Division of Research
School of Business Administration

March 1992

MANAGERIAL JUDGMENT AND FORECAST COMBINATION AN EXPERIMENTAL STUDY

Working Paper #680

Sunil Gupta
The University of Michigan

Sunil Gupta, School of Business Administration, University of Michigan, Ann Arbor, Michigan 48109-1234, (313) 764-6355

ABSTRACT

This paper examines the role of managerial judgment in forming a final forecast, or judging the achievability of a critical level of sales, when multiple forecasts or opinions are available to the decision maker. Several factors which can help improve the quality of human intervention are identified and incorporated in a decision aid. Experimental results show that aided combination can help the decision maker exploit her relevant private information, and mitigate the generally observed negative effects of human intervention. Further, the results suggest that emphasizing expected sales, even when the organization is primarily interested in achievability accuracy, helps improve performance. Several suggestions for future research are presented.

Keywords: Combining forecasts, judgment, decision making, decision support systems

"Neither hand nor mind alone, left to themselves, amount to much; instruments and aids are the means to perfection" *Francis Bacon*

INTRODUCTION

A common managerial task facing a decision maker (DM) is to forecast the level of sales that can be expected in a future period (Dalrymple 1987). Also, in situations where a GO/NO-GO decision is required, the DM must judge whether a critical level of sales can be achieved. In many such situations the DM is likely to have multiple forecasts/opinions about the expected level of sales. For example, when considering the introduction of a new product on a regional roll-out basis, forecasts for a particular region could be based upon results of a test market at a representative site (not necessarily in that region), sales force surveys of the retailers in the region, and possibly, the internal records about the company's sales of other products in the same region. Typically, these forecasts will not be in complete agreement. The DM must now decide what the expected level of sales may be and/or whether a pre-determined critical level of sales can be achieved in this region.

Faced with such a situation the literature suggests that, in the interests of accuracy, the DM should use any one of several statistical models for combining forecasts (see Clemen 1989 and Moriarty 1990), or even a simple-average heuristic. However, several authors (Ang and O'Connor 1991, Chunglo 1985, Mahmoud 1989, West and Harrison 1989) have noted that, in practice, the DM is likely to use her judgment in the process of arriving at a final forecast. That is, it is quite unlikely that when faced with differing opinions the DM will leave the final resolution entirely up to a mechanized combination rule. Further, it may well be that the DM has situation-specific information, not accounted for by the statistical models. Ignoring such information completely may forego real improvement in forecast accuracy. But, when considering the use of judgment in determining the combined forecast, one is confronted with the vast psychological literature on the various inconsistencies and biases

in human judgments (see Fischoff 1988, Hogarth and Makridakis 1981, Makridakis 1988, Tyebjee 1987). The overall conclusion of this growing literature is that human judgments are, on average, inferior to formal data-based analyses.

In this paper, human intervention in the forecast combination process is taken as a given. As Moriarty (1985) states, "Finding that systematic methods are superior, however, is not helpful for an organization that wishes to improve its forecast performance and yet chooses to depend on management judgment methods." Consequently, what is at issue is not that judgment is better or worse than using a data-based approach. Rather, following Bunn and Wright (1991), the objective is to examine whether there are ways in which the advantages and disadvantages in each approach can be resolved by allowing structured interaction of judgmental and data-based combination methods. Thus, the major objectives of this paper are to: *i)* examine the literature to identify conditions under which the use of judgment is likely to prove beneficial; *ii)* design and experimentally test a decision aid which helps the DM exploit the advantages of her special knowledge, and guards against the weaknesses of human judgment; *iii)* examine whether the nature of the DM's task, forecasting the expected level of sales versus judging the availability of a critical level of sales, affects the accuracy.

LITERATURE REVIEW

Combination of Forecasts

The literature on combining multiple estimates of an uncertain quantity has developed very rapidly in recent years (Clemen 1989, Gupta and Wilton 1987, Moriarty 1990, Schmittlein, Kim and Morrison 1990). Various schemes for combining forecasts have been

proposed. The basic formulation assumes a DM facing k forecasts, f_1 , f_2 , ..., f_k of some uncertain variable y. Most approaches yield a combined forecast of the form:

$$f_C = w_1 f_1 + w_2 f_2 + ... + w_k f_k \qquad (1)$$

The weights, w_i, are usually estimated on the basis of past performance of the constituent models with the objective of minimizing a loss function. (Hereafter, we shall refer to these w_i as the *model weights*.) Among the most commonly studied methods are variants of the minimum-variance approach (Bates and Granger 1968, Bordley 1982, Clemen and Winkler 1986, Granger and Ramanathan 1984, Winkler 1981). These methods rely on the covariance matrix of past errors to derive combination weights such that the squared error of the combined forecast is minimized. The crucial reliance of these methods on the error covariance matrix, however, has been found to result in very unstable and/or unreliable estimates of model weights, resulting in poorer performance in the holdout sample. These problems manifest themselves most prominently when past performance data are sparse, the error generating process is non-stationary or, heteroscedastic (for cross-sectional data), or the errors of the constituent forecasts show high positive correlation. Because of these problems, some have suggested operational approaches which can help improve accuracy (e.g. Gupta and Wilton 1988, 1987; Moriarty 1990; Schmittlein, Kim and Morrison 1990). Others have noted that a simple equal weighting heuristic often provides most of the benefits of combining (Makridakis and Winkler 1983).

Several approaches to incorporating *judgment* in the forecasting process have been examined in the literature. First, several authors (e.g., Carbone et. al. 1983, Edmundson, Lawrence and O'Connor 1988, Jenks 1983, Mathews and Diamantopoulos 1989, 1986, Reinmuth and Geurts 1972, Soergel 1983) have examined the effect of presenting the DM

with the model-based forecast, and then having the DM revise this forecast. The results of these studies, which do not involve the multiple forecast situation of interest in this paper, are mixed. Second, the DM could provide a subjective forecast which is then combined with the other forecasts using a data-based approach (e.g., Ashton and Ashton 1985, Blattberg and Hoch 1990, Lawrence, Edmundson and O'Connor 1986, Moriarty and Adams 1984). Results of studies using this approach have generally shown an improvement in forecast accuracy. Note that in this approach the DM, perhaps artificially, is not involved in the final resolution of the differences. Third, in two studies (Flores and White 1989, Lawrence, Edmundson and O'Connor 1986) the DM is presented with the k constituent forecasts and forms the combined forecast judgmentally. One found an improvement, the other did not. Finally, an untested approach recommended by Makridakis (1989) is one where the DM is presented with the model weights, and judgment is used to *modify* these weights and arrive at the combined forecast. The decision aid developed later in this paper essentially follows Makridakis' (1989) suggestion, and we shall refer to the judgmentally modified model weights as *modified weights.*

The objective of the rest of this literature review is to identify specific factors which can be expected to enhance the quality of judgmentally modified forecasts by appropriate inclusion in the decision aid. While the focus of the review is on the forecasting literature, we also draw from various areas of psychological research on human judgment and decision making.

Relevant Situation-Specific Information

In most studies comparing judgments with statistical models, the information available to the DM is exactly the same as that available to the model. Consequently, as Blattberg and Hoch (1990) point out, the DM's "cannot take advantage of any skill they have at identifying other information not incorporated in the model." In many practical forecasting tasks, on the other hand, it is possible that the DM will have access to situation-specific information not incorporated in the model (because it occurs only rarely), which could be relevant to improving accuracy.

