ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 7, 86–106 (1972)

Expert Measurement and Mechanical Combination[1]

HILLEL J. EINHORN[2]
Graduate School of Business, University of Chicago

The expert can and should be used as a provider of input for a mechanical combining process since most studies show mechanical combination to be superior to clinical combination. However, even in expert measurement, the global judgment is itself a clinical combination of other judgmental components and as such it may not be as efficient as a mechanical combination of the components. The superiority of mechanically combining components as opposed to using the global judgment for predicting some external criterion is discussed. The use of components is extended to deal with multiple judges since specific judges may be differentially valid with respect to subsets of components for predicting the criterion. These ideas are illustrated by using the results of a study dealing with the prediction of survival on the basis of information contained in biopsies taken from patients having a certain type of cancer. Judgments were made by three highly trained pathologists. Implications and extensions for using expert measurement and mechanical combination are discussed.

In an important article, Sawyer (1966) reviewed the issue of clinical vs statistical prediction that was first summarized by Meehl (1954). Sawyer made the very important point that one must distinguish between the mode of *data collection* and the *combination method* used to deal with the data once collected. Both of these factors could involve clinical or statistical methods and Sawyer labeled the combining process as being either clinical or mechanical. With regard to the method of data collection, the measuring instruments could also be either clinical or nonclinical. The Sawyer review essentially confirmed the earlier review of Meehl with respect to the best method for *combining data*, viz., the mechanical mode of combination is clearly superior to the clinical mode

[1] I would like to thank my colleagues, Dr. L. Warwick Coppleson, Department of Surgery, and Jerry Rose, Department of Health Administration Studies, for their help with the larger study dealing with the analysis of diagnosis. I would also like to thank the following people for reading an earlier draft of the manuscript: Alan R. Bass, Nicholas J. Gonedes, L. Richard Hoffman, and Paul Slovic. Finally, special thanks are due Zvi Lieber for his help with the programming needed for the analyses.

[2] This research was partially supported by grant 1-R03 MH-18 599-01 from the National Institutes of Mental Health.

of combination. This result has already had important implications for the aggregating of information in many different areas. The use of "machines" for aggregating information includes the "Bayesian" approach to information processing, where the phenomenon of "conservatism" is thought to be due, in part, to man's inability to aggregate information in an optimal fashion as defined by Bayes' theorem, as well as regression approaches where computers combine information in some optimal fashion (for a comprehensive review of both the Bayesian and regression approaches to information processing in judgment, see Slovic & Lichtenstein, 1970).

The fact that man cannot combine information as efficiently as a computer, for example, does not imply that man can be replaced by machines. It does imply that the necessity for a man-computer system is at hand and efforts in this direction are likely to lead to optimal uses of information (Yntema & Torgerson, 1961). The question that is of concern here does not involve the issue of clinical vs statistical combination; the fact that a machine is more efficient for the process of combining information is accepted. The major interest here is in the use of expert measurement, i.e., information collected in a clinical fashion, as input to be used in the mechanical combination process. As Sawyer pointed out (1966),

> In the *measurement* part of the clinical-statistical question, the comparison is by no means as clear. The best method from Table 4—the mechanical composite—includes data collected both clinically and mechanically. (So does the mechanical synthesis which predicts as well but requires more data.) This suggests that the clinician may be able to contribute most not by direct prediction, but rather by providing, in objective form, judgments to be combined mechanically [p. 193].

The use of the clinician (or the more inclusive term "expert") has been relatively neglected with regard to his role as a possible provider of inputs for a mechanical combination. This neglect may be partially explained by the hostility experts may feel regarding the mechanical means to "replace" them. In addition, they may feel that the important task in decision making is in the combining process and, therefore, they do not see themselves as playing a subsidiary role (as they see it) of provider of input to a computer. However, the use of both man and machine is absolutely necessary. This kind of system is exemplified by the PIP system (probabilistic information processing) developed by Edwards and his associates (e.g., Edwards, 1962; Edwards & Phillips, 1964; Edwards, Phillips, Hays, & Goodman, 1968). The basic idea of this system is that men supply the inputs which are in the form of conditional probabilities [p(Data/Hypothesis)] or in the form of likeli-

hood ratios. Once the men have specified the input, it is the machines that do the aggregating on the basis of Bayes' theorem. This system is exactly the kind where one has expert measurement and mechanical combination. If it is true, as Peterson and Beach (1968) have claimed, that man is an intuitive statistician, then the use of men to provide the inputs to such a system may be quite an efficient strategy, especially in cases where one cannot get "objective" information.

