Journal of the Korean
OR/ MS Society
Vol. 21, No. 3, December 1996

Combining Judgments for Better Decisions: A Study for Investigating Effective Combining Schemes*

HoonYoung Lee**

Abstract

Facing decision-making tasks, managers frequently make judgments. However, since managers are human beings, the efficiency of their judgments is limited. Two major sources of inefficiency in their judgments have been recognized: one is systematic deviations from normatively preferred decisions, so called bias or incorrect intuition, and the other is inconsistency in their judgments, i. e. erratic decision making variance. Rather than bias, variance is really expensive or damaging. Thus, if the inconsistency in managers judgments is removed, performance could be by far improved by virtue of the reduced random variance.

One of the approaches to improve managerial judgment is to simply bring managers together by effectively moderating the random variance due to inconsistency. Focusing on combining judgments, this paper addresses many relevant issues such as why combining and how to combine judgments, and suggests methods and models to effectively aggregate subjective judgments. We conduct an experiment to validate the effectiveness of combining judgments over individual judgments. Various combining schemes are also evaluated in terms of their predictive accuracy. Among them, mean bias based weighting scheme turns out the best. However, when available information is not enough to estimate the expertise of judges, simple and robust equal weighting might be more efficient and productive. This urges an imperative future research on the issue of 'how many and which ones to combine from a large set of experts.'

* 이 연구는 1995학년도 경희대학교 교내 연구비에 의해 연구되었음.
** 경희대학교 경영학부

I. Introduction

In the modern business environment, managers frequently face decision-making tasks of extreme complexity, ambiguity, and consequence. They have a number of cognitive abilities to make effective decisions. Since managers are human beings, however, their judgments and decision-making processes are inevitably affected by the limitations of human information processing capacity (Tversky and Kahneman 1974; Simon 1976; Hogarth 1980). Managers are subject to a number of biases and constraints. For example, because of wishful thinking, dissonance reduction, selective attention, irrational persistence of false beliefs, etc., they may (intentionally or unintentionally) bias their judgments to produce the desired outcomes (Lee et al 1987).

Moreover, managers are not or cannot be consistent in their judgments and decisions (Bowman 1963; Einhorn and Hogarth 1978; Makridakis and Hibon 1979; Hogarth and Makridakis 1981). Many works in both management and psychology have emphasized the importance of consistency in decision making. In an examination of decision fallacy due to inconsistency, Bowman(1963) proposed *management coefficient theory*. In the theory, Bowman(1963) indicated two causes of inefficiency in decision making. One is systematic deviations from normatively preferred decisions, so called bias or incorrect intuition, and the other is inconsistency in a manager's responses, what he called erratic decision making variance. The theory also asserts that it is the far out examples of behavior rather than biases which are really expensive or damaging. He also showed that if the inconsistency in a manager's decisions was removed, performance could be improved by virtue of the reduced random variance. That is, more improvement can be expected when the errors of bias are often less than those induced by variance.

The literature suggests approaches to improve managerial judgment and decision making by improving consistency. They are generally recognized as these two approaches: (1) modeling the process using the past judgment data, so-called bootstrapping, and (2) combining judgments from multiple sources. In many management situations, the consistent models outperformed inconsistent subjective judgments (Einhorn and Hogarth 1978; Makridakis and Hibon 1979; Hogarth and Makridakis 1981). In psychology also, Goldberg (1970) made a similar analysis showing that a model of man could outperform the person whose predictions were modeled, provided that the person's intuitive judgment had some degree of positive validity. Even arbitrary-consistent rules or models, to some extent, could outperform independent human judgments (Hogarth and Makridakis 1981). Thus, the relative predictive efficiency of

bootstrapping has been supported and proved.

In spite of its effectiveness, however, the modeling approach has not been applied to many managerial judgment situations due to unavailability of data, continuously changing problem contexts, inflexibility, time consuming, etc. Instead of modeling, another way to reduce the variance due to the inconsistency might be combining judgments (Granger and Newbold 1977; Moriarty and Adams 1984). There is a considerable opportunity to simply bring managers together by effectively moderating the random variance due to inconsistency. Granger and Newbold (1977) showed that in two person case, the variance of the combined judgments (forecasts) is equal to or less than the smaller of the two error variances.

