Feeding-Back Error Patterns to Stimulate Self-Reflection versus Automated Debiasing of Judgments

Nathalie Balla
KU Eichstätt-Ingolstadt
nballa@ku.de

Thomas Setzer
KU Eichstätt-Ingolstadt
thomas.setzer@ku.de

Felix Schulz
KU Eichstätt-Ingolstadt
felix.schulz@ku.de

Abstract

Automated debiasing, referring to automatic statistical correction of human estimations, can improve accuracy, whereby benefits are limited by cases where experts derive accurate judgments but are then falsely "corrected". We present ongoing work on a feedback-based decision support system that learns a statistical model for correcting identified error patterns observed on judgments of an expert. The model is then mirrored to the expert as feedback to stimulate self-reflection and selective adjustment of further judgments instead of using it for auto-debiasing. Our assumption is that experts are capable to incorporate the feedback wisely when making another judgment to reduce overall error levels and mitigate this false-correction problem. To test the assumption, we present the design and results of a pilot-experiment conducted. Results indicate that subjects indeed use the feedback wisely and selectively to improve their judgments and overall accuracy.

Keywords: decision support system, debiasing, automated debiasing, feedback, self-reflection

1. Introduction

A key question in current information systems research is how to achieve collaborative intelligence, i.e., how to combine the complementary strengths of machines, which are stronger in extracting regular patterns from data, and humans, which are more adept at considering novel or transferable situations, effects or unseen developments based on domain knowledge and intuition (Blattberg and Hoch, 1990; Nagar and Malone, 2011; Zellner et al., 2021).

We introduce a novel decision support system (DSS) aimed at improving estimation accuracy by fostering collaborative intelligence. The mechanism implemented by the DSS is to feed-back machine-learned personalized error patterns (biases) of an expert to that same expert who then decides how to incorporate that feedback into her or his further judgments.

Accuracy of estimations is vital for enterprises since planning and decision making usually depend on accurate estimations of (future) business figures. As of today, many respective tasks are dominated by judgmental approaches, i.e., by humans with individual backgrounds, attitudes, and estimation heuristics (Klassen and Flores, 2001; McCarthy et al., 2006; Sanders and Manrodt, 2003). A typical DSS supports such tasks by gathering, filtering, and presenting relevant information to derive informed and unbiased judgments.

However, providing additional information does not have an unambiguously positive effect on accuracy and while a huge body of work on DSSs has been published on how to integrate, aggregate, and visualize data to derive accurate estimations and beneficial decision alternatives, empirical evidence shows that the judgments derived by seemingly well-configured DSSs still come out flawed, including biases like overconfidence, mean or regression bias, optimism, over-steering or anchoring (see, for instance, the findings in Blanc and Setzer, 2016; Lawrence et al., 2006; Lawrence and O'Connor, 1993; Lawrence et al., 2000; Leitner and Leopold-Wildburger, 2011; Lim and O'Connor, 1996).

As a recent example derived from a large corporate dataset, Blanc and Setzer (2015a) analyze a set of empirical cash flow forecasts of a multinational corporation, generated by more than one hundred experts from different subsidiaries using forecast DSSs.

URI: https://hdl.handle.net/10125/102800
978-0-9981331-6-4

The authors find that, nevertheless, mean as well as regression biases exist for all business divisions of the company. Furthermore, they find that the statistically identified error patterns allow for an automated statistical correction of the patterns that increases overall accuracy. The authors also show that the estimated model parameters relate to characteristics of the business environments and argue that these provide valuable insights to better understand, quantify, and feed-back presumed biases to the experts to help them to improve the accuracy of future forecasts.

The same authors also show that, since automated correction is applied to estimates regardless of presumably different confidence in the original estimate, appropriate expert expectations are also corrected in the wrong direction. This leads to higher errors than necessary (Blanc and Setzer, 2015b).

To address this problem, for future research the authors suggest a feedback-based DSS that shows the expert, after she or he submitted a forecast, the forecast of a statistical (correction) model together with a description of the bias that might have driven the discrepancy to the expert's expectation. The authors propose to derive such a benchmark forecast by correcting time persistent biases in past expert forecasts. The expert might then be prompted to accept or overwrite the model forecast, ideally overwriting primarily the model predictions that would lead to heavy false-corrections.

The intuition of providing error pattern based feedback and the key assumption of such an approach that experts are capable to consider the error-feedback wisely and selectively seems compelling. However, this key assumption has, to our knowledge, not been tested so far. For instance, when an estimation task falls in a domain the expert is very familiar with and is sure that the error-feedback is likely not to apply to his or her current judgment, it should be neglected. In cases where an expert is less confident that no structural bias is at play, the feedback might be accepted and the estimation adjusted. Overall, an expert must be capable to make informed decisions if the structural error pattern he or she received is likely to be valid (i.e., whether a bias might indeed be at play).