Several studies have reported the advantages of using judgment when the DM has access to such "broken-leg cues" (Meehl 1954). Johnson (1988) found that with access to broken-leg cues, experts were able to outperform their own bootstrapped models (which ignored the broken-leg cues). In forecasting studies, Soergel (1983) and Jenks (1983) pointed out that only judgment could reasonably anticipate one-time events such as extraordinary competitive developments. Reinmuth and Geurts (1972) showed that when an unusual event, like a promotion, occurred judgmentally revised time-series sales forecasts substantially increased forecasting accuracy for the atypical period. Edmundson, Lawrence and O'Connor (1988) found that for those products about which the DM might be expected to have important non-time-series information, judgmental revision was particularly helpful in improving accuracy. Mathews and Diamantopoulos (1989, 1986) suggested that individual expertise in the form of situation-specific knowledge was probably the key element leading to the improved accuracy of judgmentally revised forecasts. When such private information was not available to the DM, Carbone et.al. (1983) found that

judgmentally revising a model-based forecast led to a loss in accuracy. It is important to note that none of these studies examines the case of a DM faced with multiple forecasts.

Several studies have examined the usefulness of combining judgmental forecasts with model-based forecasts. The general conclusion is that such combination helps improve accuracy because of the additional information contained in the DM's judgment (Ashton and Ashton 1985, Blattberg and Hoch 1990, Lawrence, Edmundson and O'Connor 1986, Moriarty and Adams 1984). However, the role of the DM's judgment in arriving at the final combined forecast itself is not examined in these studies.

In the two studies that have explicitly examined the role of judgment in combining forecasts, the results are mixed. Flores and White (1989) reported improvements through subjective combination, while Lawrence, Edmundson and O'Connor (1986) found that simple averages were more accurate. Neither study examined the specific role of relevant situation-specific information. This examination is important because if the DM's private knowledge is actually irrelevant (or has already been accounted for by the statistical model), relying on the DM's faulty perceptions of the environment could actually impair forecasting accuracy. As Bunn and Wright (1991) and Chakravarti, Mitchell and Staelin (1981) caution, untested *conventional wisdom* can be at best irrelevant and at worst misleading. Thus, the question of how judgment should be incorporated in the process of combining forecasts in a manner which enables the DM to take advantage of relevant situation-specific information, remains.

Wary of the generally noted fallibility of human judgment, Makridakis (1989) suggests that by using the DM's private information to modify the model weights for each component forecast "a way can be found to elicit the judgment and knowledge of a DM while still harnessing the advantage of using objective and consistent approaches to

forecasting." For example, based on the DM's situation-specific information a forecast based on a sales-force survey of retailers could be weighted more heavily for forecasting sales in one region, while test market projections could receive a higher weight in another region. To the extent that the DM's knowledge is relevant, the resulting combined forecast could actually outperform a model- or heuristic-based combination which, ignoring such information, uses the same set of model weights for both regions.

The true relevance of the DM's situation-specific information will not necessarily be apparent a-priori. Consequently, it is important to design and test a decision aid which will implement Makridakis' suggestion in a manner which allows the DM to exploit truly relevant situation-specific information, and will help mitigate the consequences of faulty perceptions.

Feedback

There is general agreement in the literature that the accuracy of judgmental prediction can be enhanced if the DM's learning of the predictive relation can be improved. Several authors (see, for example, Brehmer 1987, Castellan 1974, Goldberg 1968, Makridakis 1988, and O'Connor 1989) have suggested that learning can be improved by providing adequate feedback. Three basic types of feedback have been examined in the literature: *i) outcome feedback* - providing the correct answer (e.g., the actual sales for some previously forecasted sales, or whether the critical level of sales was actually achieved or not); *ii) task environment feedback* - in the forecast combination context, such feedback would provide the subjects with the data-based weights (determined by using a forecast combination model) and the resulting combined forecasts; *iii) subjective performance feedback* - in the combination context, such feedback would inform the DM about the weights placed by him/her on the

component forecasts. Much research indicates that outcome feedback is usually ineffective (Brehmer 1980), especially if the subjects cannot refer back to previous results and are expected, instead, to remember them (Goldberg 1968, Hogarth 1989). Both task environment and subjective feedback have been found to be beneficial (Brehmer 1987, Castellan 1974, Hammond, Summers and Deane 1973), though the evidence is not conclusive. Clearly, further examination of the effects of feedback in forecast combination tasks is needed.

Motivation

Several authors have suggested and found a relation between performance, and the DM's assessment of the pay-offs from being correct and the processing costs of the various decision making strategies available. Expanding upon a contingency model proposed by Beach and Mitchell (1978), Christensen-Szalanski (1980, 1978) showed that if the DM's did not particularly care whether their forecasts were accurate, and/or if the use of the 'optimal' strategy was too difficult or time consuming, DM's chose to use normatively 'suboptimal' strategies. Harkness, DeBono and Borgida (1985), in a different context, found that the ability to detect contingent relationships between task variables was positively related to the DM's personal stake in the task. Similar results have been reported by Corbin (1980), Huffman (1978), and Smith, Mitchell and Beach (1982). In the forecast combination context, processing costs can be reduced by providing a decision aid which makes it possible for the DM to easily assess the consequences of different weighting policies on the resulting forecast (see Brehmer 1987), and where applicable, the achievability judgment. For experimental studies of judgmental combination the above results also imply the need for significant pay-offs appropriately tied to forecast or achievability judgment accuracy.

Response Mode

In many forecasting situations, the primary interest of the DM may not be in producing an accurate forecast of expected sales, but rather in predicting the achievability of some critical level of sales. The difference between forecasting expected sales and making a GO/NO-GO decision is akin to that between preference judgments and choice. Einhorn and Hogarth (1981) suggest that while preference judgments generally aid choice, they are neither necessary nor sufficient. In the forecast combination context DM's *may* first combine the available forecasts, and then compare them to the critical level of sales. Or, they may merely compare the various forecasts directly against the critical level of sales; without formally attempting to assign combination weights, determining the combined forecast, and *then* comparing it with the critical level of sales. While much of the early research either dismisses (Slovic and Lichtenstein 1971), or disregards (Dawes and Corrigan 1974) this distinction, Einhorn and Hogarth (1981) suggest that focussing on the preference judgment (expected level of sales) may lead to a more deliberative process of reasoning and evaluation of evidence. Billings and Scherer (1988), in a rare study comparing the effects of the particular response mode emphasized (preference or choice), found that requiring only an explicit choice decision produced undesirable effects on information search and processing. They went on to suggest that those designing decision aids "would be well advised to carefully consider the choice of response mode." The important question, in the forecast combination context, is whether motivating the DM to produce accurate forecasts of expected sales, even when the primary interest centers on GO/NO-GO decisions, can produce more accurate assessments of the achievability of the critical level of sales? The existing research does not provide any results in this regard.

In summary, against the background of a generally pessimistic appraisal of the role of human judgment in prediction tasks, the preceding discussion suggests that the quality of judgmentally combined forecasts *may* be improved by:

i) designing a decision aid which: *a)* provides task-environment feedback (model weights) and subjective performance feedback (the subjective weights being used by the DM); and *b)* reduces the processing costs by allowing the DM to easily evaluate the consequences of different weighting policies on the combined forecast and achievability judgment.

ii) focussing the DM's attention on providing accurate expected sales forecasts even when the primary interest is in achievability judgments.

The review also suggests that the extent to which a decision aid with the general characteristics listed above will in fact prove helpful will depend on the relevance of the situation-specific information available to the DM, and on the manner in which payoffs are tied to forecast accuracy. In the next section we describe the experimental scenario and a decision aid which attempts to provide a favorable environment for the systematic use of judgment in combining forecasts.

THE DECISION AID

To aid understanding of the features of the decision aid, we first provide a brief description of the experimental forecasting scenario for which it was designed.