Many methods have been developed to effectively aggregate subjective judgments (Ashton and Ashton 1985; Gupta and Wilton 1987; also see Clemen 1989). In this papers, we broadly address issues related with combining managerial judgments. In Section 2, we will discuss some fundamentals such as why combining and how to combine judgments, and we propose many methods and models to effectively aggregate subjective judgments in Section 3. In Section 4, we describe an experiment to obtain judgment data. Section 5 investigates the effectiveness of combining methods discussed in Section 3. In Section 6, we conclude the paper by suggesting promising future research.

2. Combining Judgments

Two problems of managerial judgment have been generally recognized, i. e., the systematic deviations from normatively preferred decisions, so called bias, and the inconsistency in a manager's responses. According to Bowman (1963), inconsistency is often more critical to the performance of managerial judgment than bias. That is, more improvement can be expected by reducing random variance rather than by correcting bias.

There are several factors deteriorating the accuracy of individual judgments and increasing the size of errors. First, the information provided to the managers might not represent the intended construct. For example, sales increase might not correctly represent the increase of profit.Second, we can also expect some measurement errors. Measurement errors are ubiquitous due to various reasons. For instance, the student's unfortunate nightmare might cause the low English test scores. Finally, the relationship between the predicted variable and the measured data is not stable to introduce systematic errors in judgment. Combining judgments reduces such random variances considerably by averaging such errors. If the sum of such ran-

dom errors is zero, combining their judgments could significantly improve the quality of group judgment.

The effectiveness of combining judgments has been validated by analytical method as well as by numerous empirical studies (see Clemen 1989). In analytic method, for example, the validity is measured by the correlation between the judgment and criterion. The correlation between the combined judgment and criterion is greater than or equal to the average of correlations between individual judgment and criterion (i.e., $\rho_{OC} \geq \rho_{OI}$ where ρ_{OC} is the correlation between the actual outcomes and the averaged judgments, and ρ_{OI} is the average of correlations between the outcomes and each individual's judgments) (Libby and Blashfield 1978; Hogarth 1978; Ashton 1986). More precisely, in combination, if the correlation of any pair of individual judgments is less than one, then the equally weighed combined judgment will be more valid than the avenge of individuals'.

In the literature, a large number of methods have been proposed for combining judgments (see Clemen 1989). The main issue addressed in those studies has been the extraction of appropriate weights for each individual judgment, i.e., the method to effectively combine the individual judgments in a way to obtain the maximum accuracy. Those studies include the weighting schemes from simple, equal weighting to variable weighting based on arithmetic and geometric measures, from linear to non-linear, or from subjective to object ratings.

However, the search for a single, all-purpose, objective combining rule might be fruitless because the best rule to use might vary from situation to situation (Winkler 1989; Makridakis 1989). Thus, it might be a more productive strategy to study the underlying assumptions associated with combining rules, and to effectively match combining rules to each forecasting situation. One of the criteria for selecting a combining method is the existence of bias which is usually measured as follows:

$$B = |(y_t - m) \div s|$$

where B is the standardized bias, y_t is the true value to be predicted, μ is the mean of a population distribution of individual judgments, and σ is the population standard deviation (Einhorn et al 1977). If individual judgments are unbiased, the equal weighting scheme, which produces a combined judgment by averaging each individual judgment with equal weight, would be the most appropriate. Under such a circumstance, the equal weighting will generally be more accurate than most individual judgments. However, the improvement in accuracy by combining judgments will decrease to the extent that individual judgments are biased. Thus, the accuracy improvement of combining judgments in terms of equal weighting method is an inverse function of bias. If bias is not zero, variable weighting is likely to be

more accurate than simple equal weighting, and such a possibility intensifies with an increasing bias. Empirically, Einhorn et al (1977) found that the standardized bias (B) should reach around 0.7 before a variable weighting scheme outperformed the equal weighting method.