The following discussion assumes that one has some criterion measure that is quantitative in some form and is available (it does not matter if this criterion is judgmental or objective). The discussion is illustrated with examples from the medical specialty of pathology although the methods advocated should be general to any area of interest (the reader especially interested in medical decision making is referred to: Gustafson, 1969; Kleinmuntz, 1968; Lusted, 1968). The major point to be made here concerns the distinction between the global, overall judgment and the components that went into that judgment. The components may also be judgments but differ with respect to their degree of completeness. As an example, let us say that a pathologist, whose job is to determine the make-up of tissue, can give an overall global judgment as to the kind of disease or severity of disease as indicated by a particular slide of tissue. However, within this global judgment have gone judgments concerning the various cues or variables that have somehow been combined to form the global judgment. The basic point to be made here is that the overall, global judgment is itself not only a combination of the component judgments but is specifically a *clinical* combination of the components. However, we already know from most of the research that the clinical combination mode is not very good as compared to the mechanical mode of combination. What the expert has done in achieving a global judgment is to combine the components in some way using some cognitive combining rule. If one wanted to use the global judgment for predicting some criterion, one would be using a cognitive *summary measure* which might not be efficient. There are several reasons for this: (1) The clinical combination may be leaving out important components; (2) it may be weighting the components in nonoptimal ways (i.e., as a mechanical method would do) so that the global judgment would not be as effective as the properly weighted components; (3) it may be that the combining rule used by the judge helps him to simplify the situation so that he may be ignoring and weighting variables in such a way as to reduce cognitive strain. Related to this is the possibility that the combining rule imposed by the judge in combining the components may not be the best rule that combines the components with respect to the criterion to be predicted. If this occurred, i.e., if there was a difference

in the combining rule used by the judge in determining his global judgment and the best rule for combining the components in predicting the criterion, the use of the global judgment might be inferior to the use of the components combined mechanically.

The last point above can be conceptualized by using the "lens model" (Brunswik, 1952; Hammond, 1955; Dudycha & Naylor, 1966). Let the distal stimulus or true state be designated Y_e and the judgment or decision designated Y_s. Further, let the components or cues be symbolized X_i. The question that is of importance concerns the relationship between the functions that relate the X_i's to both Y_e and Y_s. If the cognitive combining rule is designated $f(X_i)$ then one would expect to be able to predict Y_s on the basis of the function rule and the values of X_i. On the ecological side of the lens, however, the best function that relates X_i to Y_e *may not be* $f(X_i)$, but rather $g(X_i)$. If this were the case, one would have nonmatching functions relating X_i to Y_e and Y_s and the use of the global judgment (Y_s) for predicting Y_e would clearly do a poor job. The nonmatching of functions also relates to the measure developed by Hursch, Hammond, & Hursch (1964) for dealing with nonlinearity in the judges' response system and the nonlinearity in the ecology. This measure, designated "C," gives the correlation between the nonlinearity in the response system and the ecology. However, "C" can be zero for several reasons (Einhorn, 1970). If the functions on the two sides of the lens do not match, even though they are both nonlinear, the "C" measure can be zero. As will be described later in this study, the functions that best predict Y_e and Y_s *were* different so that the "C" measure would not be entirely appropriate here.

The concern with the use of global judgments has been expressed by Brown (1970).

> Consequently, it is reasonable to question the continued reliance on global clinical judgments as a predictor of suicide lethality. Numerous items of information have to be reduced to a small number of cue dimensions, such as "symptoms," "resources," etc. which a judge might feasibly be able to consider. However, each of these categories may contain items which are not valid for the Suicide Prevention Center population, as well as some which are only poor predictors. By mixing them together with better items, the predictive validity of the better signs is attenuated. Moreover, the findings of this study reveal that judges tend to rely only upon two or three of these cue dimensions, possibly not even using some of the better predictors, and thus, further attenuating the validity of all the separate items. Then a scale, consisting only of valid items, could be constructed. These items could be checked off by the worker as he talks with the caller and the lethality rating determined by some empirical weighting system. If the total set of valid items is too large to be obtained during a crisis telephone call, a smaller number of better items could be used as a rough

> lethality scale, and the more complete scale could be utilized as a research instrument. On the other hand, if adequate prediction could be found with a small number of items, then these might be utilized as the predictors of suicide lethality [pp. 109–110].

