Improving Judgmental Forecasts with Judgemental Bootstrapping and Task Feedback Support

MARCUS O'CONNOR1*, WILLIAM REMUS2 and KAI LIM3
1University of Sydney, Sydney, Australia
2University of Hawaii, Manoa, Honolulu
3City University of Hong Kong, Hong Kong

ABSTRACT

This study examines the utility of two widely advocated methods for supporting judgmental forecasts—providing task feedback and providing judgmental bootstrapping support. In a simulated laboratory based experiment that focused on producing composite sales forecasts from three individual components, we compared the effectiveness of these two methods in improving final judgmental forecasts. In the presence of cognitive feedback task, feedback led to better forecasts than providing judgmental bootstrap forecasts. Simply providing bootstrap forecasts was of no additional benefit over a control condition. This was true in terms of the Brunswik Lens model measures of achievement, knowledge, and consistency, and in terms of forecast accuracy. This occurred both in stable environments and when special events (unusual one-time events requiring adjustments to the forecasts) arose. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS judgmental forecasts; statistical forecasts; task feedback; judgmental bootstrapping; feedback intervention theory; social judgment theory; Brunswik lens model

INTRODUCTION

Given the ever-changing business environment, it is important to help managers to develop accurate forecasts. Since managers generally prefer their own judgmental forecasts to statistically modeled forecasts (Dalrymple, 1987; Lawrence, 1983; Lawrence, O'Connor & Edmundson, 2000; Sanders & Manrodt, 2003), improving business forecasting should center on improving managerial judgmental forecasting skills. However, relative to statistical forecasts, judgmental forecasts are often biased (Blattberg & Hoeh, 1990:

*Correspondence to: Marcus O'Connor, Discipline of Business Information Systems, School of Business, Faculty of Economics & Business, University of Sydney 2006, Sydney, Australia. E-mail: m.oconnor@econ.usyd.edu.au

Copyright © 2005 John Wiley & Sons, Ltd.
There have been many attempts to improve judgmental forecasting skills. Some of the most popular of these approaches generally involve either the use of feedback or some type of (typically computer based) support of the process of developing the judgmental forecasts. Both approaches are predicated on the assumption that people prefer to have a substantial input into the process of producing the forecast. A key task is to identify which of these approaches work best; and this paper seeks to contribute to our understanding of this issue.

In the case of task feedback, the proposition is that detailed and informative feedback that reports actual and optimal performance will improve the way in which people make their subsequent forecasting judgments. From the computer support approach, these forecasts could best be improved by providing support that helps to reduce the inconsistency and biases that are known to characterize human judgment. In this paper we report an examination of the efficacy of these two approaches in their support of the judgmental forecasting process. We undertake this examination in two forecasting environments: a stable environment and an environment that exhibits some degree of instability. In the sections that follow, we first review the literature on feedback systems and examine the applicability of Feedback Intervention Theory (FIT) of Kluger and DeNisi (1996). We then review the literature on bootstrapping models and their impact on improving judgment. Then are details of an experiment to test the impact of task feedback and bootstrapped forecasts on judgmental forecast accuracy in both stable and unstable environments described. The results are reported and the findings discussed.

BACKGROUND LITERATURE

In this section, we first review the background literature on the types of feedback that are likely to enhance performance in tasks involving human judgment. We then review the literature on the utility of providing bootstrap forecasts to people as a means of reducing their inconsistency in judgmental forecasts.

Feedback literature

There is a rich and vast literature on improving performance by using feedback. This literature has been coherently integrated by the Feedback Intervention Theory (FIT) of Kluger and DeNisi (1996). Feedback Intervention Theory predicts the impact of feedback on the task-motivational processes, on the task learning processes, and on meta-task processes. Kluger and DeNisi (1996) support the theory with a comprehensive review of the literature and a meta-analysis of selected research. Our focus is on the effects of task feedback on the task learning processes. The task learning aspect of FIT argues that the impact of feedback intervention is a function of the feedback’s ability to support the generation and testing of hypotheses about the task. If the feedback supports the hypotheses testing, FIT predicts positive learning effects. However, if the feedback does not support the hypotheses testing by showing the correctness of the hypotheses, FIT predicts that negative or no learning at all may result.