In sum, if the errors of individual judgments are mainly caused by inconsistency rather than bias, equal weighting scheme seems more appropriate than variable weighting schemes, and otherwise variable weighting rules rather than averaging process.

Second, the variance of individual judgments, which vary with respect to many features, also has an effect on selecting combining rules. In other words, the group sharing the same set of information are more likely to make similar judgments with a high correlation each other. When judgments are highly correlated, we cannot expect much improvement by combining them because they are likely to share same kinds of biases. Accordingly, equal weighting scheme seems more safely fitting to the case, improving the accuracy to the extent that random error is reduced, but with unaffected amount of bias. However, when the variance among individual judgments is relatively large, variable weighting schemes could be more preferred. In general, small and large variance tend to support equal and variable weighting schemes respectively.

Finally, if available information is not enough to estimate the expertise of each individual judge or if there are some complicating factors such as dependency, high uncertainty, instability, etc, an unequal weighting scheme is risky and infirm. Instead, equal weighting is preferred because it is simple but rather robust against those complexities. Furthermore, complicated unequal weighting schemes considered up to date have not showed much improvement over equal weighting, but if any, a slight improvement at best. Therefore, from a practical point of view, it might be more productive to apply a simple equal weighting scheme, and to use resources to study other approach rather than to refine the combining method.

3. Combining Schemes

A variety of combining schemes have been studied, many of which are based on a strong assumption that individual judgment accuracy can be assessed in association with already-known true outcome, or the probability distribution of each individual's judgment is available. However, such an assumption could be rarely satisfied in the real judgment setting, and otherwise the judgment might lose its authentic value at the time when or if such outcome data are available. Thus we consider only the methods which can be used when the true out-

come data are not available.

Equal Weighting Scheme

It produces the combined judgment by weighting each participating individual judgment equally. Since it does not affect the judgment error due to bias, but the random error due to inconsistent judgments, when the error is mainly caused by bias rather than by inconsistency, it might not be appropriate. On the other hand, if the inconsistency dominates the judgment error, much improvement in accuracy could be expected with equal weighting scheme. Furthermore, in cases when no information is available to identify the better judges, the equal weighting scheme seems more appropriate because it works substantially well and safely enough to compensate the plausible improvement by using variable weighting schemes at the expense of taking the risk of far less accuracy by assigning weights reversely.

Median Judgment Scheme

This method is to use the median judgment of the combining group as the combined group judgment. In special case when the distribution of individual judgments are symmetric about the mean, this method yields the same result with equal weighting scheme. Such a symmetric distribution case also tells us the unbiasedness of judgments. Thus, if median method yields the same result with the equal weighting, we can assume that the judgments are unbiased. If biased, however, individual judgments may be markedly skewed, resulting in large differences between mean and median judgment schemes. In its essence, the median judgment scheme reflects a different aspect of averaging process than the equal weighting scheme. That is, it selects the compromising judgment, completely ignoring all extreme values. Thus the form of the distribution of judgments determines which weighting scheme leads to more accurate judgment. In case of asymmetric individual judgment distributions, however, we cannot tell which one is better to secure the higher accuracy.

Mean Bias Based Scheme

Even though all true values are not available, their mean value is sometimes available or can be easily guessed. In such cases, we can incooperate the mean value available to develop a weighting scheme. According to Moriarty (1985), the mean square error (MSE) of judgments can be decomposed into three components, i.e., the mean difference error, the regression pattern error, and the random error as follows:

$$MSE=E(A_i-P_i)^2=(m_A-m_P)^2+(s_P-rs_A)^2+(1-r^2)s_A^2$$

where μ_A and μ_P are the means of actual values and their corresponding judgmental predictions, σ_A and σ_A are their standard deviations, and ρ is the correlation between A_i and P_i. Moriarty (1985) called three components of the equation the mean difference error, the regression pattern error, and the random error respectively in their sequential order. The first and second terms are related with bias, and the third term more or less explains the error caused by inconsistency. Thus, if the first two terms are zero, the judgments are unbiased (see Moriarty 1985).