Furthermore, the idea of using components of a judgment has been dealt with by Sarbin, Taft, and Bailey (1960) and in a study by Blenkner (1954). Sawyer (1966) calls the use of clinical information combined mechanically as dealing primarily with trait ratings instead of overall summary measures. However, there is no reason to limit the potential use of components to trait ratings since the components of any global judgment could be tried instead of, or in addition to, the global judgment. This means that the researcher should design his study so that the components of the global judgment can be obtained.

The preceding ideas will be illustrated by a study in progress by Einhorn, Coppleson, and Rose, dealing with the analysis of the diagnosis of Hodgkin's disease (a form of cancer of the lymph system) by three highly trained pathologists. An extension of the basic ideas of multiple judges will be discussed after the description of the study and the results dealing with the individual analyses. The criterion we are attempting to explain (or predict) is survival time (in months) of patients suffering from this disease. The data were made available to us by the physicians originally involved with the study and represent a subset of all the cases of this disease at a particular large metropolitan hospital from 1930 to 1964. A total sample of 193 cases were used in this study. All of the patients used had died although none of the physicians had any knowledge of the patients other than a biopsy slide taken when the person first came to the hospital. The design of the study involved the three pathologists independently viewing each of the 193 slides over a year's time period (viewing one slide takes a considerable amount of time). All of the patients had been diagnosed as having the disease. The pathologists, for each slide, had to give their judgment as to the relative amount of nine histological characteristics that they saw in each of the slides. In addition, they also had to give an overall or global judgment as to the severity of the disease in terms of a classification scheme developed by experts in this area. The doctors themselves picked out the histological characteristics that they thought were important, a priori, and they formed the scales on which each of the signs was measured. For eight of the signs, a five-point scale was used while for the ninth sign a two-point scale was used. The overall, global judgment was made on a nine-point scale of severity. Besides the sample of 193 slides, 26 repeat slides were given so that test, retest reliability could be obtained for each of the signs as well as for the global judgment. In

addition to the data collected from the three pathologists, a composite judge was also used (Judge 4). The composite was simply the average judgment of the three doctors except in cases where they widely disagreed. In those cases the doctors met in a group to discuss their differences and came to some agreement. The agreement value was used in such cases. It should be stated that survival time for this disease is highly variable so that it is very difficult to deal with prognosis (or predict survival).

The first analysis deals with the use of the global judgment as a predictor of the survival time criterion. Because of the skewness in the criterion both raw survival time as well as a log transformation of the measure were tried. These results are presented in Table 1.

One should expect a negative correlation between the global judgment and the survival time measure since the higher the severity of the disease the lower should be the survival. The correlations obtained were quite low and for some judges the relationship was positive! However, these correlations are essentially random since none reached statistical significance using $\alpha = .01$. In defense of the doctors, it should be remembered that patients did receive treatments of different kinds so that one might question whether initial diagnosis should be related to survival (treatment information was unfortunately not available to us). Moreover, patients may have died for other reasons than the disease, and this would also tend to lower the correlation of severity with survival time (in some cases the therapy may well have caused death since it is quite powerful). Given all of the factors that may have affected the criterion it should be stressed that the criterion contamination should tend to lower the correlations, not only of severity of disease with survival, but of other variables as well. However, if the variance in the criterion were wholly unrelated to the information contained in the biopsy then one would not expect the use of the components to do any better in predicting the criterion than the use of the global judgment.

TABLE 1
CORRELATION OF GLOBAL JUDGMENT WITH SURVIVAL TIME

	Global-survival time		Global-log survival time	
Judge	r	r^2	r	r^2
1	−.002	.000	−.038	.002
2	.116	.012	.098	.010
3	−.139	.019	−.127	.016
4	.143	.020	.072	.005

Note: All r's based on $n = 193$; r needed for significance at $p < .01$ is .179.