We present the architecture of a novel DSS together with the design and the results of a first experiment to test this assumption. The DSS addresses the problem that auto-debiasing of experts' judgments leads to decreasing accuracy if the expert made the judgment knowledgeably and accurately, but the model falsely corrects it. The DSS design further aims at providing guidance on how to systematically improve further judgments, i.e., to learn based on errors made in the past. Such a type of DSS may be important for several fields in business, where decision-makers are dependent on the accuracy of estimations and predictions.

The experiment is the first in a series of experiments currently conducted to find evidence for such wise and systematic adjustments after receiving personalized error patterns as feedback, and whether this leads to error reduction. In the experiment, subjects are asked to estimate quantities from different general knowledge categories, while categories are not communicated, and error-feedback in terms of their mean bias (measured as mean percentage error, MPE) is displayed after a sequence of estimations made.

The experiment is designed to make the key assumption described above testable by few sub-assumptions (hypotheses) related to changes of the MPE in the right direction after feedback, whether change is emphasized in categories with higher before-feedback MPE, and whether accuracy improvement is achieved compared to subjects not receiving the feedback, with and without auto-correction of their estimates. Results indicate that subjects indeed seem to use the feedback wisely and selectively to improve judgments.

The rest of this article is organized as follows. In Section 2, we review previous research on auto-debiasing and feedback-based DSS with regard to whether they hint at specific feedback mechanisms promising to enable wise and selective consideration of error-correction feedback. In Section 3, we describe the DSS used as the experimental infrastructure. In Section 4, we present the design and the results of a first experiment that serves as a general proof of concept for the DSS. In Section 6, we discuss the results of our work so far, conclude, and outline future research on error feedback-based DSS.

2. Prior Work on Bias-Related Feedback vs. Auto-Correction

We start reviewing findings with auto-correction, and then review approaches to foster debiasing using feedback. Finally, we discuss their suitability to foster learning, improve judgment accuracy and mitigate the false-correction problem inherent with auto-correction.

As aforementioned, Blanc and Setzer (2015b) discuss accuracy gains through auto-debiasing, referring to the automatic correction of experts' forecasts by a statistical model learned on previous experts' errors. Figure 1 shows the distributions of absolute percentage error (APE) improvements of forecasts when using the corrected forecasts instead of the original expert forecasts per decile of the confidence interval around the

correction model's forecast. The larger the correction,

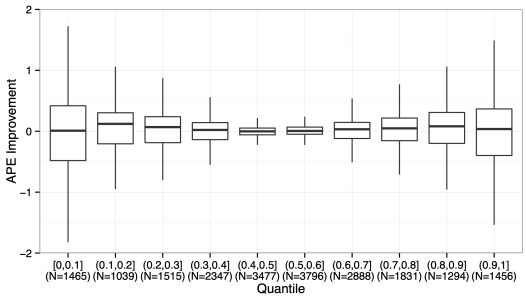


Figure 1. APE distribution by decile in the confidence interval of the auto-corrected forecast Blanc and Setzer, 2015b

the higher the variance of error differences. The most deviating decile bins contain the heaviest accuracy gains and losses.

The authors argue that, as auto-correction is applied to estimates regardless of confidence in the original estimate, originally sound expectations are also corrected in the wrong direction, leading to high errors specifically in outer decile bins. Hence, they suggest to prompt experts to accept or overwrite a model forecast if the expert forecast exceeds certain confidence bounds, as in cases of extreme deviations either a strong bias might be at play or the expert forecast might be based on specific knowledge and indeed be appropriate.

This perspective has merit as experts can be assumed to have higher confidence and knowledge in certain estimation tasks and biases are likely to depend on the type of estimation task. However, whether experts are capable of making wise feedback accept/neglect decisions depends on several factors, where one of particular importance is surely the type of feedback provided. Therefore, we now review feedback-based DSSs and whether they appear promising for the task of making wise error-feedback consideration decisions.

A common distinction of feedback types is outcome feedback (OFB) and cognitive feedback (CFB). OFB refers to "information that describes the accuracy or correctness of the response" (Jacoby et al., 1984, p. 531), and is often solely the correct answer. CFB is "information regarding the how and why that underlies this accuracy" (Jacoby et al., 1984, p. 531).

Regarding outcome feedback, Remus, O'Connor, and Griggs (1996), Balzer, Doherty, and O'Connor (1989) and Lawrence, Goodwin, O'Connor, and Önkal (2006), amongst others, show that OFB in form of providing correct answers is rather ineffective, and many studies question the usefulness of OFB of that type in general (Balzer et al., 1989). It is argued that such information is insufficient to improve judgment. It has even been shown that better performing experts avoid using OFB of that type (Lawrence et al., 2006; Remus et al., 1996).

In contrast, OFB in the form of personalized performance feedback seems more suitable. As an example, Benson and Önkal (1992) studied performance feedback in probability estimation. In their experiment, subjects made four weekly predictions of football games for the following weekend regarding the probability for a team to win. Subjects of the treatment group received performance feedback while control group subjects did not. The authors find that performance feedback helped to increase forecasting accuracy.

