer that question we use a linear regression with as dependent variable the difference between mean $ErrR$ in the second sample versus the mean $ErrR$ in the first sample. As independent variables we use the differences in the mean of $|AdjR|$, the differences in the standard deviation of $AdjR$ and the differences in how close the fraction of positive adjustments were to 0.5. Estimation results based on Ordinary Least Squares appear in Table 6.

[INSERT TABLE 6 ABOUT HERE]

The first observation from these estimation results is that there is a highly significant and positive relation between the change in mean absolute relative forecast adjustment and the change in mean relative difference in forecast accuracy. The more a manager decreased the relative adjustments in absolute sense on average, the more the manager was able to improve average relative forecast accuracy as compared to model forecasts. So indeed, smaller adjustments do better.

The next independent variable shows an interesting result. Ceteris paribus, the more the standard deviation in relative forecast adjustments increased, the more the average relative forecast accuracy as compared to model forecasts improved. This result is significant at a 5% significance-level. Hence, although the adjustments should be smaller in size on average, an increase in variation resulted in a lower $ErrR$. Apparently, managers were better able to identify when and how to adjust, instead of careless adjustments of model forecasts.

The last variable shows the expected sign. If a manager made positive and negative adjustments closer to $50/50$, forecast accuracy improved. This parameter is significantly different from zero at 14.5%, which given the small sample size of only 21

could be considered as significant.

The fact that the first variable has a significantly positive parameter and the second variable a significantly negative parameter also indicates that replacing positive with negative adjustments increases forecast accuracy as compared to making only positive or only negative adjustments. If average adjustment is positive, then the second parameter indicates that adjustments should fluctuate more around that mean. The first parameter indicates that the adjustments should be more close to zero. If the average adjustment is negative, then the second parameter also indicates that adjustments should fluctuate more around that mean, whereas the first parameter indicates that the adjustments should be closer to zero.

5 Conclusions

We analyzed forecast adjustments of experts, before and after giving these experts feedback and we examined if feedback improved subsequent forecast accuracy due to changes in behavior. We answered that question by analyzing the data from an actual natural experiment. In that experiment we considered the adjustments and forecast errors of SKU-level sales data of both before and after the managers who are responsible for the forecasts, received cognitive process feedback, performance feedback and extra information on task property feedback. The cognitive process feedback and performance feedback was based on the empirical results obtained in Franses and Legerstee (2009, 2010).

We clearly observe that the managers changed their forecast adjustment behavior significantly and in directions that could be beneficial. They adjust significantly less frequently, significantly less upwards and significantly more downwards. Furthermore, we have seen that the average of the adjustments, the average of the relative adjustments and the average of the absolute but relative adjustments decreased significantly, while the average of the absolute adjustments did not. This, together with decreased standard errors of adjustments and the changes in distributions of relative adjustments, shows that relatively extremely large positive adjustments and relatively small positive adjustments are replaced by zero-adjustments and relatively small negative adjustments. Even though many changes were significant, we concluded that feedback could have resulted in even larger changes in behavior.

If we look at forecast accuracy, we are optimistic. Average forecast accuracy of the expert compared to that of the model increased significantly and changed from poorer performance to better performance. This increase can be largely ascribed to small improvements relative to the size of the realized sales. This can be seen from the fact that relative forecast accuracy of the expert compared to the model does increase, but not significantly. Next we observe lesser volatility in the forecast errors, significantly less forecast deteriorations by the experts, significantly more no improvements and also more forecast improvements.

We also compared the behavior and accuracy of the 21 different managers separately. We see large differences in the level of adaptation of the managers to the feedback. Furthermore, the way they changed their behavior influences the change in forecast accuracy significantly. Smaller adjustments in an absolute sense and more balance between the amount of positive and negative adjustments clearly increases forecast accuracy. However, ceteris paribus, more variation in the adjustments also improves the forecast accuracy. In sum, we can conclude that cognitive process feedback, performance feedback and extra information on the statistical model used to create the model forecasts results in more accurate expert forecasts than if the forecasters only receive outcome feedback and simple task performance feedback by way of statistical model forecasts.