Subjects were informed that they were to play the role of a marketing manager for a large manufacturer of consumer products. The company marketed its products in 30 districts nationwide. For introducing new products, the company could decide to do a

national launch, roll-out district by district, or abort the product. To be able to make these decisions the marketing manager obtained three sales forecasts for each district based on: *i)* consumer surveys and testing, *ii)* sales force surveys of retailers, and *iii)* internal information from company records. Given these three forecasts the marketing manager was expected to determine: *a)* the levels of sales that could be expected in each of the districts; and *b)* whether a pre-determined critical level of sales could be achieved. (The informational structure of this scenario is quite similar to Case 1 in Moriarty 1990, p 410).

The decision aid was implemented on IBM compatible personal computers. On the main screen shown in Figure 1, the top panel showed the three component forecasts and their average. As mentioned above, equal weighting is a commonly used heuristic which performs reasonably well in many forecast combination situations.

The middle panel provided task-environment feedback by showing the model weights, derived by using Gupta and Wilton's (1988, 1987) odds-matrix (OM) approach, and the resulting combined forecast. As the subjects progressed from district to district, the weights were updated on the basis of the performance of each forecast up to that point. The OM approach was chosen because it has been shown to be particularly accurate when the available data are sparse (as is the case in the early stages of the experiment) (Gupta and Wilton 1988, 1987, Mascarenhas and Sand 1989), and because the derived weights are never negative, as can be the case when using an approach such as Winkler's (1981) minimum-variance approach. A pre-test showed that using the minimum-variance weights caused considerable confusion among the subjects. The weights were displayed using color-coded bars and numerically.

The third panel of the main screen showed three color-coded bars corresponding to the weights used by the DM for the most recent forecast, thus providing subjective performance feedback. Subjects could change the weights in this panel by adjusting the bar lengths. As the bar lengths were adjusted, the implied modified weights and the resulting combined forecast were continuously displayed. Alternately, subjects could switch to a pairwise comparison screen which asked them to provide the odds of forecast i outperforming forecast j for the current district. The implied combination weights were calculated using the OM approach, and the resulting bar lengths and forecast were shown in the third panel of the main screen. Some authors (Armstrong et. al. 1975, Edmundson 1990) have found that eliciting decomposed, rather than global, judgments improves the quality of the information elicited. The pairwise comparison screens permitted such decomposition. However, the subjects were not *required* to use the pairwise approach, primarily because after the first few forecasts, the task tends to become somewhat tedious. Finally, the DM was given the option of accepting the combined forecast based either on model weights (in the middle panel), the modified weights (in the bottom panel), or enter any other value.

As part of the cover story, subjects were told that due to competitive pressures the company had recently introduced a product nationally before completing the normal decision making procedures described above. Thus, the subjects were told, the three input forecasts, the critical level of sales and the actual sales for each of the thirty districts were available for the experiment. This setup allowed credible presentation of feedback to the subjects. At the touch of a single function key (indicated at the bottom of the main screen),

subjects could bring up a screen of histories of their own and each forecast's past errors, and hits and misses in judging achievability of critical sales.

THE EXPERIMENT

Manipulations

Situation Specific Information (Info): Recall that the scenario involved three types of districts, those with low (range: 2,500-9,977, average: 5943), medium (range: 10,050-38,917, average: 22,967), and high levels (range: 39,523-51,657, average: 45,957) of expected sales. Subjects were told that actual sales had exceeded critical sales in some of the districts and fallen short in others. Therefore, the subjects had no information about whether the introduction had generally been successful or not. Finally, they were told that it was *possible,* though *not necessarily true* that some of the input forecasts were more accurate when predicting one level of sales than another.

The relevance of this situation-specific information was manipulated by generating input forecast errors based either on a single error covariance matrix for all types of districts *(Irrelevant Situation-Specific, ISS),* or on the basis of separate error covariance matrices for each district type *(Relevant Situation-Specific, RSS).* Because the model weights depend on the error covariance matrix, subjects' hypothesis that the appropriate weights changed with district type was false for *ISS,* but was true for *RSS.* For both conditions, the model weights shown to the subjects in the middle panel of the main screen of the decision aid were not adjusted for district type. Thus, a DM could benefit from the situation-specific information by modifying weights, but only in the *RSS* condition. In the *ISS* condition, relying on this hypothesis should impair forecast accuracy. Table 1 shows the error covariance and

correlation matrices, the model weights based on the OM approach, and for comparison, Winkler's (1981) minimum-variance approach.

Combination Process (Combo): To test the efficacy of the decision aid, half the subjects used the decision aid and the other half did not. The procedure for those who used the decision aid to modify the model weights *(MMW)* has already been described. For those who did not have access to the decision aid and combined entirely on the basis of their judgments *(JC)*, Figure 2 shows the main screen. The average of the three input forecasts was still provided, because this is a commonly used heuristic. Also, the history screens were available. Based on the literature review, we should expect more accurate forecasts for the *MMW* condition. Further, because of better feedback, subjects in the *MMW* condition should be able to detect the relevance of the situation-specific information better. Consequently, these subjects should be more able to exploit their private information.

Task Emphasis (Task): Based on the three forecasts, the subject's task was to predict the expected level of sales, *and* the achievability of a pre-determined critical level of sales. However, the relative importance of these tasks were manipulated.

In both cases subjects were told that they could expect to win as much as fifty four dollars, depending on the accuracy of their forecasts. The accuracy of an expected sales forecast was measured as the absolute deviation between the subject's forecast and actual sales, summed across the thirty districts. The accuracy of the achievability judgment was assessed as Net Hits = (Number of Hits - Number of Misses).

The desired emphasis was manipulated by changing the judgment asked for first, and by altering the payoff structure. To emphasize the expected sales forecast *(ES)*, subjects were *first* asked to forecast the expected level of sales. *Then*, the critical level of sales for the

district was revealed and they were asked to judge whether it could be achieved. Note that subjects were unaware of the critical sales when predicting the expected sales. Also, thirty-six of the possible fifty-four dollars were dependent on the accuracy of the expected sales judgment. To emphasize the achievability judgment *(AJ)*, subjects were given the critical level of sales, asked to judge whether it could be achieved, and then asked to forecast the expected sales. In this case, a perfect score of thirty hits was worth thirty-six dollars. Pre- and post-experimental checks showed that the manipulation was successful in heightening subjects' concern regarding the emphasized judgment. Creyer, Bettman and Payne (1989) used a similar payoff approach to emphasize accuracy and effort goals in a multiattribute alternatives experiment.

Based on the literature review, subjects emphasizing sales *(ES)* should be more deliberative and process the information more appropriately. Consequently, they should be more accurate in their predictions of both expected sales and achievability of critical sales.

Experimental Procedure

Subjects were recruited by posting announcements in prominent locations in two business schools. Ten graduate students (MBA's and PhD's), with background in statistics and/or forecasting, were recruited for each of the eight cells (2 *Info* × 2 *Combo* × 2 *Task*), for a total of eighty subjects. Each participant was randomly assigned to one of the eight cells and provided with printed material to study before coming for the experiment. The printed materials described the scenario and provided basic information about the usefulness of combining forecasts. When subjects arrived for the experiment they were asked if they had any questions. After these questions had been answered, they were seated at a computer terminal. They were encouraged to ask questions whenever needed.

Before starting the actual experiment, subjects were given the opportunity to go through fifteen trial forecasts. The trial forecasts were created to resemble the actual error covariances as closely as possible. Subjects could continue in the trial phase as long as they needed. The trial forecasts permitted the subjects to familiarize themselves with the use of the computer, to assess whether the accuracy of the forecasts indeed changed with district types, and to form priors about the three forecasts.