This bias decomposition is initially proposed to analyze manager's judgmental bias, to find the main source of errors, and potentially to educate the manger to correct them in the subsequent, similar judgments and decision making. The decomposition provides an intuition for developing an weighting scheme, i. e., assigning weights according to the inverse proportion of the size of errors. However, since we assume that the actual values are not known a priori, the ρ cannot be estimated beforehand, and thus the regression pattern error and the random error cannot be appraised. If the mean of the actual values could be approximated, however, we can estimate the mean difference errors, and incooperate them to assess the weight for each individual expert. That is, a relatively higher weight will be assigned to an expert showing the smaller mean difference error, in his/her judgments.

Confidence Based Scheme

Another way to access the accuracy of experts is to look at their confidence on their judgments. The confidence assessment in combining judgments has been used in some studies (Larreche and Moinpour 1983; Sniezek and Henry 1989). However, the interindividual validity of the confidence assessments is the question to be answered before using them to determine weights for combining individual judgments. If the individual confidence assessments do have interindividual validity, then they would provide a valid cue to assess the relative accuracy of the individual judgments, and to be used as a basis for weight assessments.

Consistency Based Scheme

Consistency could be used as an indication of the level of expertise. Experts tend to be more consistent in evaluating information available. Their judgments should not be easily biased depending on the contexts presenting problems (e.g., primacy and recency effects). However, this approach also falls on the problems of measuring each individual's consistency and its interindividual validity.

One of possible methods is to use 'goodness of fit' of regression model. If we can repeatedly measure the individual judgments in a similar context, with different sets of information,

then R^2 in regression analysis could represent the level of judge's consistency in evaluating each information. For example, when evaluating an applicant for Ph. D. admission, an expert is more likely to give a consistent importance weight on the score of GRE through applicants. Thus, the consistency in judgments is correlated with the level of expertise. However, note that in order to look for interindividual validity, a same set of questions should be equally provided to the judges who are supposed to participating in combining process.

4. An Experiment

An experiment was conducted to obtain judgment data to examine some combining schemes discussed in Section 3.

Questionnaire Design

We used the historical data of Ph.D. admission in one of the Ivy league schools. In order to reduce the sampling bias, we selected the applicants with US citizenship, who had the average GRE score greater than 500 and no missing attributes. Fifty seven sample applicants were selected. Among them, 26 applicants were actually accepted, and 31 rejected. We randomly divided them into two groups of 37 and 20 applicants. The first group was used as the estimation sample whose data were available to decision makers. The second was considered as the holdout sample, i.e., the problem cases on which subjects were supposed to predict the committee's decision. Thus, the task of subject is to speculate and judge the probability that the committee accepted each holdout sample applicant on the 0 to 10 scale. They are also asked to mark the confidence level of their judge on each question. Each subject judges the same twenty holdout applicants.

Among 43 attributes describing an applicant, three attributes were selected by the experts (admission coordinators and office employees) as the most important in admission decision: GRE score, undergraduate school, and recommendations. These variables were quantified for representation and analysis. Two important GRE scores were averaged to range from the minimum of 200 to the maximum of 800. We ranked the undergraduate schools from 1 to 4 scales based on the classification in Barron's Profiles of American Colleges: 'competitive,' 'very competitive,' 'highly competitive,' and 'the most competitive' schools were quantified into 1, 2, 3, and 4 respectively. The recommendations were ranked based on the summary evaluations at 'good,' 'unusual,' 'outstanding,' and 'truly exceptional.' They are scaled to 1, 2,

3 and 4 respectively. The average of three recommendations for each applicant was used in the experiment.

Manipulation

By designing four different types of questionnaires, the information given to subjects was manipulated for the experiment. Subjects were randomly assigned into four different groups. The first group was given no information. They were supposed to predict based on their personal experience only. The second group was given all the estimation sample data. The third group was provided with the three similar applicants' records to each questioned applicant, who were selected from the estimation sample by a simple similarity calculation method. For the fourth group, we provided the correlation matrix and the discriminant and regression models estimated using the estimation sample. In short, by manipulating the availability of past data as well as the methods to use and present them, we could simulate the judges' different quality of expertise and knowledge about the problems and investigate their effects on the subjects' judgments.