Fischer and Harvey (1999) observed that feedback originating from performance on one trial increases motivation of the subject in the next trial. Here, subjects were asked to combine sales forecasts of others, where the treatment group received feedback on their first trial before the their second trial. The feedback showed the own forecast, the actual outcome, and the respective error. The results indicate that such feedback does help to learn and also induces motivation through goal-setting as the feedback functions as a goal to outperform.

Although we do not find studies focusing on selective incorporation of feedback and adjustment of estimations, based on prior research, actionable error-feedback seems to be a promising candidate for our setting.

Concerning cognitive feedback, Sengupta and Abdel-Hamid (1993) published an article in Management Science that presents an experiment integrating CFB in DSSs. 47 subjects performed tasks as project managers in terms of staffing decisions for a software project, which involved trade-offs between cost and time plan. After making decisions, every subject received outcome feedback in terms of a report on the current stage of the project. In the CFB group, CFB was available in form of task information through plots of variables over the project's life span (such as information on the perceived cost and size of the project) and a summary of the past interval. Experimental results show that subjects with access to CFB (in addition to OFB) performed best compared to the group receiving only OFB.

Sengupta (1995) conducted further experiments, where subjects had to conduct personnel screening. The treatment group received OFB as well as CFB whereas the control group received OFB only. OFB was shown as the rating decisions made by the expert commitee and CFB as the committee's decision strategy regarding similar jobs as well as consistency scores and information on a subject's own decision strategy. The findings show that subjects receiving OFB together

with CFB tend to outperform those receiving OFB only. Combining performance with cognitive feedback therefore seems like an approach worthwhile to be pursuit for our purpose.

However, a severe challenge is the acceptance of feedback by an expert in general, as it has been found that experts are usually overconfident in their own expectations even if their ability is shown to be inferior to the estimate provided by software (Leitner and Leopold-Wildburger, 2011). Therefore, a challenge is fostering a self-reflective process, i.e., the interpretation and assessment of own thoughts, emotions, and actions, required for directed change and key to wise decision making (Grant et al., 2002; Sasse-Werhahn et al., 2020).

For instance, in an experiment by Goodwin (2000), prompting forecasters to revise judgmental forecasts after statistical information has been provided did not improve accuracy, whereas asking forecasters to adjust a forecast while requiring reflection by providing a reason for the adjustment performed best. It has also been found that specifically feedback like error-feedback drives reflective processes, which in turn affects if and how the feedback is accepted and used. For example, Sargeant, Mann, van der Vleuten, and Metsemakers (2009) conducted interviews with physicians who evaluated assessment feedback they received. This reflection was useful in terms of how to apply the feedback.

Overall, in search for a promising feedback-type, previous work on feedback, debiasing, and self-reflection encourages the usage of an expert's own error pattern – relating to a potential bias that can be understood and corrected – as performance related feedback type. In addition, it seems suitable to induce self-reflection as it is different to external feedback often adopted insufficiently. Thereby we provide both, promising types of OFB and CFB.

3. Experimental Infrastructure and Procedure

We now introduce the DSS infrastructure used for our experiments from a procedural perspective together with key considerations, while keeping technical details short. We illustrate several components by providing examples of their implementation in the first experiment.

Technically, the DSS is developed as a Web-App using Dynamic HTML (PHP) as frontend, and a Relational Database Management Server (MySQL) as backend containing the parameterization of the experiments, storing outcomes, and used for analyzing the answers and reactions of the subjects. The error pattern derivation, its presentation as feedback as well as the calculation of loss functions are provided by tools written in PHP and *R*.

An overview of the steps supported by the DSS is depicted in Figure 2. First, an experiment is

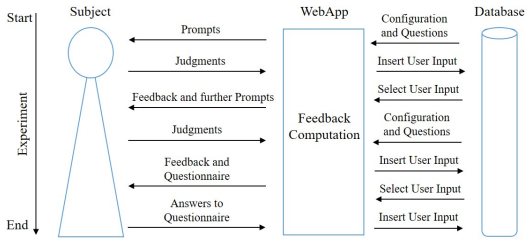


Figure 2. Experimental Infrastructure and Processes

configured using a Web-based tool and the configuration is stored in a database. Configuration items are pages for briefing/debriefing, comprehensibility questions, estimation questions to be answered by treatment and control group subjects, the loss function that measures performance, rules when feedback of what form is provided, texts and visuals provided with a question, rules when an experiment terminates, and a final questionnaire form.