In order to explore the utility for using the components of the global judgment, the following procedure was used. Each doctor was analyzed separately so that the results are based on individual analyses. The judgmental components (judgments of histological signs) were used as the predictors while survival time (in months) was used as the criterion. A second analysis was performed (again, for each doctor) in which the global judgment was included as a predictor variable. The reason for doing this was due to the possibility that the global judgment might contain something more than what is included in the components. For example, the global judgment might include information not contained in the components and/or it might also contain validity due to the "clinical combination" method used by the judge. In order to deal with the possible function rules that might model the judges, three different models were tried. These models were: (1) the linear model, (2) conjunctive model, and (3) disjunctive model. The latter two models have been conceptualized by Coombs (1964), Dawes (1964), and given a mathematical approximation by Einhorn (1970). The conjunctive model differs from the linear model since no compensation is allowed for, i.e., it is a multiple cutoff procedure where one must have certain minimums on a number of dimensions. The disjunctive model, on the other hand, says that in order to have a high utility, the multi-attribute stimulus should have one attribute which has a very high level (an example of the disjunctive model would be in the selecting of football players—one would want outstanding ability on kicking, or passing, or running). These models have been used in two experimental situations where they provided a better fit for the data than the linear model for many subjects (Einhorn, 1971). The results for the analyses are presented in Table 2.

Examination of Table 2 shows that the disjunctive model provided the best fit for the data for all the judges. However, of more importance is the fact that when one looks at the correlations for the components *alone*, there is a large increase in the correlations and the amount of explained variances. For example, using the first judge, the amount of variance explained from using the global judgment is 0% but using the components of the global judgment the R^2 goes to .185 or 18.5% of the variance accounted for. The same general picture emerges for the other judges. The second half of Table 2 shows the results when the global judgment is added to the components. It can be seen that the adding of the global judgment does not add very much to the components although this varies somewhat for the different doctors. Both Judges 2 and 4 seem to benefit most by the addition of the global judgment while Judges 1 and 3 do not seem to gain very much by the addition of this other variable.

TABLE 2

INITIAL FIT OF COMPONENTS WITH AND WITHOUT GLOBAL JUDGMENT IN RELATION TO SURVIVAL TIME

	Components alone						Components + global judgment					
	Linear		Conj.		Disj.		Linear		Conj.		Disj.	
Judges	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
1	.378	.143	.407	.166	.430	.185	.386	.149	.409	.167	.439	.193
2	.258	.066	.246	.060	.297	.088	.335	.112	.290	.084	.355	.126
3	.302	.091	.313	.098	.333	.111	.353	.125	.357	.128	.381	.145
4	.364	.132	.388	.151	.408	.166	.425	.181	.414	.172	.452	.204

Note: All R's based on $n = 193$.

The next question that is raised by these results concerns how well one can predict using the components (with and without the global judgment) vs using the global judgment alone. This question can be answered by using a cross-validation scheme to test for the amount of "shrinkage" one gets using the fit on another sample. The following procedure was used: The sample was randomly split into two sets consisting of 100 and 93 observations. The first set was used to obtain the regression weights for the particular function rule used and the regression equation was then used to obtain predicted survival time scores for the second set of 93 observations. The correlations between the observed and predicted scores give one an indication of the predictive efficiency of each model for predicting the criterion. The results are shown in Table 3.

Examination of Table 3 shows that there was "shrinkage" as one might expect. However, the disjunctive model still provided the highest predictive efficiency of any of the models. It can also be seen that the amount of shrinkage that one gets is not uniform for the four different judges. For example, Judge 2 does not seem to be as affected by the cross-validation procedure as does Judge 3. Exactly why this occurs is not clear since the same number of variables were used for each of the physicians and the sample sizes were also held constant for each doctor. Another interesting point to be seen in Table 3 concerns the relative improvement in prediction using the global judgment when added to the components. In some cases, the predictive validity is lower when the global judgment is added to the components, while in other cases the predictive validity is higher. The reason for the lower correlation comes from the fact that in the initial fit one is adding an extra predictor variable and therefore one might expect a greater amount of shrinkage. However, for Judges 2 and 4, the inclusion of the global judgment with the components does add to explaining the variance of the criterion so that it should be included for predictive purposes.