At the end of the trial phase, their priors about the accuracy of the constituent forecasts in each of the three district types were elicited. Then the program proceeded to the thirty actual districts, ordered randomly. In both phases, subjects could get context-sensitive help regarding every element displayed on the screen. Also, at the end of each forecast, their current payoff and the payoff they would receive if they continued to forecast at the current accuracy level was displayed. Pre-tests showed that providing this information served to increase their involvement with the task. After forecasts and judgments had been elicited for each of the thirty districts, some post-experimental questions were asked, and then the screen displayed the final score and payoff. They were paid and thanked for their participation. On average, subjects spent between 1-½ to 2 hours to complete the task and earned $27.43. Post-experiment tests revealed that they found the task involving.

RESULTS

Accuracy of Expected Sales Forecasts

The error variances of the three input forecasts in the two *Info* conditions, *ISS* and *RSS* were not equal. Consequently, it is not meaningful to compare the subjects' forecasting errors across the two conditions. To provide some comparability, each subjects' mean absolute error

was divided by the mean absolute error that would have resulted from using the OM model weights. For the *RSS* condition, the model weights used to form the baseline combined forecasts and the resulting errors were calculated on the basis of a *single* odds-matrix. The separate sets of weights, shown in Table 1, for high, medium and low sales districts were not used. Thus, the situation-specific information available to the subjects in the *RSS* condition *could* help them outperform the baseline OM forecasts which do not utilize this information.

Figure 3 shows the average ratio of errors for each of the eight cells of the experiment. Numbers smaller than 1 imply that the subjects' observed error was less than would have resulted from using the baseline OM weights. The results show that for each of the four *Info* and *Task* pairs, subjects using the decision aid *(MMW)* outperformed those forming judgmental combinations *(JC)*. Further, across the two *Info* conditions, only subjects using the decision aid were able, always, to perform at least as well as the baseline forecast. Thus, the decision aid is helpful. However, the ability to *outperform* the baseline forecasts also depends on the relevance of the DM's situation-specific information. In the *ISS* condition, the best the subjects could do was to be almost as good as model forecasts. But in the *RSS* condition, with the decision aid available, they could actually outperform the less well informed baseline forecasts.

The results also show a significant interaction between *Combo* and *Task*. The task emphasis does not matter for those subjects using a decision aid. However, for those forming judgmental combinations, emphasizing the expected sales *(ES)* helps improve accuracy in both *Info* conditions.

Together, these results imply that: *i)* introducing judgment into the forecast combination task can improve accuracy only when the DM has some relevant situation-specific information, not available to the baseline model; *ii)* however, if judgment is going to be introduced anyway, using a decision aid, such as that developed in this paper, never impairs accuracy and may even help outperform the baseline model; and *iii)* emphasizing the expected sales forecast is helpful only when the DM does not have access to a decision aid. This last result is somewhat at variance with expectations based on the literature review. When the decision aid is available, task emphasis does not matter. This may have been because using the decision aid causes the DM to engage in more detailed analysis, yielding most of the possible improvement. The additional improvement from emphasizing expected sales is, therefore, negligible.

Bias in the Expected Sales Forecast

Tables 2a and 2b show the biases in the subjects' forecasts of the expected sales. The input forecasts were unbiased, and consistent combination weights should have resulted in unbiased forecasts. The results in Table 2a show that on average there were large positive biases in the sales forecasts, though compared to the total error they were not inordinately large. The biases are much larger when *Info=RSS*, probably because the average error of the input forecasts in this condition was larger. The table also shows that having the decision aid helps reduce the bias. Table 2b allows a more detailed analysis of the biases. Here we note a tendency to overestimate sales when they are expected to be high and to underestimate them when they are expected to be low. This pattern is similar to that observed by Mathews and

Diamantopoulos (1989). Finally, the table also shows that having the decision aid is only of limited help in mitigating these biases.

Accuracy of Achievability Judgments

Figure 4 shows the results for accuracy of achievability judgments. The cell means shown in the figures are the subjects' net hits = (number of hits - number of misses) divided by the net hits that would have resulted from using the baseline OM forecast. In this figure higher cell means imply greater accuracy. (Other measures of discrimination skill which examine the pattern of hits and misses in more detail have been suggested in the literature, e.g., Birnbaum and Mellers 1983, Murphy 1973, Yaniv, Yates and Smith 1991. However, for this experiment, these measures were monotonically related to net hits. Because subjects' payoffs were based on net hits, that measure is reported here.)

Some of the results are similar to those for expected sales accuracy. For example, subjects with decision aids *(MMW)* usually outperformed and never did worse than those without *(JC)*. Further, compared to the baseline forecast, subjects with the decision aid never did significantly worse and the improvements were most pronounced when the DM had relevant situation-specific information (the *Info*×*Combo* interaction is significant).

However, there are some important differences. First, in the *ISS* condition, subjects using the decision aid were able to outperform the baseline forecast. For expected sales, none of the *ISS* cell means were better than the OM forecast. Thus, the task emphasis and use of decision aid seem to be more helpful when the primary interest is in achievability accuracy. Second, compared to the baseline, the worst performance was for judgmental combination under the more complex *RSS* environment. For expected sales accuracy, the worst

performance was under the *ISS* condition. These differences imply that the decision aid does help exploit the advantages of using the DM's private information and guards against the drawbacks of faulty perceptions. Together, the results also show that emphasizing expected sales matters, especially when the primary interest is in achievability accuracy. While we had speculated about this result, the existing literature had not examined this possibility. Our result implies that the distinction between the two tasks can have real implications for decision making, one that the forecasting and marketing research literatures need to pay more attention to than has been the case in the past.

DISCUSSION

The evidence from much of the past literature on judgment and decision making suggests that when models can be used, human intervention generally impairs accuracy. In this paper, by contrast, we have argued that some effort should also focus on what DM's can do under more favorable conditions. Increasingly the need for such research has been emphasized in the literature (e.g., Bunn and Wright 1991, Hogarth 1989, Mahmoud 1989, Makridakis 1981, Moriarty 1985, Phillips 1987). Our literature review identified several factors which can help improve the quality of human intervention in the forecasting process. These factors were then incorporated in a decision aid. The experimental results, in agreement with the weight of past literature, showed that unaided combination can indeed be inferior to baseline model forecasts, especially when the decision maker does not have an informational advantage over the model. But, perhaps more importantly, the results also showed that using the decision aid helped mitigate such accuracy losses, and actually helped the DM exploit the

private information when it was relevant. Since the DM may often be expected to be unaware of the true relevance of the private information, using a decision aid such as that developed here is highly advisable. Future research should also examine the causes for, and approaches that might help mitigate the large positive biases for forecasts of high levels of sales and negatives biases for low levels observed in this study.

Much of the past literature has also ignored the importance of the particular task emphasized on the accuracy of the resulting forecasts. Our results show that emphasizing expected sales versus achievability judgment generally resulted in greater accuracy of both types of forecasts. Consequently, it is advisable that forecasting systems build procedures and provide incentives for the DM to provide the most accurate point forecast, even when the primary organizational interest centers on the achievability judgment.

The ultimate hope of our research, of course, is that building on the positive results reported here, it will be possible to construct decision aids which will in fact become an integral part of organization forecasting systems. Yet, several organizational constraints were not faithfully reflected in the experiment.