Samples

The sample consisted of 40 graduate students in the Ph.D. program (most majoring in business or economics) of the university, who experienced the application procedure at least once. Forty questionnaires (10 of each type) were randomly distributed to the subjects, and each group thus had 10 subjects. Each subjects were asked to predict 20 holdout samples. However, we were cautious in assigning the type IV questionnaire requiring some statistic background, only to the subjects who at least finished the statistics requirements in the doctoral program.

Manipulation Check

In order to evaluate the performance of each group, we first averaged the subject's judgments on each questioned applicant in each group. Each group had 20 averaged judgments on 20 questioned applicants. We measured the absolute differences (errors) between those judgments and the actual decisions (1 is acceptance, and 0 rejection) for each questioned applicant in each group. Each group then had 20 absolute errors between its averaged judgments and actual committee's decision. Using those absolute errors (differences), we performed the paired t-tests between group 1 and each of the other 3 groups. Table 1 summarizes the results of three paired t-tests.

Table 1: Paired T-Tests of Absolute Errors Between Group 1 and The Other 3 Groups

Groups	Mean	Std Dev	Mean Diff. with Group 1	t-Value	1-tail Prob.
No Past Data (Group 1)	0.4535	0.1995	-	-	-
All Past Data (Group 2)	0.4395	0.2192	-0.0140	-0.9934	0.1665
Similar Past Data (Group 3)	0.4160	0.2534	-0.0375	-1.2899	0.1063
Statistical Summary (Group 4)	0.4360	0.2421	-0.0175	-1.2003	0.1224

* 40 subjects were randomly assigned to each group (10 subjects for each group), and each subjects was asked 20 questions.

The availability of information, manipulation somewhat differentiates the performance of subjects' judgments among different groups. Although the differences are not statistically significant (one tail probabilities are 0.1665, 0.1063, and 0.1224), the mean absolute errors of groups are different (Group 1 = 0.4535; Group 2 = 0.4395; Group 3 = 0.4160; Group 4 = 0.4360). Manipulation is not a perfect success. However, the data are useful enough for the combining study.

5. Comparison of Combining Schemes

Using the data obtained in a specific experiment setting, we attempts to describe and evaluate combined judgments. The accuracy of combined judgments is compared to that of individuals'. Five weighting scheme models--equal, median, mean-bias, variance, consistency based weighting--are evaluated using the experimental data, following a brief discussion with presentation of the results. Prior to such discussions and illustrations, we provide the mathematical expressions of judgmental errors and weighting schemes.

Representation and Measure of Judgment Error

Judgment errors are measured by mean squared errors (MSE). Thus, the judgment errors of individual, group, and total subjects of all groups are represented in functional forms as follows:

$$MSE_i = \frac{\sum_{K=1}^{M} (A_k - P_{ik})^2}{M} \quad \cdots\cdots (1)$$

$$MSE_G=\frac{\sum_{K=1}^{M}(A_k-P_{Gk})^2}{M} \quad \cdots\cdots(2)$$

$$P_{Gk}=\frac{\sum_{i=1}^{N_G}W_iP_{ik}}{N_G} \quad \cdots\cdots(3)$$

$$MSE_T=\frac{\sum_{K=1}^{M}(A_k-\sum_{i=1}^{N}W_iP_{ik})^2}{M} \quad \cdots\cdots(4)$$

where MSE_i (1), MSE_G (2) and MSE_T (4) represent the prediction mean square errors of individual i, group G and total subjects respectively; M is the total number of problems answered by each judge; A_k is the actual value of the problem k; P_{Gk} (3) is the combined judgment of group G to predict the actual value of the problem k; W_i is the weight on the judgments of individual i in the combining process; N_G is the number of individuals belonged to group G ; N is the total number of individuals.