Our study uses the Balzer, Sulsky, Hammer and Sumner (1992) typology for organizing the various types of information feedback. The typology categorizes information feedback as task information, cognitive information, or functional validity information feedback. Task information feedback provides information about the task; this is typically in the form of optimal task cue (or information) weights. Cognitive information feedback provides information about one’s cognitive strategy; typically this is the form of guidance based on the decision maker’s cue weights. Functional validity information feedback provides information on the relationship between the task system and cognitive strategy.
Balzer et al. (1992) conducted an experiment using a Multiple Cue Probability Learning (Brunswik Lens Model; Brunswik, 1952) task to examine the merits of the various feedback types discussed above. They found that all treatments providing task information improved performance over the no feedback condition. They also found that task feedback provided with cognitive feedback outperformed cognitive feedback alone. Thus, they argued that task information was the crucial type of feedback needed to improve performance. On the other hand, Tape, Kripal, and Wigton (1992) found that, contrary to much of this literature, task and cognitive feedback was not as effective as their outcome feedback. Noting the difference in their result from most of the others, they suggest that it may have been due to the nature of their task. They used a five cue linear-additive task to predict cardio-vascular failure where the cues were all positive—unlike many of the other studies which have incorporated negative cue weights in their tasks.

Remus, O’Connor, and Griggs (1996) replicated and extended Balzer et al. (1992) using a judgmental forecasting task. Task feedback with the cognitive feedback led to significantly better forecasting accuracy than cognitive feedback alone or the other feedback treatments. Thus, task feedback appeared to be an excellent way to improve judgmentally generated forecasts. In this paper, we extend the managerial findings of Remus et al. (1996) to the task of combining forecasts from a number of estimates made by various departments in the organization. As Harries, Yaniv, and Harvey (2004) point out, such a task is often characterized by realistic decision making where the judgments of various advisors need to be “combined”.

**Judgmental bootstrapping**

Judgmental bootstrapping is a technique based on the idea that people usually have of important factors in a task (Tape et al., 1992), but often do not know explicitly how much weightage to give to each factor and are inconsistent in the weights they actually use. If the latter is true, then appropriate weights can be extracted through regression on the past history of the task and used in a linear model to suggest good decisions or forecasts. Judgmental bootstrapping entered the managerial literature through the work of Bowman (1963) who argued that judgmental bootstrapping feedback led to better decisions primarily through the reduction of erratic decision making. He supported his argument with an analysis comparing the performance of actual decisions with the performance of decisions based on the bootstrapped model.

There is a lot of empirical support for judgmental bootstrapping. Kleinmuntz (1990) provides an excellent review of this vast literature. Of particular interest is the work of Camerer (1981) who developed the conditions for judgmental bootstrapping models to outperform people; those conditions covering a wide variety of situations. Also worthy of note is Armstrong (2001a) who provides a meta-analysis of empirical comparisons on the value of judgmental bootstrapping as a means of improving judgmental forecast accuracy. In essence, judgmental bootstrapping seems to be of benefit because judgmental inconsistency is eliminated (Newton, 1965; Stewart, 2001). Note, however, that in this study we do not simply use the bootstrap forecasts—we merely provide such forecasts to the decision maker, who can choose to use it in any way they feel. Again, our assumption is that the decision maker wants to be in control and so we simply provide bootstrap forecasts to them to assist them in their judgments. We do not replace human judgment; even though there is a substantial literature to suggest this is effective in improving performance (e.g., Cooksey, 1996). We simply provide the bootstrap forecast to guide them in the hope that their consistency and performance will be improved.

**Forecasting environments**

Sometimes forecasting environments are characterized by stability and sometimes they are unstable. Goodwin and Fildes (1999) noted that the on-going forecasting process is occasionally interrupted by exogenous influences such as a competitor’s actions, and have termed these ‘special events.’ When a special event occurs, the forecast has to be adjusted to reflect the impact of the event. They reported an experiment...
on the impact of such special events; they found that the adjustments made were far from optimal. Not un-
expectedly, people seemed to perform better in a stable forecasting environment.

Since the introduction of "special events" will reduce the predictability of the task feedback and the uti-
licity of the bootstrap forecasts provided, we anticipate that there will be a reduction in all measures of per-
formance (discussed later).

RESEARCH DESIGN

Participants
The 77 participants in the experiment were undergraduate students at a major university. They were recruited
from a junior level statistics course that covered forecasting during the three weeks prior to the experiment;
this course was required of all business students. In the weeks prior to the experiment, regression was
explained in terms that were consistent with the weighting performed in this experiment. Participation in
the experiment was rewarded with enough points to make one level of difference in the final course grade
for many students (e.g., change a B to an A). All the students enrolled in the course participated in the
experiment.

Task description
The experimental task was to develop a demand forecast for a product for the next period given the forecasts
that the marketing, planning, and sales departments each had developed. The underlying demand equation in
standardized form had a 0.5 weight on the marketing department’s forecast \( (b_1) \), a 0.26 weight on the plan-
ning department’s forecast \( (b_2) \), and a 0.24 weight on the sales department’s forecast \( (b_3) \). Thus, the market-
ing department’s forecast was roughly twice as important as the forecasts from the planning and sales
departments.