Then, the subjects are randomly assigned to the treatment or control group, shown information on the experiment, asked comprehensibility questions and the experiment itself starts by prompting for judgments. An example prompt is shown in Figure 3. Here, the task is

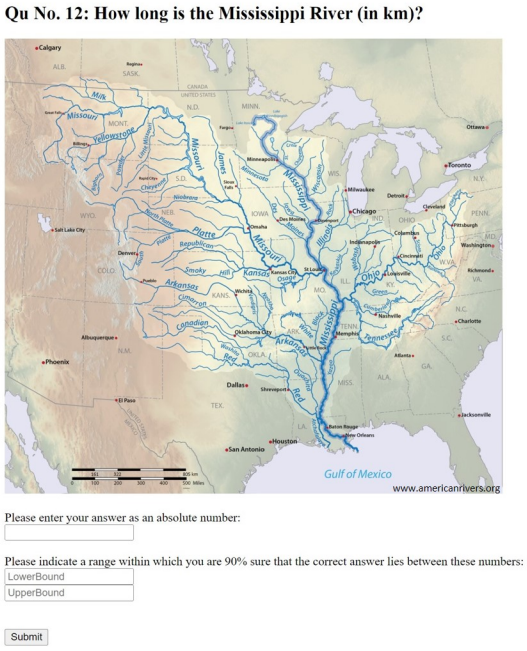


Figure 3. DSS Interface – User Prompt Example

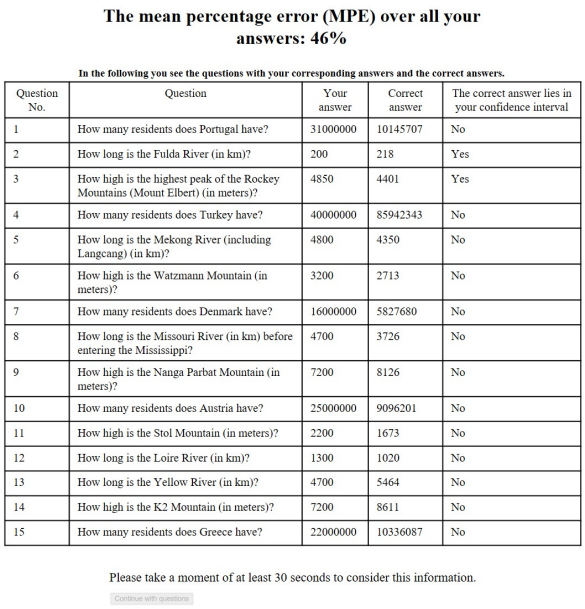
to estimate the length of the Mississippi in kilometers, where guidance is provided by a map and a legend indicating the scale. A subject is also asked to indicate her or his 90% confidence interval – the interval to which a subject is 90% sure that the correct answer lies within. After the answer is submitted, the next prompt is displayed, and after a defined number of questions either feedback is provided for 30 seconds (treatment group) or a blank page is shown inviting to take a 30 second break (control group). In case feedback is given, the subject's error pattern (bias) is computed and displayed together with her or his individual estimation errors.

In our first experiment, the subject's mean bias computed as her or his mean percentage error (MPE) across all the previously given answers, is shown as feedback. As this DSS is meant to demonstrate general functionality as a proof of concept, the MPE is only a simple example of a statistical model. Other models can be used to calculate indications of other biases. MPE is computed as follows: per estimate made, the difference between the estimate and the actual (the actually correct answer) is computed and that difference is divided by the actual and multiplied by 100. MPE is then the mean of these values and therefore also calculated across all categories. MPE is chosen for reasons of comprehensibility and ease of applicability for debiasing. For example, a MPE of 0.5 means that estimates exceed actuals by 50% on average, and correction means to take only $\frac{2}{3}$ of a further estimate. Figure 4 shows an example feedback page with the (potential) mean bias of the subject.

The intention of the feedback is to make a subject aware of a potential mean bias, possibly derived by previously given answers. A potential cognitive bias might be mentally corrected by a subject when providing further novel questions. This may be category-specific, although categories are not mentioned or used by the DSS. Thus, the aim of the feedback is to make subjects reflect on previous error patterns to improve future estimations. Hence, the subjects must make novel estimations applying the generic feedback that is computed across all their previous answers and needs to be cognitively wisely applied.

After the feedback or the blank page, a subject is faced with another sequence of novel judgments from the same categories and the experiment terminates with a final feedback and a user survey.

MAPE, the mean absolute percentage error, is used as the performance and accuracy criterion to determine improvement or deterioration between question sequences and for the payouts that depend on MAPE values. MAPE is calculated similarly to MPE, but taking the absolute differences between



The mean percentage error (MPE) over all your answers: 46%

In the following you see the questions with your corresponding answers and the correct answers.

Question No.	Question	Your answer	Correct answer	The correct answer lies in your confidence interval
1	How many residents does Portugal have?	31000000	10145707	No
2	How long is the Fulda River (in km)?	200	218	Yes
3	How high is the highest peak of the Rockey Mountains (Mount Elbert) (in meters)?	4850	4401	Yes
4	How many residents does Turkey have?	40000000	85942343	No
5	How long is the Mekong River (including Langcang) (in km)?	4800	4350	No
6	How high is the Watzmann Mountain (in meters)?	3200	2713	No
7	How many residents does Denmark have?	16000000	5827680	No
8	How long is the Missouri River (in km) before entering the Mississippi?	4700	3726	No
9	How high is the Nanga Parbat Mountain (in meters)?	7200	8126	No
10	How many residents does Austria have?	25000000	9096201	No
11	How high is the Stol Mountain (in meters)?	2200	1673	No
12	How long is the Loire River (in km)?	1300	1020	No
13	How long is the Yellow River (in km)?	4700	5464	No
14	How high is the K2 Mountain (in meters)?	7200	8611	No
15	How many residents does Greece have?	22000000	10336087	No

Please take a moment of at least 30 seconds to consider this information.