Assessment of Weights

Individual weights for combining judgments are estimated based on each proposed scheme. The weights are represented in functional forms as follows:

$$W_i=\frac{1}{N_G} \quad \cdots\cdots Equal\ Weight$$

$$W_i=\frac{log\left(\frac{1}{|\bar{P}_i-\bar{A}|}\right)}{\sum_{j=1}^{N_G}log\left(\frac{1}{|\bar{P}_j-\bar{A}|}\right)} \quad \cdots\cdots Mean\ Bias\ Based\ Weight$$

$$W_i=\frac{S_i^2}{\sum_{j=1}^{N_G}S_j^2} \quad \cdots\cdots Confidence(Variance)\ Based\ Weight$$

$$W_i=\frac{R_i^2}{\sum_{j=1}^{N_G}R_j^2} \quad \cdots\cdots Confidence(R^2)\ Based\ Weight$$

where W_i is the weight assigned to the judgments of individual subject i; N is the number of total subjects; $\bar{P}_i$ and $\bar{A}$ are the mean values of subject i's judgment on multiple questions and their actual outcomes respectively; N_G is the number of subjects in group G;σ_i^2 is the variance of individual i's judgments; R_i^2 is the R^2 of individual i, measured using twenty observations. In estimating the weights based on mean biases, we use log, because

otherwise the value could be an infinity. In the confidence based weight measure, we assume that the variance of individual judgments may be correlated with the level of subject's confidence on their judgments. That is, the more confident on their judgments subjects are, the more they are likely to choose extreme values, close to either 0 or 1 of binary decision values, on the probability scale. Thus, the variance of the judgments of a more confident subject tends to be larger than those of subjects with less confidence on their judgments. Assuming that goodness of fit in a linear model represents the consistency of judgments, we used R^2 of each subject to represent the consistency in his/her judgments.

Comparison of Judgment Errors

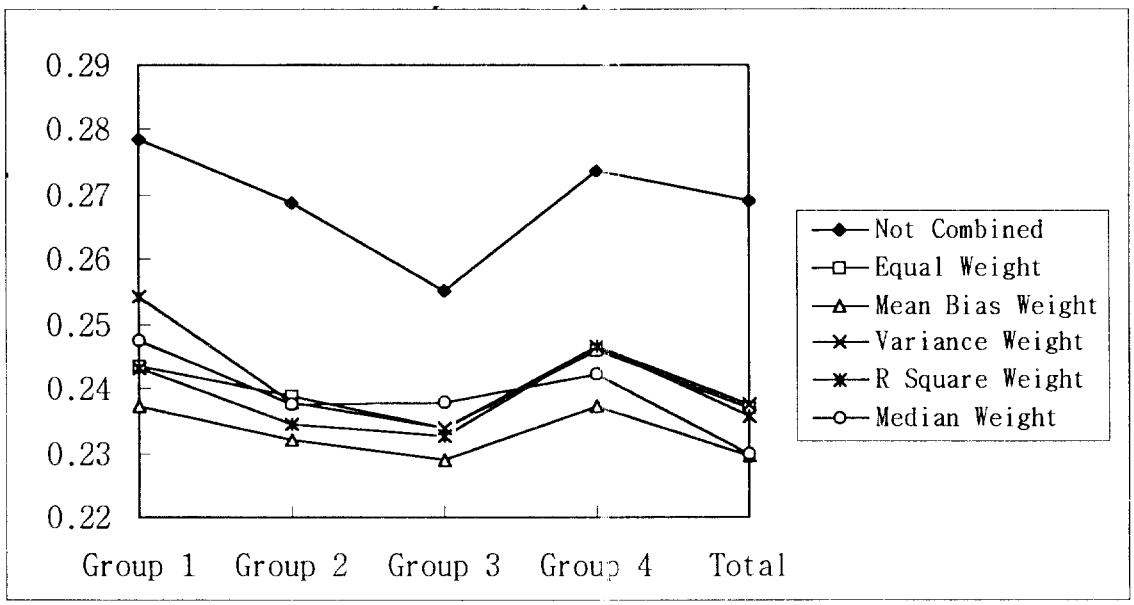
We measure MSE_G and MSE_T for each proposed combining scheme, and compare their values with the averages of MSE_S' for each group and total subjects, which are considered as the basis. We summarize the results in Table 2, based on which a graph is generated as shown in Figure 1.