A Visual Basic program was developed to model the above forecasting task. The Marketing, Planning, and
Sales Departments’ forecasts were generated using normal distributions generated by a random number gen-
erator; the random number generator seed was the participant’s social security number. Thus, each partici-

pant had unique variable values for the forecasts. Each of those forecasts had a mean of 2500 units. The
standard deviation of the Marketing forecasts was 200 units, the standard deviation of the Planning forecasts
was 200 units, and the standard deviation of the Sales forecast was 400 units. The forecasting task error was
normally distributed with mean zero and standard deviation of 343. For each trial, the three component fore-
casts were provided at the same time. Subjects could also observe (through a tabular presentation) past trials,
the history of their judgments, and the actual value (past 10 only), including a measure of absolute error.

The first 40 trials were in a stable forecasting environment. The participants first viewed the Sales, Market-
ing, and Planning Departments’ forecasts. The participants were required to combine these forecasts to cre-
ate an overall demand forecast. When making their demand forecasts, the participants needed only to use the
mouse to move a slider to a demand value and ‘click’ to record that value. The value of the actual demand
was then displayed.

Beginning in trial 41, special events were presented to the participants; these were intended to evoke an
adjustment of 500\(^1\) units to the forecasts. In the special event environment, the software provided a balanced
presentation of 6 upward special events and 6 downward special events were introduced. The downward

\(^1\)The guidelines we used to establish this value were that (1) the special event would have to have a substantial effect on demand, (2) the
industry would be relatively small with a few known competitors, and (3) the event would seem realistic in terms of our consulting
experience in such situations. In such a case we thought 20% of the average demand of 2500 would be appropriate. There was no
additional noise added to this special events adjustment.
special event messages occurred in trials 42, 49, 56, 58, 64, and 66. The upward special event messages occurred in trials 44, 47, 52, 54, 62, and 68. The experiment terminated with trial 70. Participants were not informed when the special events would occur.

The procedure for special events was the same as used for other forecasts except that after the participants clicked on a forecast, a message box popped up with 'Message from our CEO: Our competitor’s recent marketing campaign has succeeded/failed. How much do you want to adjust your forecast downward/upward?' The participants were then able to adjust their forecast upward or downward using a second slider contained in the message box. This procedure uses forecast decomposition since the study by Webby, O’Connor and Edmundson (2005) had found decomposition to be effective in improving forecast accuracy in the special event forecasting environment.

Four treatments were needed to test the hypotheses described earlier; these four treatments led to four versions of the software. Each treatment contained three forecasting segments. Twenty or more trials in each segment are sufficient to allow stable Brunswik Lens model performance measures to be calculated (Stewart, 1988; Tape et al., 1992).

Treatments 1a and 1b—The control groups

In Treatment 1a, the forecasting task as described above lasted for 70 trials in a stable environment; that is, no special events were presented. In Treatment 1b, the first 40 trials were in stable environment but the next 30 trials were in the special events environment. Thus, treatments 1a and 1b were identical for the first 40 trials and are combined to form Treatment 1 when testing the hypotheses in a stable environment during the first 40 trials. The distinction between Treatment 1a and 1b was in trials 41 to 70. As is described later, these treatments provide two alternative control groups for the special events environment during trials 41 to 70.

Treatment 2—Task feedback

Task information feedback is information about the true nature of the task. As in Balzer et al. (1992), the cue weights ($r_{c_1}$ to $r_{c_3}$) used to generate the demand were given to the subjects at the end of block 1 (period 20) and block 2 (period 40). Cognitive information feedback is information about the participant’s cognitive strategy, in particular, the relationship between cues and the participant’s judgments. As in Balzer et al. (1992), the participant’s subjective cue weights ($r_{s_1}$ to $r_{s_3}$) (i.e., the weights that capture the participant’s behavior in the previous 20 periods) were also provided for the forecasting task. It is important to note that in many practical situations like forecasting, task and cognitive feedback are often commingled because the two concepts are closely linked (Doherty & Balzer, 1988); they are also allowed to commingle in this experiment for the experimental control reasons described later. As noted earlier, Balzer et al. (1992) and Remus, O’Connor and Griggs (1996) found that no other kind of feedback outperformed task and cognitive information feedback; although we note the contrary finding of Tape et al., 1992.

As is standard practice in the Brunswik Lens model literature, the task and cognitive feedback weights were displayed as standardized coefficients. Each feedback weight had a button assigned to it which when 'clicked' would provide an explanation (in simple terms) of the meaning of that feedback measure. These explanations were pre-tested to assure that the participants would understand the explanation. The feedback messages were similar to those used by Balzer et al. (1992) and Remus et al. (1996). Where the subjective weight was smaller (larger) than suggested for a provided cue, they were encouraged in these messages to increase (decrease) their reliance on it; where it appeared that the cue was being used approximately as suggested, they were encouraged to continue with their weight for that cue. In Treatment 2, the first 40 trials were in a stable environment and the next 30 trials contained special events as described above; task feedback was provided at period 20 and again at period 40.
Treatment 3—Judgmental bootstrap forecasts

In this technique, the cue weights that the participants implicitly used to make a forecast were extracted using regression. These cue weights were then used to produce a forecast for the next period. In this experiment, the software automatically extracted the weights uniquely for each subject after trial 20 (based on trials 1 to 20). These weights were used to provide feedback from trials 21 to 40. After trial 40, the weights were again extracted; this time based on trials 21 to 40. This second set of weights was used to give feedback from trial 41 onward.