Continue with questions

Figure 4. DSS Interface – Feedback Page Example

estimates and actuals. Information on how to interpret and apply MPE for debiasing is given in the briefing phase (without telling subjects that they will receive feedback) together with information on MAPE used as performance measure for payouts.

The infrastructure and the scenario characterized above is the one used in our first experiment. The experiment itself will be described in the subsequent section. A broader picture of our research, including other scenarios that will be considered in our research and how the first scenario is embedded in our research plan will be provided in Section 6.

4. Experiment

We first describe the research design in terms of the experiment's configuration (the general experimental procedure including the feedback provided and the loss function is described in Section 3). Second, we present the assumptions explored in the experiment and the measures used to analyze whether we find support for the assumptions. Third, we provide the results.

4.1. Research Design

In the experiment we have 74 subjects (34 in the treatment, 40 in the control group), of which 39 are female and the rest male. 41 subjects are business students and 33 subjects are (school) students.

Subjects are prompted for point estimates of quantities together with a 90% confidence interval

from general knowledge categories, namely *number of residents of a country*, *river length*, and *mountain height*. Example questions are: "How many residents does France have?", "How long is the Hudson River (in km)?", "How high is the Mount Everest (in meters)?". The experiment contains two sequences of 15 questions, 30 questions in total. Categories are neither communicated nor used by the correction model, but easy to anticipate by humans.

Estimation tasks are supported by cues: maps of the respective country including the ten largest cities with an indication of a range of their size; maps of the rivers with a scale in the legend; topographical maps of the mountains with a reference mountain height. These visual aids shall reduce error variance, but are also useful for heuristics applied and might trigger specific biases to be recognized as mean bias error patterns. For example, a subject may underestimate the additional river length stemming from the river loops and bends. The subject might then apply the error-feedback to debias her or his estimates only for questions of that type in case other categories seem unbiased.

The scenario mimics experts' environments where experts have expertise and basic confidence in all categories they are prompted for estimates, consider different types of visual cues and information for different types of estimation tasks, while expertise and heuristics applied might vary amongst categories. A human expert will typically be faced with categories or types of questions where he or she is particularly prone to biases. These types can be human-specific and a machine or statistical method would likely not be able to recognize the same types a particular human might have in mind. For instance, in our experiment it may be that a subject knows river lengths, mountain heights, and population sizes of north and middle European countries well but might be less familiar with other regions and then apply a geographical categorization.

After the first sequence of questions, a subject in the treatment (control) group receives feedback in terms of her or his mean bias measured as MPE (a blank page with the prompt to pause for 30 seconds). The MPE is displayed as inverse performance feedback to the subject, which can be easily applied for debiasing, together with the individual answers given by the respective subject and the actual correct answers per estimation question. The errors per question provide further hints by which categories the MPE might be driven, or where over- or underestimation is identifiable to foster reflection on how to further adapt judgments. We note that feedback is strictly related to patterns in a subject's own error history.

Following the feedback or the blank page, a subject answers the second sequence of questions, which are completely new to the subject. These questions are from the same categories as used in the first sequence.

After the experiment, a subject receives a debriefing and her or his MAPE is computed. A subject receives a payout for participation and has the chance to additionally win one of two prizes per treatment group. The lower the MAPE of a subject, the higher the chance to receive a prize. This incentivization is meant to increase the motivation and performance of subjects.

The overall experimental procedure is depicted in Table 1. In the following, we will describe the assumptions tested in the experiment.

Table 1. Experimental Design and Procedure

	Treatment Group	Control Group
15 Questions	x	x
Feedback	Yes	No
15 Questions	x	x
Feedback and Demographic Questions	x	x

4.2. Assumptions Studied

We split up the key assumption that one's own error patterns can foster wise and selective consideration of the feedback, into the (sub) assumptions A1–A5.

A1: MPE-feedback impacts judgment behavior resulting in MPE changing in the right direction.

The direction of change of a subject's MPE is analyzed to study if a reaction to the feedback can be assumed that leads in the right direction. If a subject receives a negative MPE, her/his subsequent MPE should be less negative or slightly positive and vice versa.

To test A1, per subject we determine the MPE over the answers in the first sequence (before the feedback or blank page) and the answers in the second sequence (after the feedback or blank page). Per subject we then determine whether her or his MPE changed in the right direction, and compare the ratio of right-direction MPE changes in the treatment versus the control group. The assumption is that the ratio is higher in the treatment than the control group and around 50 % in the control group (where no feedback is provided that might cause systematic MPE change).