Table 2: Mean Squared Errors (MSE's) of Not Combined and 5 Combining Schemes by Each Group and Total

Weighing Schemes	Group 1	Group 2	Group 3	Group 4	Total
Not Combined (Individual)	0.2787	0.2687	0.2551	0.2737	0.2690
Equal Weighting	0.2434	0.2388	0.2340	0.2459	0.2370
Mean Bias Based Weighting	0.2372	0.2320	0.2288	0.2372	0. 2296
Variance Based Weighting	0.2541	0.2478	0.2339	0.2466	0.2375
R^2 Based Weighting	0.2430	0.2345	0.2327	0.2465	0.2357
Median Based Weighting	0.2475	0.2376	0.2378	0.2423	0.2298

As we expected, the combined judgments outperform the averages of MSE_S' of individual subject judgments in all cases (four groups and total group). Their differences are statistically significant in all comparisons ($p < 0.01$). Thus, regardless the methods used to combine judgments, the combined judgments are significantly better than individual judgments. This difference mainly comes from the variance, which has been by far reduced in terms of combining process. In comparison with the significant differences between the combined and the not-combined, the differences among the combined judgments of different methods are not so much meaningful. It is because their differences should rely only on the rest components of

Figure 1: Mean Squared Errors (MSE's) of Not Combined and 5 Combining Schemes by Each Group and Total



error, excluding the variance.

Among the tested weighting schemes, the weighting based on the size of mean bias shows the best performances in all cases. It seems because the mean difference (so-called mean bias) between the judgments and the actual values is another major source of judgment errors. In fact, when we investigated the mean of each individual's judgments, all subjects' judgments are positively biased. That is, in comparison with the actual mean of 0.45 (9 out of 20 questioned applicants are accepted and the rest 11 are rejected), no subject's judgment means are less than 0.45, ranging from 0.5 to 0.7. The subjects in Group 3 showed the least mean bias and those in Group 4 showed the highest bias. The means and the standard deviations of subjects' judgment averages in each group are shown in Table 3.

When we measure the standardized biases, their averages range around 2.0 (see the large

Table 3: Group Means and Standard Deviations of Subjects' Judgment Averages

	Group 1	Group 2	Group 3	Group 4	Total Subjects
Group Mean	0.6345	0.6125	0.5770	0.6700	0.6235
Standard Deviation	0.0930	0.0839	0.0735	0.0862	0.0910

biases in Table 3). Thus variable weighting scheme such as mean bias based weighting consistently outperforms equal weighting. However, equal weighting scheme also shows the reasonable performance in comparison with other weighting schemes except the mean bias based one. When available information is not enough to estimate the expertise of judges, simple and robust equal weighting might be more efficient and productive.

6. Concluding Remarks

Managers frequently make judgments. However, the effectiveness of their judgments is bounded due to bias and inconsistency. In the literature, inconsistency has been recognized more critical than bias(Bowman 1963), and many approaches have been proposed to improve the quality of judgments by reducing inconsistency. Among the proposed, this paper focuses on combining judgments, which has received the recent attention due to many successful empirical results in combining forecasts from multiple models. Since the concept of combining models is almost identical to that of combining judgments, we applied the theory and methodology developed for combining models to combining judgments with minimum modifications necessary, and suggested several combining schemes.

An experiment was conducted to gather judgment data to test the schemes proposed. In the experiment, different expertise of judges were enforced by manipulating the information provided to each subject. The combined judgments of all combining schemes significantly outperformed the individual judgments. In the experiment judgment data set, a variable weighting scheme, mean bias based weighting consistently provided better predictions than any other variable and equal weighting methods tested.

However, simple equal weighting scheme also showed a relatively good and robust performance. When available information is not enough to estimate the expertise of judges, simple and robust equal weighting might be more efficient and productive. If complicated variable weighting schemes could not significantly improve the performance over the simple, equal weighting method, we had better use our resources to investigate other convincing techniques to improve the quality of combined judgments rather than to search for the better weighting schemes. Instead of cultivating weighting rules, for instance, we had better develop a method to effectively select the most appropriate judges to combine from a large pool of experts with unpredictable expertise. This suggests the very important but rather untouched future research topics of 'how many and which ones to combine when combining judgments.'