The participant’s cue weights ($r_{1}$ to $r_{3}$) were displayed and explained to the participants after trials 20 and 40. Thereafter, a judgmental bootstrapping forecast was provided prior to each forecast; that is, the participant’s regression rule (using the participant’s cue weights from the previous block) automatically suggested forecasts for each new trial. Thus, participants in this treatment were provided with a forecast based on their past forecasts.

By design, both the task feedback and judgmental bootstrap forecast treatments included feedback of the participant’s actual cue weights ($r_{1}$ to $r_{3}$). Thus, the feedback of the participant’s actual cue weights was not a confounding variable when comparing Treatment 2 and 3. The resulting comparison thus is only between the impact of task feedback and the impact of the forecasted values generated with judgmental bootstrapping feedback.

Performance measures

We used four measures to compare the performance of the subjects. Our measure of consistency ($R_5$) is from the Brunswik Lens Model, it reflects the consistency with which a subject applies his or her judgment. The measures of performance also include the Brunswik Lens knowledge ($G$) and achievement ($R_6$) measures. Knowledge ($G$) measures the extent to which a person matches the systematic features of his or her judgment policy to those of the task. Achievement ($R_6$) measures the match between the judgments and the actual outcomes (Brehmer, 1988, p. 23). We also used a common measure of performance used in forecasting, namely, absolute percentage forecasting error (Makridakis, 1993).

Procedure

The participants were trained in the use of software, immediately before beginning the experiment. The context for the forecasts was given (sales forecasting) and the forecasting task explained. The participants were also told that they would be provided with helpful, valid feedback on their forecasting accuracy to aid them in improving their forecasts. The experimenter demonstrated the software and participants then made two practice forecasts to finish the training session.

The participants completed 70 forecasting trials. In all trials, their forecast was taken using a slider as an input device. The participants were randomly assigned to a treatment and feedback was provided as described above. The average time taken to complete all the trials was less than 1 h.

Hypotheses

As the study is concerned with examining the impact of intervention techniques (task feedback and bootstrap forecasts), the hypotheses concern the improvement (or otherwise) in the various performance measures across time period blocks. We anticipate that there will be different effects depending on whether stable or unstable environments are being considered.

For stable forecasting environments, Feedback Intervention Theory suggests that the task feedback will increase one’s knowledge about the task and permit examination of hypotheses, which will concomitantly increase performance (using the principles of the Brunswik lens equation; Cooksey, 1996). Thus, we expect...
that knowledge of the task (G) and hence performance will be improved over Control with the provision of task feedback. Since the bootstrap forecasts are unique to that group, we expect that this group will display forecasts that are more consistent (and hence have higher performance) than the Control group. We have no information as to whether Task feedback or Bootstrap forecasts will have a superior effect on performance. However, despite all these expectations, since no research has been done on this matter before, the hypotheses are expressed in the null form. Specifically, although we expect that the performance of the Task feedback and Bootstrap forecast to be superior to the Control group, our null hypotheses for the period of stability is:

\[ H_{1_0}: \] There will be no difference in the forecast performance (achievement, \( r_a \) and absolute forecast error, \( APE \)) between the Bootstrap, Task feedback, and Control groups.

Similarly, although we expect that the forecast consistency of the bootstrap forecasts will be greater than that of the Task Feedback and Control groups, our null hypothesis on consistency is:

\[ H_{1_0}: \] There will be no difference in forecast consistency (\( R_c \)) between the Bootstrap, Task feedback, and Control groups.

And, whilst our expectation is that the Task Feedback groups will exhibit greater knowledge about the task (G) compared to the Bootstrap and Control groups, our null hypothesis is:

\[ H_{1_0}: \] There will be no difference in knowledge (G) between the Bootstrap, Task feedback, and Control groups.

For unstable forecasting environments (characterized by special events), there does not appear any evidence to suggest that one group will be better than another, on any of the dependent variables. The only expectation is that all groups will reduce their performance, knowledge, and consistency with the introduction of special events. Thus, our last hypothesis is:

\[ H2: \] The introduction of special event information will reduce performance, knowledge, and consistency for all treatment groups.

RESULTS

Before any testing of the hypotheses was made, we needed to ensure that all treatment groups displayed the same results for the first block of trials (periods 1–20). MANOVA was used to test whether there were any differences between the groups for all four dependent variables. Results indicate that there were no differences.