Our study clearly shows that it is useful to examine what forecasters do and what the results are in terms of forecast accuracy. Presenting this information to the forecasters appears useful in practice. Of course, analyzing more data sets, analyzing forecasts from other companies and other forecasting areas, would result in even more reliable conclusions.

The fact that we analyze a natural experiment is the strength and novelty of this research, but of course also implies some limitations. We were not able to set the experimental design. Hence, we are not able to make a distinction between the individual effects of each of the feedback types, that is, of the information provision on the statistical model, the performance feedback and the cognitive process feedback. We were also not able to have a control group. We hope to have a chance to run a natural experiment again in the future where we can accommodate these limitations.

References

Armstrong, J. (2001). *Principles of Forecasting*. Kluwer Academic Publishers, Boston, MA.

Armstrong, J. and Pagell, R. (2003). The ombudsman: Reaping benefits from management research: Lessons from the forecasting principles project. *Interfaces*, 33:91–97.

Athanasopoulos, G. and Hyndman, R. (2011). The value of feedback in forecasting competitions. *International Journal of Forecasting*, 27:845–849.

Balzer, W., Sulsky, L., Hammer, L., and Sumner, K. (1992). Task information, cognitive information, or functional validity information: Which components of cognitive feedback affect performance? *Organizational Behavior and Human Decision Processes*, 53:35–54.

Björkman, M. (1972). Feedforward and feedback as determiners of knowledge and policy-notes on a neglected issue. *Scandinavian Journal of Psychology*, 13:152–158.

Blattberg, R. and Hoch, S. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, 36:887–899.

Bolger, F. and Önkal-Atay, D. (2004). The effects of feedback on judgmental interval predictions. *International Journal of Forecasting*, 20:29–39.

Fildes, R. and Goodwin, P. (2007). Good and bad judgement in forecasting: lessons from four companies. *Foresight: The International Journal of Applied Forecasting*, 8:5–10.

Fildes, R., Goodwin, P., Lawrence, M., and Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25:3–23.

Franses, P. and Legerstee, R. (2009). Properties of expert adjustments on model-based SKU-level forecasts. *International Journal of Forecasting*, 25:35–47.

Franses, P. and Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality? *Journal of Forecasting*, 29(3):331–340.

Gönül, S., Önkal, D., and Goodwin, P. (2009). Expectations, use and judgmental adjustment of external financial and economic forecasts: An empirical investigation. *International Journal of Forecasting*, 28:19–37.

Goodwin, P. (2000). Improving the voluntary integration of statistical forecasts and judgement. *International Journal of Forecasting*, 16:85–99.

Goodwin, P. and Fildes, R. (1999). Judgmental forecasts of time series affected by special events: Does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making*, 12:37–53.

Lawrence, M., Goodwin, P., O'Connor, M., and Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22:493–518.

Legerstee, R., Franses, P., and Paap, R. (2011). Do experts incorporate statistical model forecasts and should they? Tinbergen Institute Discussion Paper.

Lim, K., O'Connor, M., and Remus, W. (2005). The impact of presentation media on decision making: Does multimedia improve the effectiveness of feedback? *Information and Management*, 42:305–316.

Mathews, B. and Diamantopoulos, A. (1992). Judgmental revision of sales forecasts: The relative performance of judgementally revised versus non revised forecasts. *Journal of Forecasting*, 11:569–576.

Remus, W., O'Connor, M., and Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, 66:22–30.

Sanders, N. (1992). Accuracy of judgemental forecasts: a comparison. *Omega*, 20:353–364.

Stone, E. and Opel, R. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes*, 83:282–309.

Trapero, J., Fildes, R., and Davydenko, A. (2010). Nonlinear identification of judgmental forecasts effects at SKU level. *Journal of Forecasting*, online publication.

Wackerly, D., Mendenhall III, W., and Scheaffer, R. (2002a). *Mathematical Statistics with Applications*, chapter 10.3, pages 467–473. Duxbury Advanced Series, 6 edition.