First, the most obvious shortcoming lies in the immediacy and high quality of the feedback provided to the subjects. In most real situations knowledge of actual outcomes is likely to be delayed and error-prone. The extent to which feedback delays and fallible outcome feedback hinder the quality of modified forecast combinations needs further examination. Second, the environment of the experiment was much more controlled than any real forecasting situation is likely to be. For example, the only "pressure" on the DM was that of the explicitly stated payoffs. In most organizational contexts the relationship between

performance and compensations is more murky. Also, several other pressures, political and otherwise, are likely to be present. The informational control in the experiment was also more extreme than is likely to be the case in real situations. Our subjects had only one piece of external information to cloud (or illuminate) their judgments; whether forecast accuracy actually changed with district types. Typically, there should be many more "broken-leg cues" available. With a less well-defined incentive structure and a multitude of possible cues to adjust for, will the DM still be able to take advantage of her private information?

Our defense for these shortcomings is twofold. First, even if the added complexity reduces or eliminates the gains in comparison to the baseline methods, there seems to be little reason to expect that the advantage over unaided judgmental combinations will also be lost. Since judgment will often be used anyway, the objective of beginning the task of identifying how best to use it has been advanced. Second, given the vast amount of literature arguing against the use of human judgment, often obtained in at least as controlled an environment, the mere demonstration that accuracy can be improved through careful human intervention is an important objective.

Finally, some important areas for future research deserve mention. First, our focus has been on accuracy. Thus, the extent to which using the decision aid or emphasizing the expected sales improved learning itself was not studied. Similarly, the effect of the various factors on the DM's confidence in the forecasts and the forecasting systems was not examined. These aspects are likely to prove important in the organizational implementation of decision aids and deserve more careful examination. Second, the decision aid incorporated all of the various factors identified through the literature search. Thus we do not know, for example,

how important task-feedback is relative to subjective-performance feedback. Identifying the particular sources of improvement should prove beneficial. Finally, the underlying reasons for the task-emphasis effects found in our experiment need to be explored. Three theoretical frameworks, suggested by Payne (1982), which may account for the differences have been suggested in the literature: a cost/benefit approach (e.g., Beach and Mitchell 1978), a framing approach (e.g., Tversky and Kahnemann 1981) and a production systems approach (e.g., Pitz 1977). The differences and overlaps among these approaches in terms of the DM's awareness of decision strategies, extent of learning, and effects of incentives should prove useful avenues for future research.

REFERENCES

Ang, S. and M. O'Connor (1991), "The Effect of Group Interaction Processes on Performance in Time Series Extrapolation," *International Journal of Forecasting*, 7, 141-149.

Armstrong, J.S., W.B. Denniston Jr. and M.M. Gordon (1975), "The Use of the Decomposition Principle in Making Judgments," *Organizational Behavior and Human Performance*, 13, 257-263.

Ashton, A.H. and R.H. Ashton (1985), "Aggregating Subjective Forecasts: Some Empirical Results," *Management Science,* 31(12), 1499-1508.

Bates, J.M. and C.W.J. Granger (1968), "The Combination of Forecasts," *Operations Research Quarterly,* 20, 451-468.

Beach, L.R. and T. R. Mitchell (1978), "A Contingency Model for the Selection of Decision Strategies," *Academy of Management Review*, 3, 439-449.

Billings, R.S. and L.L. Scherer (1988), "The Effects of Response Mode and Importance in Decision Making Strategies: Judgment and Choice," *Organizational Behavior and Human Decision Processes*, 41(1), 1-19.

Birnbaum, M.H. and B.A. Mellers (1983), "Bayesian Inference: Combining Base Rates with Opinions of Sources who Vary in Credibility," *Journal of Personality and Social Psychology*, 45(4), 792-804.

Blattberg, R.C. and S.J. Hoch (1990), "Database Models and Managerial Intuition: 50% Model + 50% Manager," *Management Science*, 36(8), 887-899.

Brehmer, B. (1987), "Social Judgment Theory and Forecasting," in G. Wright and P. Ayton (eds.), *Judgmental Forecasting*, Wiley: Chichester, 199-214.

——— (1980), "In One Word: Not From Experience," *Acta Psychologica*, 45, 223-241.

Bordley, R.F. (1982), "The Combination of Forecasts: A Bayesian Approach," *Journal of the Operations Research Society*, 33(2), 234-249.

Bunn, D.W. and G. Wright (1991), "Interaction of Judgmental and Statistical Forecasting Methods: Issues and Analysis," *Management Science*, 37(5), 501-518.

Carbone, R., A. Andersen, Y. Corriveau and P.P. Corson (1983), "Comparing for Different Times Series Methods the Value of Technical Expertise, Individualized Analysis and Judgmental Adjustment," *Management Science*, 29(5), 559-566.

Castellan, J.N. Jr. (1974), "The Effect of Different Types of Feedback in Multiple-Cue Probability Learning," *Organizational Behavior and Human Performance*, 11(1), 44-64.

Chakravarti, D., A. Mitchell and R. Staelin (1981), Judgment Based Marketing Decision Models: Problems and Possible Solutions," *Journal of Marketing*, 45(Fall), 13-23.

Christensen-Szalanski, J.J.J. (1980), "A Further Explanation of the Selection of Problem-Solving Strategies: The Effects of Deadlines and Analytic Aptitudes," *Organization Behavior and Human Performance*, 25, 107-122.

——— (1978), "Problem-Solving Strategies: A Selection Mechanism, Some Implications, and Some Data," *Organization Behavior and Human Performance*, 22, 307-323.

Chunglo, F.J. (1985), "Developing Sales Forecasting - Master Scheduling Software," *P & I M Review*, 5(2), 56-60.

Clemen, R.T. (1989), "Combining Forecasts: An Annotated Bibliography," *International Journal of Forecasting*, 5, 559-583.

——— and R.L. Winkler (1986), "Combining Economic Forecasts," *Journal of Business and Economic Statistics*, 4(1), 39-46.

Corbin, R.M. (1980), "Decision might not get made," in T.S. Wallsten (ed.), *Cognitive Processes in Choice and Decision Behavior*, Erlbaum: Hillsdale, N.J.

Creyer, E.H., J.R. Bettman and J.W. Payne (1990), "The Impact of Accuracy and Effort Feedback and Goals on Adaptive Decision Behavior," *Journal of Behavioral Decision Making*, 3(1), 1-16.

Dalrymple, D.J. (1987), "Sales Forecasting Practices: Results from a United States Survey," *International Journal of Forecasting*, 3, 379-391.

Dawes, R.M. and B. Corrigan (1974), "Linear Models in Decision Making," *Psychological Bulletin*, 81, 95-106.

Edmundson, R.H. (1990), "Decomposition: A Strategy for Judgmental Forecasting," *Journal of Forecasting*, 9, 303-314.

———, M.J. Lawrence and M.J. O'Connor (1988), "The Use of Non-Time Series Information in Sales Forecasting: A Case Study," *Journal of Forecasting*, 7, 201-211.

Einhorn, H.J. and R.M. Hogarth (1981), "Behavioral Decision Theory: Processes of Judgment and Choices," *American Review of Psychology*, 32, 53-88.

Fischoff, B. (1988), "Judgmental Aspects of Forecasting: Needs and Possible Trends," *International Journal of Forecasting*, 4, 331-339.

Flores, B.E. and E.M. White (1989), "Subjective versus Objective Combining of Forecasts: An Experiment," *Journal of Forecasting*, 8, 331-341.

Goldberg, L.R. (1968), "Simple Models or Simple Processes? Some Research on Clinical Judgments, *American Psychologist*, 23, 483-496.

Granger, C.W.J., and R. Ramanathan (1984), "Improving Methods of Combing Forecasts," *Journal of Forecasting*, 3, 197-204.

Gupta, S. and P.C. Wilton (1988), "Combination of Economic Forecast: An Odds-Matrix Approach," *Journal of Business and Economic Statistics*, 6(3), 373-379.

——— and ——— (1987), "Combination of Forecasts: An Extension," *Management Science*, 33(3), 356-372.