The impact of feedback on judgmental forecasting in stable forecasting environments\(^2\)

The four forecast performance measures from the stable forecasting environment were analyzed using MANOVA, the mean value of these measures is shown in Table 1. Recall that Treatment 1a and 1b are combined to form Treatment 1. As Table 1 shows, there were significant main effects between the treatments for

\(^2\)Unless otherwise noted, we used MANOVA for the overall test and ANOVA with multiple comparisons (Pedhazur & Schmelkin, 1991, pp. 482–490) to explore significant MANOVA results. Since correlation measures are known to be non-normal, all Brunswik Lens model-related measures were transformed to normally distributed variables using the Fisher z transformation; this is customarily done in the Brunswik Lens model literature (Stewart, 1988).
Table 1. Performance measures for forecasting in a stable environment

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>After forecasting support began (Trials 21–40)</th>
<th>Across treatments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute percentage error</td>
<td>T1: 116.8</td>
<td>F = 5.28</td>
</tr>
<tr>
<td></td>
<td>T2: 96.0</td>
<td>p = 0.007</td>
</tr>
<tr>
<td></td>
<td>T3: 115.4</td>
<td>df = 2.74</td>
</tr>
<tr>
<td></td>
<td>Ave: 111.1</td>
<td></td>
</tr>
<tr>
<td>Achievement (r_a)</td>
<td>T1: 0.629</td>
<td>F = 4.57</td>
</tr>
<tr>
<td></td>
<td>T2: 0.718</td>
<td>p = 0.013</td>
</tr>
<tr>
<td></td>
<td>T3: 0.587</td>
<td>df = 2.74</td>
</tr>
<tr>
<td></td>
<td>Ave: 0.642</td>
<td></td>
</tr>
<tr>
<td>Consistency (r_c)</td>
<td>T1: 0.857</td>
<td>F = 5.89</td>
</tr>
<tr>
<td></td>
<td>T2: 0.918</td>
<td>p = 0.004</td>
</tr>
<tr>
<td></td>
<td>T3: 0.849</td>
<td>df = 2.74</td>
</tr>
<tr>
<td></td>
<td>Ave: 0.871</td>
<td></td>
</tr>
<tr>
<td>Knowledge (G)</td>
<td>T1: 0.897</td>
<td>F = 3.54</td>
</tr>
<tr>
<td></td>
<td>T2: 0.945</td>
<td>p = 0.034</td>
</tr>
<tr>
<td></td>
<td>T3: 0.841</td>
<td>df = 2.74</td>
</tr>
<tr>
<td></td>
<td>Ave: 0.896</td>
<td></td>
</tr>
</tbody>
</table>

Measures shown are by treatment followed by the overall average.

Legend: T1: control group—the participants were not provided forecasting support (n = 38) (Note: Treatment 1 is a pooling of control groups la and lb). T2: the participants were provided with task and cognitive feedback for forecasting support (n = 20). T3: the participants were provided with judgmental bootstrap forecasting support (n = 19).

all measures. Thus, contrary to the null hypotheses H1a, H1b, and H1c, there were differences in performance, consistency and knowledge between the treatment groups.

To further explore these differences, post hoc tests were performed between the three treatment groups with a view to compare the results with our expectations (presented earlier) which were variously derived from Feedback Intervention Theory and Social Judgment Theory using the Brunswikian methodology. Table 2 presents the hypotheses, expectations, and the results of these post hoc tests.

As Table 2 shows, results did not support the null hypotheses; and only some of our expectations. We expected that the Task feedback group (through improved knowledge, G) and the Bootstrap forecast group (through improved consistency, r_c) would have produced superior performance compared to the Control

Table 2. Comparison of hypotheses, expectations and results

<table>
<thead>
<tr>
<th>H</th>
<th>Null Hypothesis</th>
<th>DV</th>
<th>Expectation</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1a</td>
<td>Control = TFB = BS</td>
<td>r_a</td>
<td>TFB = BS &gt; Control</td>
<td>H1a not supported</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Expectation not supported TFB &gt; BS = Control</td>
</tr>
<tr>
<td>H1b</td>
<td>Control = TFB = BS</td>
<td>r_c</td>
<td>BS &gt; TFB = Control</td>
<td>H1b not supported</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Expectation not supported TFB &gt; BS = Control</td>
</tr>
<tr>
<td>H1c</td>
<td>Control = TFB = BS</td>
<td>G</td>
<td>TFB &gt; BS = Control</td>
<td>H1c not supported</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Expectation supported TFB &gt; BS = Control</td>
</tr>
</tbody>
</table>

Legend: TFB denotes Task and Cognitive Feedback; BS denotes Bootstrap forecasts. Where ‘ = ’ denotes statistical equivalence of groups and ‘ > ’ denotes the superiority of one group over another.
group. And we also expected that consistency (Rc) would have been greatest for the Bootstrap forecast group, with knowledge (G) highest for the Task feedback group. However, as the final column of Table 2 shows, the Task feedback group performed best for all dependent variables. Furthermore, there were no differences between the Control and Bootstrap forecast groups for all the dependent variables. This was unexpected. Clearly, the provision of bootstrap forecasts which were intended to increase consistency did not have their desired effect. Thus, for the stable forecast periods (periods 21–40 only), there were differences in performance due to treatment groups, with the Task feedback group performing best.