Wackerly, D., Mendenhall III, W., and Scheaffer, R. (2002b). *Mathematical Statistics with Applications*, chapter 10.9, pages 498–505. Duxbury Advanced Series, 6 edition.

Welch, E., Bretschneider, S., and Rohrbaugh, J. (1998). Accuracy of judgmental extrapolation of time series data. characteristics, causes, and remediation strategies for forecasting. *International Journal of Forecasting*, 14:95–110.

Tables

Table 1: Summary statistics for forecast adjustments Adj (see equation (1)) and relative forecast adjustments $AdjR$ (see equation (2)). The second column shows the statistics for the first sample (September 2006 - September 2007) and the third column those for the second sample (October 2007 - December 2007). The final column gives one-sided p-values of the test for the difference between the statistics in the two samples (if available).

	Sample 1	**Sample 2**	**p-value**
Mean Adj	212.000	44.041	0.006
Std Adj	2832.890	2634.060	.
Mean $\|Adj\|$	940.431	936.987	0.478
Std $\|Adj\|$	2680.616	2462.075	.
Mean $AdjR$	0.154	0.070	0.003
Std $AdjR$	2.699	0.399	.
Mean $\|AdjR\|$	0.275	0.202	0.008
Std $\|AdjR\|$	2.689	0.351	.
Mean $Adj(Adj \neq 0)$	212.606	44.756	0.007
Std $Adj(Adj \neq 0)$	2836.919	2655.355	.
Mean $\|Adj(Adj \neq 0)\|$	943.122	952.195	0.444
Std $\|Adj(Adj \neq 0)\|$	2683.976	2479.065	.
Mean $AdjR(Adj \neq 0)$	0.155	0.071	0.004
Std $AdjR(Adj \neq 0)$	2.703	0.402	.
Mean $\|AdjR(Adj \neq 0)\|$	0.276	0.205	0.010
Std $\|AdjR(Adj \neq 0)\|$	2.693	0.353	.

Table 2: Distribution of relative forecast adjustments $AdjR$ (see equation (2)). The second column shows the distribution of the first sample and the third column the distribution of the second sample. The final column gives one-sided p-values of the test for the difference between the fractions in the two samples. $I[\cdot]$ is an indicator function which takes a value of 1 if the statement between brackets is true and is 0 otherwise.

	Sample 1	**Sample 2**	**p-value**
Mean $(AdjR > 0)$	0.571	0.536	0.003
Mean $(AdjR < 0)$	0.426	0.448	0.040
Mean $I[AdjR < -1]$	0.000	0.000	1.000
Mean $I[-1 \leq AdjR < -0.75]$	0.003	0.003	0.466
Mean $I[-0.75 \leq AdjR < -0.5]$	0.010	0.011	0.258
Mean $I[-0.5 \leq AdjR < -0.25]$	0.056	0.060	0.218
Mean $I[-0.25 \leq AdjR < 0]$	0.357	0.373	0.100
Mean $I[AdjR = 0]$	0.003	0.016	0.000
Mean $I[0 < AdjR < 0.25]$	0.404	0.366	0.001
Mean $I[0.25 \leq AdjR < 0.5]$	0.098	0.107	0.130
Mean $I[0.5 \leq AdjR < 0.75]$	0.032	0.032	0.482
Mean $I[0.75 \leq AdjR < 1]$	0.011	0.015	0.119
Mean $I[AdjR \geq 1]$	0.026	0.016	0.003

Table 3: Summery statistics for the differences between absolute forecast errors of the expert and absolute forecast errors of the model forecast, Err (see equation (3)), and the relative differences in absolute forecast error $ErrR$ (see equation (4)). The second column shows the statistics for the first sample and the third column shows the statistics for the second sample. The final column gives one-sided p-values of the test for the difference between the statistics in the two samples (if available).

	Sample 1	**Sample 2**	**p-value**
Mean Err	9.652	-69.891	0.066
Std Err	2218.144	2066.510	.
Mean $ErrR$	0.018	0.002	0.289
Std $ErrR$	1.304	1.028	.