For A1 we conduct a Fisher's exact test of independence between the results of the treatment and the control group for the ratio of right-direction MPE changes to detect a significant difference between the proportions of the two categorical variables. The test

is one-sided to test if the proportion of cases where the MPE changed in the right direction is higher in the treatment than in the control group. The treatment and control group are independent, relatively small samples, for which reason Fisher's exact test is a suitable non-parametric test.

A2: MPE-feedback induces emphasized adaptation of judgment in the category with the highest MPE, resulting in larger MPE change in this category.

The rational of A2 is that subjects know in which categories they are biased the most and use this knowledge wisely instead of blindly applying the feedback across all categories, as auto-debiasing unaware of categories would do.

To test A2, per subject and category MPE before and after the feedback (or blank page) is calculated. Then, the percentage of matches of the category with the highest absolute MPE in the first sequence and the category with the largest MPE change in the right direction from the first to the second sequence for the subjects in the treatment versus the control group are computed. If the percentage value in the treatment group exceeds the one in the control group, and in the treatment group both categories match in more than $\frac{1}{3}$ of cases (the baseline ratio in case of randomness), selective application of the feedback to specific categories can be assumed.

As for A1, we also test the significance of the difference in the results between treatment and control group with the Fischer's exact test for A2.

A3: MPE-feedback leads to higher MAPE reduction compared to no feedback given.

This assumption differs from A1 as it is related to accuracy improvements as a result of adapted judgment compared to A1 that studies solely MPE changes in the right direction. We note that MAPE might increase although MPE changes in the right direction when changes lead to absolute percentage errors exceeding the MAPE in the first sequence (if the absolute percentage errors increase).

After determining the difference of a subject's MAPE in the first and the second sequence, we calculate the ratio of MAPE improvements of subjects in the treatment versus the control group. We assume this ratio to be higher for the treatment group and again expect a ratio of around 50% in the control group due to random MAPE increases or decreases.

As for previous assumptions, we conduct a Fisher's exact test, again to find a significant difference between the results of treatment and control group.

A4: MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE leading to larger MAPE decrease in that category.

The rational of A4 is that subjects selectively apply the feedback to certain categories with high error levels (high bias) such that, respectively, the MAPE declines most strongly in these categories with high MAPE before the feedback.

The MAPE reduction after the feedback or blank page is computed per subject and category to determine the correspondence between the category with the highest MAPE in the first sequence and the category with the largest MAPE reduction from the first to the second sequence. If the ratio in the treatment group exceeds the one in the control group, and in the treatment group both categories match in more than $\frac{1}{3}$ of cases (baseline in case of randomness), wise, category-specific application of the feedback that leads to MAPE reduction can be assumed.

Again, we examine the significance of the difference in the results between treatment and control group by performing a Fisher's exact test.

A5: MPE-error-feedback leads to higher MAPE reduction particularly in categories with high MAPE in the first sequence compared to auto-correction.

Support for this assumption would indicate that feedback-based adjustment can mitigate strong false-corrections inherent when using auto-correction.

To test A5, the percentage of MAPE improvements in the treatment group between sequence one and two is compared to the percentage of hypothetical MAPE improvements through auto-correction in the control group. The improvements by auto-correction in the control group are computed by taking the answers of a subject of the second question sequence and including the MPE of the answers of the first questions sequence in the calculation of all hypothetically corrected answers for the second sequence. Then the MPE over these auto-corrected answers is computed. The Fisher's exact test is again used to test the significance of the difference between the results. Furthermore, per subject we determine the category with the highest MAPE in the first sequence and study whether feedback is beneficial in reducing these high error levels (MAPE) compared to auto-correction. For this specific category we compute how high the improvement is by subtracting the MAPE in the second sequence from the MAPE in the first sequence and taking the average per group thereof. We assume this value to be higher in the treatment group versus the control group (with auto-correction).

To test the significance of the difference between the results of treatment and control group for the level of MAPE reduction in percentage points, we conduct a non-parametric Wilcoxon-Test as we cannot assume a normal distribution and we have two independent groups. Here we cannot use the Fisher's exact test as our

target variable is numeric and not categorical as before.

5. Results

First, results are presented per assumption. Second, we summarize and discuss the results in an aggregated fashion and relate them to the key assumption.

A1 (MPE-feedback impacts judgment behavior resulting in MPE changing in the right direction):

In the treatment group, the relative frequency of MPE changing in the right direction after feedback is 91.2%. For the control group the corresponding value (after the blank page) is 65%. This strongly hints toward a consideration of the feedback leading to a systematic adaptation of the judgments that resulted in respective changes of the MPE observed afterwards: if a subject received a negative MPE as feedback, she or he typically gave higher responses to the following questions and vice versa.

The p-value of the Fisher's exact test is 0.0071, thus the result is highly significant at a 1% significance level.