The impact of feedback on judgmental forecasting when special events occurred

Table 3 displays various forecast performance measures in the stable forecasting environment (trials 21 to 40) and in the special events environment (trials 41 to 70). Using MANOVA, we found significant main effects across the four treatments in consistency (Rc) (F(3, 73) = 5.77, p = 0.001) and achievement (ra) (F(3, 73) = 6.35, p = 0.001) in the special events environment. The absolute forecast error across the treatments (F(3, 73) = 3.91, p = 0.012) was also significant. The main effect across treatments on knowledge (G) (F(3, 73) = 2.55, p = 0.062) was almost significant using a two tail test. The second hypothesis tests whether there was any reduction in any of the dependent variables with the introduction of special events in periods 41–70. As Table 3 shows, there was no difference in any of the dependent variables across the time blocks.

Table 3. Performance measures for forecasting when special events occur

<table>
<thead>
<tr>
<th></th>
<th>Before special events (Trials 21–40)</th>
<th>During special events (Trials 41–70)</th>
<th>Across treatments</th>
<th>Across time interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute percentage error</td>
<td>T1a 114.6</td>
<td>T1a 118.3</td>
<td>F = 2.38</td>
<td>F = 3.91</td>
</tr>
<tr>
<td></td>
<td>T1b 119.0</td>
<td>T1b 122.9</td>
<td>p = 0.128</td>
<td>p = 0.012</td>
</tr>
<tr>
<td></td>
<td>T2 96.0</td>
<td>T2 106.1</td>
<td>df = 1.73</td>
<td>df = 3.73</td>
</tr>
<tr>
<td></td>
<td>T3 115.4</td>
<td>T3 116.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ave 111.1</td>
<td>Ave 115.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achievement (ra)</td>
<td>T1a 0.643</td>
<td>T1a 0.604</td>
<td>F = 3.60</td>
<td>F = 6.35</td>
</tr>
<tr>
<td></td>
<td>T1b 0.614</td>
<td>T1b 0.540</td>
<td>p = 0.062</td>
<td>p = 0.001</td>
</tr>
<tr>
<td></td>
<td>T2 0.718</td>
<td>T2 0.692</td>
<td>df = 1.73</td>
<td>df = 3.73</td>
</tr>
<tr>
<td></td>
<td>T3 0.587</td>
<td>T3 0.594</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ave 0.642</td>
<td>Ave 0.609</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistency (Rc)</td>
<td>T1a 0.851</td>
<td>T1a 0.822</td>
<td>F = 0.02</td>
<td>F = 5.77</td>
</tr>
<tr>
<td></td>
<td>T1b 0.863</td>
<td>T1b 0.863</td>
<td>p = 0.877</td>
<td>p = 0.001</td>
</tr>
<tr>
<td></td>
<td>T2 0.918</td>
<td>T2 0.914</td>
<td>df = 1.73</td>
<td>df = 3.73</td>
</tr>
<tr>
<td></td>
<td>T3 0.849</td>
<td>T3 0.822</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ave 0.871</td>
<td>Ave 0.871</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge (G)</td>
<td>T1b 0.890</td>
<td>T1b 0.853</td>
<td>F = 3.40</td>
<td>F = 2.55</td>
</tr>
<tr>
<td></td>
<td>T1a 0.905</td>
<td>T1a 0.841</td>
<td>p = 0.069</td>
<td>p = 0.062</td>
</tr>
<tr>
<td></td>
<td>T2 0.945</td>
<td>T2 0.919</td>
<td>df = 1.73</td>
<td>df = 3.73</td>
</tr>
<tr>
<td></td>
<td>T3 0.841</td>
<td>T3 0.854</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ave 0.896</td>
<td>Ave 0.867</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Measures shown are by treatment followed by the overall average.

Legend: T1a: control group—the participants were not provided forecasting support and no special events occurred. (n = 19).
T1b: control group—the participants were not provided forecasting support and special events occurred. (n = 20).
T2: the participants were provided with task and cognitive feedback for forecasting support and special events occurred. (n = 19).
T3: the participants were provided with bootstrapping forecasting support and special events occurred. (n = 19).