Hammond, K.R., D.A. Summers and D.H. Deane (1973), "Negative Effects of Outcome Feedback on Multiple-Cue Probability Learning," *Organizational Behavior and Human Performance*, 9, 30-34.

Harkness, A.R., K.G. DeBono and E. Borgida (1985), "Personal Involvement and Strategies for Making Contingency Judgments: A Stake in the Dating Game Makes a Difference," *Journal of Personality and Social Psychology*, 49, 22-32.

Hogarth, R.M. (1989) *Judgment and Choice: The Psychology of Decision*, 2nd edition, Wiley: Chichester.

——— and S. Makridakis (1981), "Forecasting and Planning: An Evaluation," *Management Science*, 27(2), 115-138.

Huffman, M.D. (1978), "The Effect of Decision Task Characteristics on Decision Behavior," *Technical Report 78-16*, Department of Psychology, University of Washington, Seattle, Washington.

Jenks, J.M. (1983), "Non-Computer Forecasts to Use Right Now," *Business Marketing*, 68, 82-84.

Johnson, E.J. (1988), "Expertise and Decision under Uncertainty: Performance and Process," in M.T.H. Chi, R. Glaser, and M.J. Farr (eds.), *The Nature of Expertise*, Erlbaum, Hillsdale, N.J.

Lawrence, M.J., R.H. Edmundson and M.J. O'Connor (1986), "The Accuracy of Combining Judgmental and Statistical Forecasts," *Management Science,* 32(12), 1521-32.

Mahmoud, E. (1989), "Combining Forecasts: Some Managerial Issues," *International Journal of Forecasting*, 5, 599-600.

Makridakis, S. (1989), "Why Combining Works?," *International Journal of Forecasting,* 5, 601-603.

——— (1988), "Metaforecasting: Ways of Improving Forecasting Accuracy and Usefulness," *Internal Journal of Forecasting,* 4, 467-491.

——— (1981), "Forecasting Accuracy and the Assumption of Constancy," *OMEGA: The International Journal of Management Science*, 9(3), 307-311.

——— and R.L. Winkler (1983), "Averages of Forecasts: Some Empirical Results," *Management Science*, 29(September), 987-996.

Mascarenhas, B. and O.C. Sand (1989), "Combining Forecasts in the International Context: Predicting Debt Reschedulings," *Journal of International Business Studies*, 20, 539-552.

Mathews, B.P. and A. Diamantopoulos (1989), "Judgmental Revision of Sales Forecasts: A Longitudinal Extension," *Journal of Forecasting*, 8, 129-140.

——— and ——— (1986), "Managerial Intervention in Forecasting: An Empirical Investigation of Forecast Manipulation," *International Journal of Research in Marketing*, 3, 3-10.

Meehl, P.E. (1954), *Clinical Versus Statistical Prediction*, University of Minnesota Press, Minneapolis.

Moriarty, M.M. (1990), "Boundary Value Models for the Combination of Forecasts," *Journal of Marketing Research*, 27(November), 402-17.

——— (1985), "Design Features of Forecasting Systems Involving Managerial Judgments," *Journal of Marketing Research*, 22(November), 353-364.

——— and A.J. Adams (1984), "Management Judgment Forecasts, Composite Forecasting Models and Conditional Efficiency," *Journal of Marketing Research*, 21(August), 239-250.

Murphy, A.H. (1973), "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology*, 12, 595-600.

O'Connor, M. (1989), "Models of Human Behavior and Confidence in Judgment: A Review," *International Journal of Forecasting*, 5(2), 159-169.

Payne, J.W. (1982), Contingent Decision Behavior," *Psychological Bulletin*, 92, 382-402.

Phillips, L. (1987), "On the Adequacy of Judgmental Forecasts," in G. Wright and P. Ayton (eds.), *Judgmental Forecasting*, Wiley: Chichester, 11-30.

Pitz, G.F. (1977), "Decision Making and Cognition," in H. Jungermann and G. deZeeuw (eds.), *Decision Making and Change in Human Affairs*, Reidel: Dordrecht, Holland.

Reinmuth, J.E. and M.D. Geurts (1972), "A Bayesian Approach to Forecasting Effects of Atypical Situations," *Journal of Marketing Research*, 9(August), 292-297.

Schmittlein, D.C., J. Kim and D.G. Morrison (1990), "Combining Forecasts: Operational Adjustments to Theoretically Optimal Rules," *Management Science*, 36(9), 1044-1056.

Slovic, P. and S. Lichtenstein (1971), "Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment," *Organization Behavior and Human Performance*, 6, 649-744.

Smith, J.F., T.R. Mitchell and L.R. Beach (1982), "A Cost-Benefit Mechanism for Selecting Problem Solving Strategies: Some Extensions and Empirical Tests," *Organizational Behavior and Human Performance*, 29, 370-396.

Soergel, R.F. (1983), "Probing the Past for the Future," *Sales and Marketing Management*, 130, 39-43.

Tversky, A. and D. Kahnemann (1981), "The Framing of Decisions and the Psychology of Choice," *Science*, 211, 453-458.

Tyebjee, T.T. (1987), "Behavioral Biases in New Product Forecasting," *International Journal of Forecasting*, 3, 393-404.

West, M. and J. Harrison (1989), "Subjective Intervention in Formal Models," *Journal of Forecasting*, 8, 33-53.

Winkler, R.L. (1981), "Combining Probability Distributions from Dependent Information Sources," *Management Science*, 27(4), 479-488.

Yaniv, I., J.F. Yates and J.E.K. Smith (1991), "Measure of Discrimination Skill in Probabilistic Judgment," *Psychological Bulletin,* 110(3), 611-617.

Table 1

Error Covariance and Correlation Matrices and Combination Weights

Covariance Matrix				*Correlation Matrix*		
			Info=ISS			
45301	42867	12844		1.00	0.78	0.38
42868	66239	7041		0.78	1.00	0.17
12844	7041	25359		0.38	0.17	1.00
			Winkler Weights	*0.07*	*0.19*	*0.74*
			OM Weights	*0.27*	*0.17*	*0.56*
			Info=RSS			
District Type: Low						
286585	232811	11891		1.00	0.86	0.06
232811	256448	56961		0.86	1.00	0.28
11891	56961	160147		0.06	0.28	1.00
			Winkler Weights	*0.50*	*-0.19*	*0.69*
			OM Weights	*0.19*	*0.26*	*0.55*
District Type: Medium						
4422033	777523	3167597		1.00	0.24	0.84
777523	2349561	1168287		0.24	1.00	0.43
3167597	1168287	3214853		0.84	0.43	1.00
			Winkler Weights	*0.21*	*0.66*	*0.13*
			OM Weights	*0.15*	*0.66*	*0.19*
District Type: High						
21054403	266646	5895217		1.00	0.01	0.35
266646	23165709	15747536		0.01	1.00	0.89
5895217	15747536	13440600		0.35	0.89	1.00
			Winkler Weights	*0.38*	*0.14*	*0.48*
			OM Weights	*0.24*	*0.27*	*0.48*

Table 2a

Bias in Forecast of Expected Sales

		Combo			
		JC		*MMW*	
Task		*AJ*	*ES*	*AJ*	*ES*
Info	*ISS*	480	400	15	-9
	RSS	2817	2759	2595	2559

Significant Main and Interaction Effects at the 0.1 Level: Info, Combo, Info×Combo

Table 2b

Bias in Forecasts by Type of District

		DISTRICT TYPE		
INFO	*COMBO*	High	Medium	Low
ISS	*JC*	633	282	-474
	ES	356	10	-362
RSS	*JC*	9358	-3987	-2793
	ES	8074	-3450	-1835