A2 (MPE-feedback induces emphasized adaptation of judgment in the category with the highest MPE, resulting in larger MPE change in this category):

The percentage of MPE changes in the right direction in the category in which the MPE was the highest in the first sequence is 76.5% for the treatment group after the feedback and 50% in the control group after the blank page. The results for A2 hence provide underpinning that subjects selectively make MPE changes and support the presumption that subjects are aware of their category-specific estimation capability and use the feedback in those categories in which they assume their performance to be low, i.e. those with an emphasized mean bias.

For the results of A2, the p-value of the Fisher's exact test is 0.017, which indicates significance of the difference of results between treatment and control group at a 5% significance level.

A3 (MPE-feedback leads to higher MAPE reduction compared to no feedback given):

In the treatment group, 67.6% of the subjects reduced their MAPE after the feedback, compared to 50% of the control group after the blank page (no feedback). This result indicates that subjects seem to reflect on the feedback and use it to change their judgmental behavior in a way that their MAPE decreased after the feedback, in contrast to the control group that did not receive feedback and did not reduce their MAPE on average.

For these results the p-value of the Fisher's exact test is 0.097, thus the results are significant at a 10% significance level.

A4 (MPE-feedback induces emphasized adaptation of judgment in the category with the highest MAPE leading to larger MAPE decrease in that category):

In the treatment group, 58.8% of subjects made the highest MAPE improvement after the feedback in the category in which the MAPE was the highest in the first sequence, compared to 47.5% of the control group subjects after the blank page (no feedback). Furthermore, for 83.3% of those subjects in the treatment group, where the categories of highest MAPE in sequence one and highest MAPE improvement matched, the total MAPE considering all categories was improved after the feedback. This speaks for a wise and selective usage of the feedback, leading to increased accuracy of estimations.

The p-value of the Fisher's exact test here is 0.23, therefore the difference in the results between treatment and control group are not significant at a 10% significance level. However, due to the small sample sizes, the power of the test is obviously low, and larger sample sizes are required to achieve significant results here.

A5 (MPE-error-feedback leads to higher MAPE reduction particularly in categories with high MAPE in the first sequence compared to auto-correction):

In 67.6% of cases in the treatment group the feedback lead to MAPE improvements, whereas in 57.5% the auto-correction lead to MAPE improvements in the control group. For these results the p-value of the Fisher's exact test is 0.26, indicating that the results are not significant at a 10% level.

We find an average MAPE reduction of 12.48 percentage points in the category with the highest MAPE in the first sequence after the feedback in the treatment group compared to 3.45 percentage points after the blank page when applying auto-correction in the control group. The p-value for the Wilcoxon test is 0.092, for which reason the difference of the results is significant at a 10% significance level.

Although the results are not highly significant, they indicate subjects' capability of reducing the highest errors better (or more) by applying the feedback compared to a non-selective auto-correction.

Overall, we find indication that the group receiving error-feedback considers it selectively to improve judgments and judgmental accuracy. This is the case for overall error reduction as well as for category specific application of the feedback. In addition, the comparison with auto-correction already supports the assumption that large false-corrections with auto-debiasing can be mitigated with an error-feedback approach as proposed.

6. Discussion, Conclusion, and Outlook

The results provide strong support for our key assumption of wise consideration of feedback related to one's own error-pattern.

In particular, the high degree of matches between categories of highest MPE and correct MPE changes as well as MAPE and MAPE improvement represents the capability of humans to recognize error patterns or structures and being able to selectively adapt judgmental behavior accordingly. Reviewing the motivation of this paper, the category matches contribute to the aim of reducing strong false-corrections as errors are decreased the most where necessity for error reduction is the highest. This demonstrates that such a combination of the machine's and human's strengths – the computation and feedback of the MPE through the machine and the usage of the feedback by the human – achieves collaborative intelligence and is a promising direction of future research.

Considering the comparison of feedback versus auto-correction, we can hypothesize that humans applying feedback based on their own error compared to statistical models blindly applying learned error patterns can reduce large false-corrections. This relates to the research by Blanc and Setzer (2015a) who recommend to feed-back the supposed bias to the expert based on estimated model parameters to improve accuracy. Furthermore, it concerns their future research proposition to show experts bias-related feedback of their past forecasts and the forecast of a statistical model and give the expert the opportunity to act upon the feedback to reduce strong false-corrections. In our experiment, we obtained results indicating that this is supported by providing respective feedback to experts.

Our research has the limitation that, due to the COVID-19 pandemic, it has been challenging to conduct experiments with larger numbers of subjects, to be done in presence and not possible online as of the risk of subjects using search engines.

Regarding our future research plan, additional to running more experiments to further support our assumptions with the scenario used in our first experiment, Figure 5 shows further scenarios that will be considered, and how the first scenario is embedded.