In contrast with the stable events period, all four treatment groups were included in the analysis.
Thus, H2 is not supported; although we note that for most of the treatment groups there was a reduction in the means for the dependent variables. Furthermore, Table 3 also reports the results of the comparison between the treatment groups for the two forecast blocks. The overall results suggest that the superiority of Task feedback over the other treatments was carried over to the unstable special event periods (block 3).

DISCUSSION

Using Feedback Intervention Theory we expected that people in the Task feedback group would have a mechanism to amend their hypotheses (the foundation of FIT) which would lead to superior performance in performance (achievement \( r_a \), and forecast error, APE) and knowledge (\( G \)) than the Control group. Similarly, with a reliance on Social Judgment Theory and the Brunswikian methodology, we expected that consistency (\( R_c \)) would improve for the Bootstrap forecast group over the Control group. As reported in Table 2, our expectations were not substantiated by the results. Only the Task feedback group overall performed better than the Control group.

The nature of the experimental task was to simply understand and appropriately weight the three forecasts that were provided to them. The optimal weighting system was \( b_1 = 0.5 \), \( b_2 = 0.26 \), and \( b_3 = 0.24 \). An excellent approximation to these weights would simply be a two stage approach: the first stage being an average of the second and third forecasts, and the second stage then to simply average the first forecast with this other average. Clearly, very few participants picked up on this simple approximation. Across all treatment groups for block 1 (periods 1–20), the average weight for the first forecast \( (b_1) \) was 0.326, for the second forecast \( (b_2) \) was 0.211, and for the third forecast \( (b_3) \) was 0.463. Each of these weights was significantly different from each other. In many ways, this result is somewhat puzzling. Why should there be significant differences between the forecast weights in this initial period? More specifically, why should the third forecast be given such a high weight relative to the others? As the third forecast was the last to be provided, perhaps some form of recency effect may have been in operation. However, since our three forecasts were always presented in the same order, we are unable to determine whether this is a satisfactory explanation. Figure 1 shows the relative weights for the three forecasts across the three treatment groups for the first and second forecasting blocks. Considering the first forecasting block (periods 1–20) only, one can clearly see that there seems to be a degree of accordance in weights between the Control and the BS groups. However, the weights employed by the Task feedback group are somewhat different from the others. In particular, the relative weight given to \( b_3 \) by the Task feedback group is significantly different from the other two \( (t = 2.7, p < 0.01) \). As one would have expected all groups to employ similar weights for the initial 20 periods (before any treatments had been enacted), this result is unexpected and without explanation. However, we note that the weight on \( b_3 \) by the

![Figure 1. Relative weights for the different treatment groups for the first two blocks](image-url)
Task feedback group for this initial period is in the wrong direction, away from the optimal weights. Thus, the first period of forecast weighting revealed significant differences between the three weights: with the Task feedback group’s weights quite different (and inferior) from the other two groups.

When comparisons are made between periods 1 and 2, a different picture emerges. For the Control group, the weights have changed a little between the forecast blocks. Whilst the weight of the second forecast ($b_2$) remained relatively constant, a greater weight was appropriately given to the first forecast ($b_1$) with a concomitant reduction in the third weight ($b_3$). Although this was nowhere near optimal, it was at least a movement in the right direction. Thus, the Control group appears to display a degree of learning at the task, even though it should be remembered that NO task feedback was given at the end of the first block for this group.

In contrast, the Task feedback group received task and cognitive feedback at the end of the first block. This entailed some information on the optimal weights as well as specific and individual information as to the weights they had employed in the previous 20 periods. A simple comparison revealed the difference between the optimal and in-use weights, enabling them to make amendments for the remaining forecast blocks. Past research (e.g., Balzer et al., 1992; Remus et al., 1996) has demonstrated this to be most effective in improving performance (cf. Tape et al., 1992). As Figure 1 illustrates, people in the Task feedback group have responded vigorously to the feedback. Whilst for block 1 the highest weight was (inappropriately) given to the third forecast, for block 2 the highest weight was (appropriately) given to the first forecast. This large change was highly significant ($t(19) = 5.8$, $p < 0.001$), and there was no difference in the weight given to the second forecast between the blocks. It appears that for this Task feedback group, they have correctly swapped their major weight from the third to the first forecast. Although this is also true for the Control group, they have done so to a much larger extent.

As for the Bootstrap forecast group, the weights for the three forecasts are really very similar between the two blocks (paired t-tests revealed no differences in any of the weights between the blocks). This is a priori somewhat surprising; but, given the performance results reported in Table 2, it was to be expected. We expected that the Bootstrap forecast group would tend to use the provided forecasts, as this took some of the guesswork out of the weighting system. If this was the case, one would find that consistency ($R_i$) would have increased substantially for this group (as implied by Social Judgment Theory). As Table 3 reveals, consistency actually fell, it did not increase. Thus, despite the provision of forecasts that were based on the individual’s past weighting scheme, people in that group seemed to ignore such information and continue to experiment with the weighting system. Whilst it is disappointing to note this lack of improvement in consistency, we note that Tape et al. (1992) also found that consistency did not improve over extensive (170+) trials in a cardio-vascular risk determination task using doctors as subjects. As Stewart (2001) comments, people seem to know they are inconsistent in their judgments and seem to accept it as a characteristic of their behavior.