Table 4: Distribution of the relative differences in absolute forecast error $ErrR$ (see equation (4)). The second column shows the distribution of the first sample and the third column shows the distribution of the second sample. The final column gives one-sided p-values of the test to see if there is a difference between the fractions in the previous two columns. $I[\cdot]$ is an indicator function which takes a value of 1 if the statement between brackets is true and is 0 otherwise.

	Sample 1	**Sample 2**	**p-value**
Mean $(ErrR > 0)$	0.502	0.483	0.069
Mean $(ErrR < 0)$	0.494	0.499	0.349
Mean $I[ErrR < -1]$	0.010	0.013	0.141
Mean $I[-1 \leq ErrR < -0.75]$	0.006	0.004	0.063
Mean $I[-0.75 \leq ErrR < -0.5]$	0.015	0.012	0.114
Mean $I[-0.5 \leq ErrR < -0.25]$	0.055	0.059	0.239
Mean $I[-0.25 \leq ErrR < 0]$	0.408	0.411	0.390
Mean $I[ErrR = 0]$	0.004	0.018	0.000
Mean $I[0 < ErrR < 0.25]$	0.409	0.374	0.002
Mean $I[0.25 \leq ErrR < 0.5]$	0.057	0.066	0.084
Mean $I[0.5 \leq ErrR < 0.75]$	0.018	0.023	0.075
Mean $I[0.75 \leq ErrR < 1]$	0.007	0.010	0.069
Mean $I[ErrR \geq 1]$	0.011	0.010	0.350

Table 5: Summary statistics of the distributions of relative adjustment statistics and relative difference in error statistics across 21 managers. Columns 2 and 3 give the minimum and maximum of the statistics calculated over the first sample. Columns 4 and 5 give the minimum and maximum calculated over the second sample. Columns 6 and 7 give the minimum and maximum of the difference between the statistics (statistic second sample minus statistic first sample). The last column shows the number of times that the difference is negative, except for Mean $I[AdjR = 0]$ for which it shows the number of times that the difference is positive and for Mean $I[AdjR > 0]$ for which it shows the number of times that it approaches 0.5.

	Sample 1		**Sample 2**		**Diff.**		**Diff. Opt.**
	min	**max**	**min**	**max**	**min**	**max**	**nr.**
Mean $AdjR$	0.008	0.921	-0.078	0.314	-0.812	0.096	14
Std $AdjR$	0.156	9.309	0.171	0.942	-9.112	0.077	15
Mean $\|AdjR\|$	0.103	1.060	0.103	0.422	-0.790	0.120	14
Std $\|AdjR\|$	0.118	9.304	0.119	0.879	-9.157	0.090	16
Mean $I[AdjR = 0]$	0.000	0.031	0.000	0.301	-0.031	0.299	4
Mean $I[AdjR > 0]$	0.470	0.667	0.307	0.769	-0.189	0.237	11
Mean $ErrR$	-0.125	0.582	-0.388	0.253	-0.603	0.168	13
Std $ErrR$	0.160	5.846	0.115	3.275	-5.218	2.575	12
Mean $(ErrR > 0)$	0.423	0.677	0.323	0.795	-0.253	0.280	12

Table 6: This table shows the estimated parameters with p-values, F-statistic with p-value and R-squared statistic of a linear regression of the difference in mean of relative differences in absolute forecast error (equation (4)) (dependent variable) and the differences in mean of absolute relative forecast adjustment (equation (2)), the differences in standard deviation of relative forecast adjustment and the differences in how close the fraction of positive adjustments is to 0.5. Differences are as measured between the second and first sample of the data and estimation is done for 21 observations using OLS.

Variable	**Coef.**	**Prob.**
Constant	-0.020	0.485
Diff. Mean $\|AdjR\|$	0.968	0.001
Diff. Std $AdjR$	-0.051	0.024
Diff. |Mean $(AdjR > 0) - 0.5$|	0.437	0.145
F-statistic	9.381	0.001
R-squared	0.623	.