FIGURE 1
Main screen for Modified Model Weights (MMW)

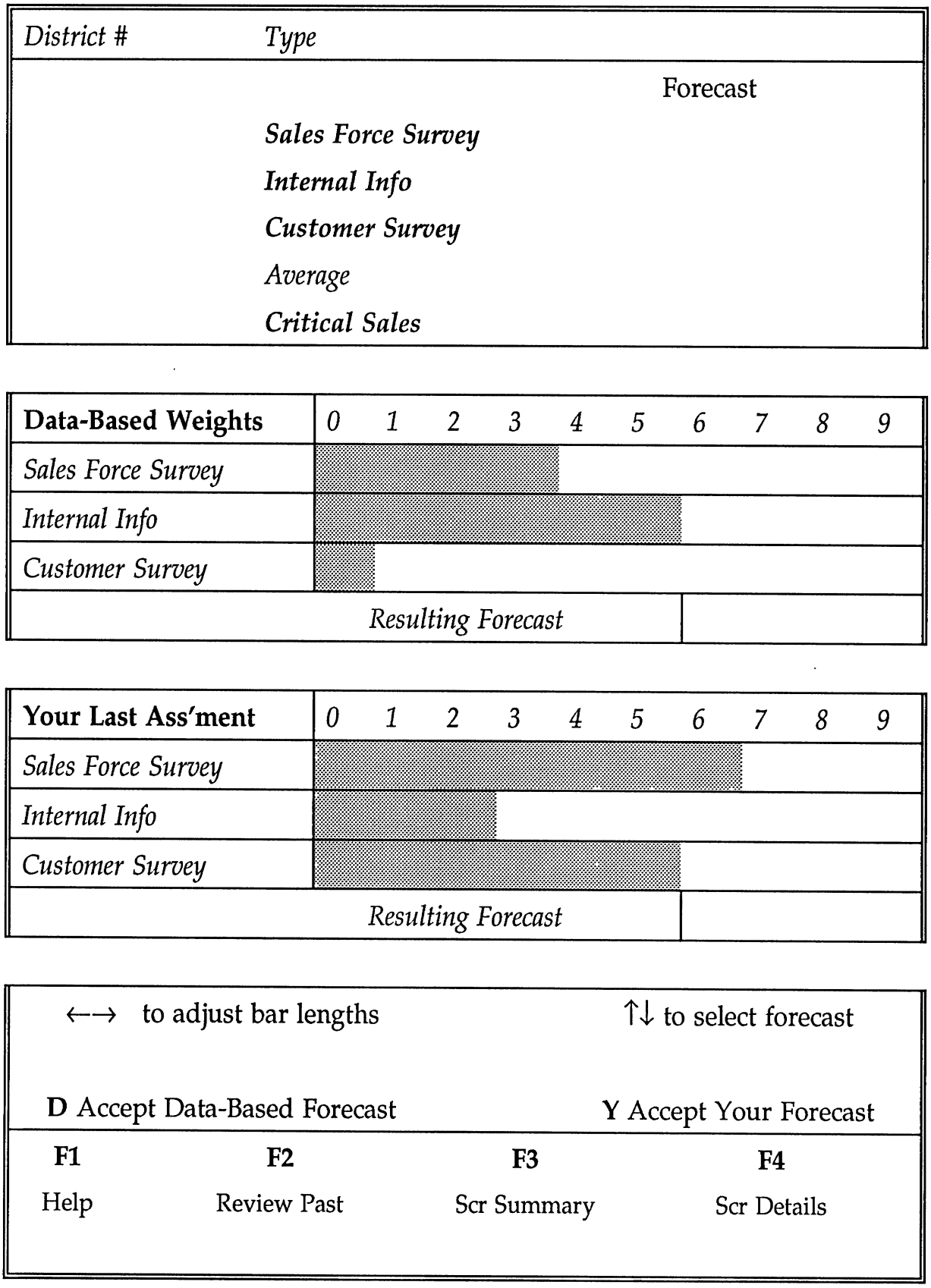


Figure 2

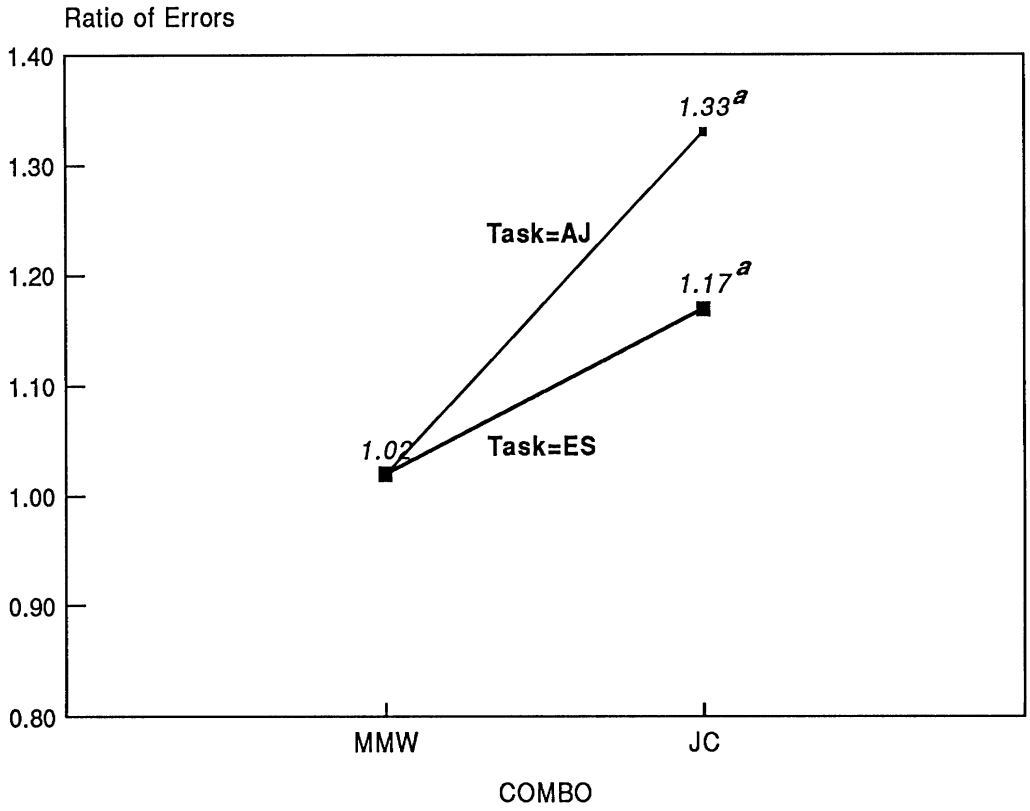
Main screen for Judgmental Combination (JC)

District #	*Type*	
		Forecast
	Sales Force Survey	
	Internal Info	
	Customer Survey	
	Average	
	Critical Sales	