In the third block (periods 41–70), special events were introduced. This consisted of a message window indicating that a marketing campaign had either proved successful or unsuccessful, a successful (unsuccessful) campaign warranting an upward (downward) revision of about 20% in the forecast. Contrary to $H_3$, where we expected performance to deteriorate in this block, people in all groups seemed to be able to appropriately accommodate this extra information. Perhaps this was due to the fact that the experimental procedure necessitated a two-stage process of firstly estimating their forecast as normal, with a further adjustment for any special event. Figure 2 shows the relative weights for all four treatments for periods 2 and 3.

One of the most interesting observations from Figure 2 is the relative weights used by people in the Task feedback group. Whilst there was no difference in performance between the two forecast blocks, as Figure 2 shows there has been a change in the weights for the Task feedback group. Interestingly, there has been a significant downward revision of the relative weight for the first forecast from 54.7% to 43.7% ($t = 3.02$, $p < 0.01$). As the optimal weight is 50%, this represents a deterioration in performance. The reason this deterioration was not reflected in the performance statistics is that there was a concomitant change in the relative weights for the other forecasts. Still, a substantial change in the relative weights for this group is interesting.
Given the two stage process of forecast estimation for this period, it suggests that the special event information made much more of a difference to the best performing group, leading to a deterioration in performance.

It is a common task in business forecasting to combine the judgments from a number of sources. Of course, there can be statistical advantages from combining through simple averaging (see, Clemen, 1989; Lawrence, Edmundson, & O’Connor, 1986). But, given that people from various parts of the organization have access to differing amounts of information, from various sources and perspectives, and with varying degrees of relevance and validity, an unequal weighting of the individual component forecasts seems necessary. Dalrymple’s (1987) survey of sales forecasting practices reveals that about two-thirds of businesses report some degree of combining of forecasts, typically using informal processes. Our study has shown that if such combinations are completed using informal judgmental processes, there may be difficulties in achieving accurate composite forecasts. Consistency in the judgmental combination process seems to be a major problem. Our study has shown that even providing people with a statistically derived composite forecast (in the form of a bootstrap forecast) using their own individual weights does not seem to improve consistency or performance. People seem to want to “tinker” with the statistical composite: a fact that has been observed in time series forecasting studies (e.g., Lim & O’Connor, 1996). Clearly, the message seems to be that, once a forecast weighting system has been determined (possibly a difficult social process!), the resulting forecast should only be amended or adjusted in extreme circumstances. As shown in the special events section of our study, a two stage decomposition approach to special events (Webby et al., 2005) does not lead to a great deterioration in forecast accuracy or consistency. The key seems to be to find an acceptable weighting system and to stick to it—even in situations characterized by special events. As Armstrong (2001b, p. 433) comments ‘For combining to be effective, one should have independent forecasts that are systematically combined. When this is not the case, combining is expected to have little value.’

There are several limitations to our work. First, the experiment focused on task learning. Hence, the results of this experiment only generalize to learning a new task. There are also limitations imposed by the use of the task cue weights of 0.5, 0.26, and 0.24. In particular, people often have difficulties with tasks that have negative weights or that are non-linearly related to the forecasted variable (Sniezek & Naylor, 1978). Thus, our conclusions are untested with such problems. However, positive weights do characterize a wide variety of
forecasting tasks, so the use of these weights in the experiment does not greatly limit the generalization of the results. One of the obvious limitations of our study is that it is clearly an abstraction from the rich environment of sales forecasting. Importantly, immediate trial-by-trial feedback does not characterize practical forecasting situations, and our results need to be understood in that light. An interesting future study may be to examine the way people react to changing their weights when immediate feedback is not given or delayed. Lastly, we did not include an experimental condition where people received both task and cognitive feedback as well as bootstrap forecasts.

CONCLUSIONS

This study has demonstrated that Task and cognitive feedback, which allows a person to compare the actual and optimal weights, was most beneficial in a forecast combination task. But simply providing bootstrap forecasts in an effort to improve consistency did not produce the expected result. The task used in this study is representative of many businesses not only where a number of departments or individuals make estimates of sales for the next period, but also where a final forecast needs to be set in concrete based on these individual forecasts. Because information is asymmetrically distributed amongst the individuals and departments (the marketing department will arguably have more information on the market than a production department), a simple average will not be beneficial. People will need to utilize a differential weighting system to best integrate these forecasts. They need to be able to appropriately weight each forecast, and they need to do it consistently.

REFERENCES