The first scenario (X1), the one considered in this article's experiment, considers situations with low complexity for the human and high complexity for the machine. Therefore, in our first experiment the latent topics (here categories) can be considered to be easily detectable by humans, while the machine is unaware of the categories, can only provide aggregated feedback and is also merely able to auto-correct future estimations uniformly. The resulting assumption for X1 is that humans know when to integrate error feedback into their subsequent estimates as they know in which categories they are biased and might perform better when considering the feedback.

The second scenario (X2) considers situations with low complexity for both human and machine, i.e., here the latent topics are known by the machine and, for example, category-specific auto-correction can be applied. Due to human biases that might also be category-specific, the auto-correction performance of the machine might benefit from this information. An option for X2 would be giving feedback for each question with category awareness, in which case it might be more appropriate for the human to generally follow the machine feedback (X2a).

The third scenario (X3) covers situations with high complexity for both human and machine. Experiments with this scenario will contain questions that cannot be clearly assigned to a category. Furthermore, the categories will be latent in nature and the questions will have a rather vague reference to each other so that categories are not obvious. Here, it will be challenging for a machine to provide specific feedback, while the assumption is that the human might still be able to apply the general feedback wisely and selectively based on her or his domain knowledge and latent categories she or he has in mind.

Overall, the intention of the scenarios and our research is to better understand the situations in which feedback of what type can be expected to be beneficial, shedding light on the applicability of the approach in real-world settings.

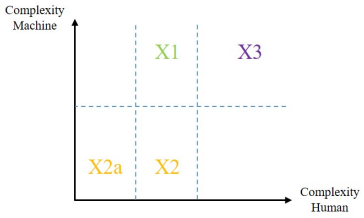


Figure 5. Experimental Scenarios Considered

References

Balzer, W. K., Doherty, M. E., & O'Connor, R. J. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin*, (106(3)), 410–433.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of

probability forecasters. *International Journal of Forecasting*, (8), 559–573.

Blanc, S., & Setzer, T. (2015a). Analytical debiasing of corporate cash flow forecasts. *European Journal of Operational Research*, (243(3)), 1004–1015.

Blanc, S., & Setzer, T. (2015b). Improving forecast accuracy by guided manual overwrite in forecast debiasing. *Twenty-Third European Conference on Information Systems (ECIS)*, Paper 66.

Blanc, S., & Setzer, T. (2016). When to choose the simple average in forecast combination. *Journal of Business Research*, (69), 3951–3962.

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science*, (36(8)), 887–899.

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, (15), 227–246.

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

Grant, A., Franklin, J., & Langford, P. (2002). The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality*, (30(8)), 821–836.

Jacoby, J., Mazursky, D., Troutman, T., & Kuss, A. (1984). When feedback is ignored: Disutility of outcome feedback. *Journal of Applied Psychology*, (69(3)), 531–545.

Klassen, R. D., & Flores, B. E. (2001). Forecasting practices of canadian firms: Survey results and comparisons. *International Journal of Production Economics*, (70), 163–174.

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

Lawrence, M., & O'Connor, M. (1993). Scale, variability, and the calibration of judgmental prediction intervals. *Organizational Behavior and Human Decision Processes*, (56), 441–458.

Lawrence, M., O'Connor, M., & Edmundson, B. (2000). A field study of sales forecasting accuracy and processes. *European Journal of Operational Research*, (122), 151–160.

Leitner, J., & Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information – a review of the literature. *European Journal of Operational Research*, (213(3)), 459–469.

Lim, J. S., & O'Connor, M. (1996). Judgmental forecasting with interactive forecasting support systems. *Decision Support Systems*, (16(4)), 339–357.

McCarthy, T. M., Golicic, S. L., & Mentzer, J. T. (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices. *Journal of Forecasting*, (25), 303–324.

Nagar, Y., & Malone, T. (2011). Making business predictions by combining human and machine intelligence in prediction markets. *ICIS 2011 Proceedings*, Paper 20.

Remus, W., O'Connor, M., & Griggs, K. (1996). Does feedback improve the accuracy of recurrent judgmental forecasts? *Organizational Behavior and Human Decision Processes*, (66(1)), 22–30.

Sanders, N. R., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega*, (31(6)), 511–522.

Sargeant, J. M., Mann, K. V., van der Vleuten, C. P., & Metsemakers, J. F. (2009). Reflection: A link between receiving and using assessment feedback. *Advances in Health Sciences Education*, (14), 399–410.

Sasse-Werhahn, L. F., Bachmann, C., & Habisch, A. (2020). Managing tensions in corporate sustainability through a practical wisdom lens. *Journal of Business Ethics*, (163), 53–66.

Sengupta, K. (1995). Cognitive feedback in environments characterized by irrelevant information. *Omega*, (23(2)), 125–143.

Sengupta, K., & Abdel-Hamid, T. K. (1993). Alternative conceptions of feedback in dynamic decision environments: An experimental investigation. *Management Science*, (39(4)), 411–428.

Zellner, M., Abbas, A. E., Budescu, D. V., & A., G. (2021). A survey of human judgement and quantitative forecasting methods. *Royal Society Open Science*.