References

[1] Ashton, A. H. and Ashton R. H. (1985), "Aggregating Subjective Forecasts: Some Empirical Result," *Management Science*, 1499-1508.

[2] Ashton, A. H. (1986), "Combining the Judgements of Experts: How Many and Which Ones?," *Organizational Behavior and Human Performance*, 38, 405-414.

[3] Bowman, E. H. (1963), "Consistency and Optimality in Managerial Decision Making," *Management Science*, Vol. 9, 310-321.

[4] Clemen, R. T. (1989), "Combining Forecasts: A Review and Annotated Bibliography," *Internal Journal of Forecasting*, 5, 559-583.

[5] Einhorn, H. J., and Hogarth, R. M. (1978), "Confidence in Judgment: Persistence of the Illusion of Validity," *Psychological Review*, Vol. 85, 395-416.

[6] Einhorn, H. J. (1974), "Expert Judgement: Some Necessary Conditions and An Example," *Journal of Applied Psychology*, Vol. 57, 562-571.

[7] Einhorn, H. J., and Hogarth, R. M. (1977), "Quality of Group Judgement," *Psychological Bulletin*, Vol. 84, 158-172.

[8] Goldberg, L. R. (1970), "Man Versus Model of Man: A Rationale, plus Some Evidence, for a Method of Improving on Clinical Inferences," *Psychological Bulletin*, Vol. 73, 422-432.

[9] Granger, C. W. J. and Newbold, P. (1977), "Some Comments on the Evaluation of Economic Forecasts," *Applied Economics*, 5, 35-47.

[10] Hogarth, R. M. (1978), "A Note on Aggregating Opinions," *Organizational Behavior and Human Performance*, 21, 40-46.

[11] Hogarth, R. M. (1980), "Judgement and Choice: The Psychology of Decision," *Wiley*, Chichester, England.

[12] Hogarth, R. M. and Makridakis, S. (1981). "The Value of Decision Making in A Complex Environment: An Experimental Approach," *Management Science*, Vol. 16, 93-107.

[13] Hogarth, R. M. and Makridakis, S. (1981). "Forecasting and Planning: An Evaluation," *Management Science*, Vol. 27, 115-138.

[14] Larreche, J. and Moinpour, R. (1983), "Managerial Judgment in Marketing: The Concept of Expertise," *Journal of Marketing Research*, 110-121.

[15] Lee, H., Acito, F., and Day, R. L. (1987). "Evaluation and Use of Marketing Research by Decision Makers: A Behavioral Simulation," *Journal of Marketing Research*, Vol. 14, 187-196.

[16] Libby, R and Blashfield, R. K. (1978), "Performance of a Composite as A Function of The Number of Judges," *Organizational Behavior and Human Performance*, 21, 121-129.

[17] Makridakis, S., and Hibon, M. (1979), "Accuracy of Forecasting: An Empirical Investigation," *Journal of Royal Statistic Society*. 97-145.

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

[19] Moriarty, M. M. (1985), "Design Features of Forecasting Systems Involving Management Judgments," *Journal of Marketing Research*, 353-364.

[20] Moriarty, M. M. and Adams, A. J. (1984), "Management Judgment Forecasts, Composite Forecasting Models, and Conditional Efficiency," *Journal of Marketing Research*, 239-250.

[21] Simon, H. A. (1976), "Administrative Behavior," *The Free Press*, New York.

[22] Sniezek, J. A. and Henry, R. A. (1989), "Accuracy and Confidence in Group Judgement," *Organizational Behavior and Human Performance*, 43, 1-28.

[23] Tversky, A. and Kahneman, D. (1974), "Judgement under uncertainty: Heuristics and Biases," *Science*, 185, 1124-1131.

[24] Wilton, Peter C. and Gupta, Sunil (1987), "Combination of Forecasts: An Extension," *Management Science*, 356-372.

[25] Winkler, R. L. (1989), "Combining Forecasts: A Philosophical Basis and Some Current Issues," *International Journal of Forecasting*, Vol. 5, 605-609.

* Questionnaires are available from the author upon request.