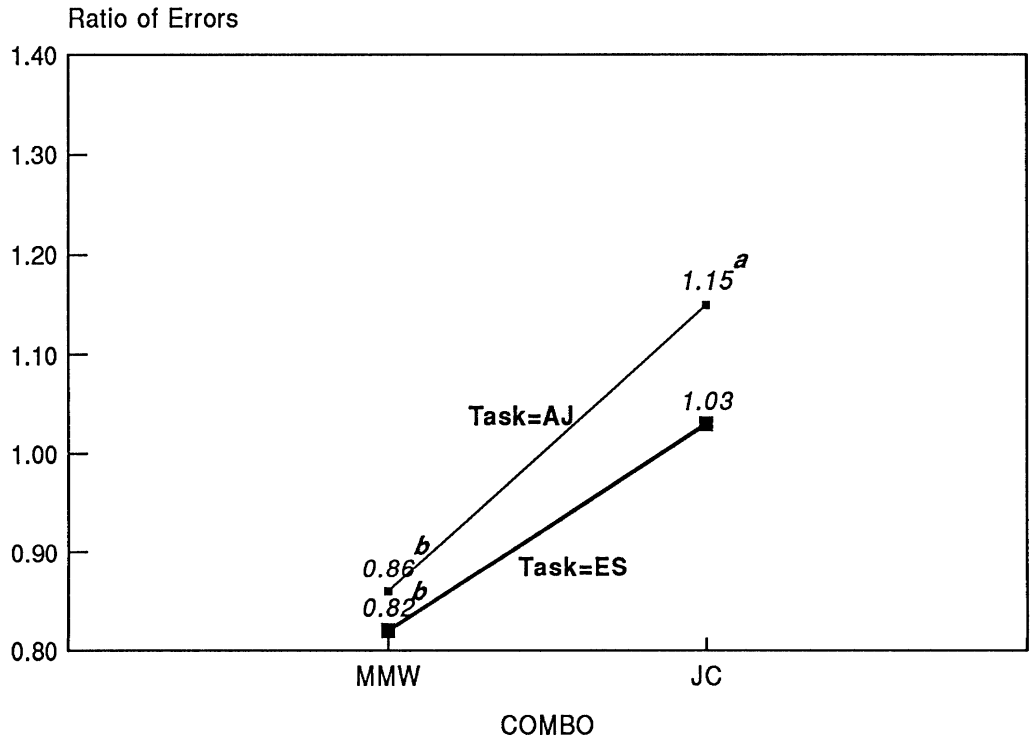
F1 Help	**F2** Review Past	**F3** Scr Summary	**F4** Scr Details

Figure 3

Expected Sales Accuracy: Info=ISS
Observed to Odds Matrix



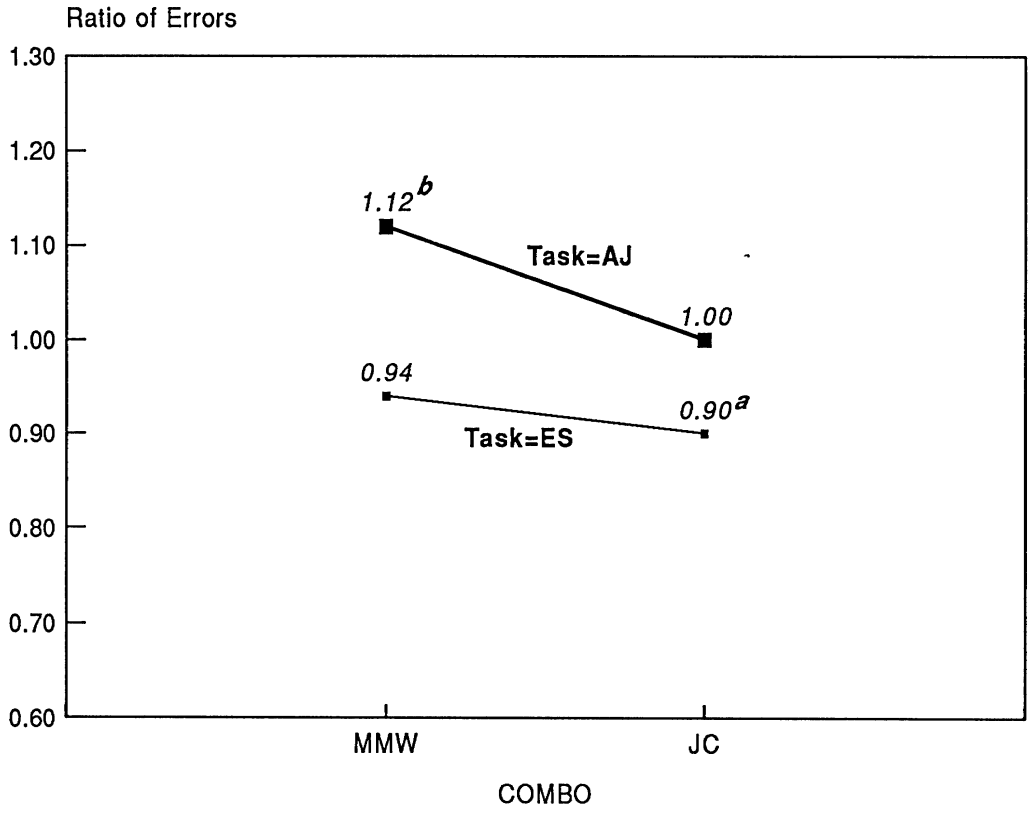
Expected Sales Accuracy: Info=RSS
Observed to Odds Matrix



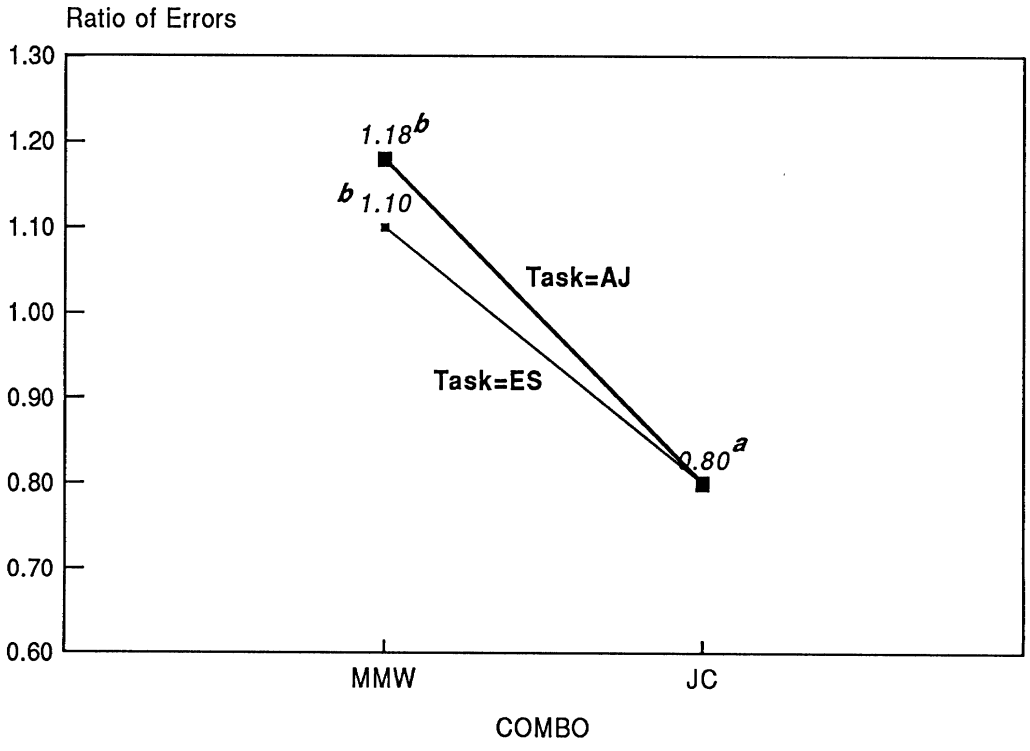
a - significantly worse than baseline OM forecast at 0.1 level
b - significantly better than baseline OM forecast at 0.1 level
Significant effects - Combo, Task, Combo*Task

Figure 4

Achievabilty: Info=ISS
Normalized Hit Score



Achievabilty: Info=RSS
Normalized Hit Score



a - significantly worse than baseline OM forecast at 0.1 level
b - significantly better than baseline OM forecast at 0.1 level
Significant effects at 0.1 level - Combo, Task, Combo*Task