It should be noted that the cross-validity for using the global judgment alone is close to zero. This is due to the fact that the initial fit using the global judgment was not significant and, therefore, any shrinkage that one would obtain would mean that the predictive validity would be essentially zero. If one compares the amount of variance accounted for by using the components alone as compared with the use of the global judgment, it is clear that the components do a much better job of predicting the criterion. The correlations obtained for Judges 1, 2, and 4 are significantly different from zero (using the disjunctive model and $\alpha = .01$) so that the use of the components alone does a superior job to using just the global judgment alone. The question as to whether the global judgment adds anything over and above what is already

TABLE 3

CROSS-VALIDATION FOR THREE MODELS FOR PREDICTING SURVIVAL TIME USING COMPONENTS WITH AND WITHOUT THE GLOBAL JUDGMENT

	Components alone						Components + global judgment					
	Linear		Conj.		Disj.		Linear		Conj.		Disj.	
Judges	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
1	.258*	.066	.320*	.103	.380*	.145	.150	.023	.292*	.086	.359*	.129
2	.220	.048	.171	.029	.293*	.086	.331*	.111	.250	.063	.340*	.116
3	.201	.040	.195	.038	.229	.052	.149	.022	.135	.018	.191	.036
4	.249	.062	.273*	.074	.350*	.122	.295*	.087	.278*	.077	.377*	.142

Note: The original sample was $n = 100$ and the cross-validated sample was $n = 93$.

* $p < .01$.

contained in the components is given an equivocal answer from these results. It seems that in certain cases the global judgment does add to the components and should be included in the prediction equation, while in other cases its inclusion only tends to lower the predictability. This is obviously an empirical question that can only be answered by doing the research in the particular situation.

Predicting the Global Judgment from the Components

The next issue concerns the relationship between the components of the global judgment and the global judgment itself. It was stated earlier that the combining rule (or function) used in the cognitive combining of the components might not be the same rule as the best combining function used to predict the criterion. This was investigated here in the following way: The global judgment was used as the criterion and the components as the predictors for each judge. Again, three different models (linear, conjunctive, and disjunctive) were used as possible combining rules for representing what the judge might be doing. While there has been a great deal of research concerned with the finding of "configural" judges, little has been said about the normative question as to whether judges *should be* configural. Using the lens model conceptualization, this question becomes one of finding out if the best function for predicting Y_e is a configural function and if man's judgment, Y_s, is also a configural function of the cues. However, it is necessary that the two functions be the same since the notion of configurality may include many different types of functional rules. Using the notation introduced before, man should be configural if $f(X_i) = g(X_i)$, i.e., if the same configural function is used to combine components in the ecology and the judge's response system. The results for predicting the global judgment from the components (based on double-cross validation) are presented in Table 4.

The first thing that can be seen is that conjunctive model is the best model for predicting the global judgment. Although the correlations are low (in comparison to other regression studies) the nonlinear, noncompensatory models explain twice the amount of variance as does the linear model. Since the conjunctive model is a multiplicative model this finding suggests that these judges are more appropriately modeled by configural, interactive models rather than by the linear model. Furthermore, these results are consistent with the notion of the nonlinear, noncompensatory models serving as a possible cognitive simplification mechanism (Einhorn, 1971). Although no ratings of the difficulty of the task were obtained from the pathologists, this task is considered extremely difficult and complex by the judges. Given the complexity of the task and the relative superiority of the nonlinear models to the linear model, the

TABLE 4
PREDICTION OF THE GLOBAL JUDGMENT FROM THE COMPONENTS

	Linear		Conj.		Disj.	
Judges	R	R^2	R	R^2	R	R^2
1	.340*	.115	.447*	.200	.425*	.181
2	.019	.000	.073	.005	.070	.005
3	.330*	.109	.445*	.198	.275*	.076
4	.230	.053	.332*	.110	.318*	.101

Note: All correlations are based on the average correlation obtained on double cross-validation for the two samples. These correlations were first transformed by Fisher's z and then re-transformed back to correlations. The sample sizes for the two sets of data were, $n_1 = 97$, $n_2 = 96$.

* $p < .01$.

idea that the nonlinear, noncompensatory models serve as a cognitive simplification device is given support by these data.

The fact that the conjunctive model provides the best prediction of the global judgment while the disjunctive model is best for predicting survival time from the same components illustrates that different functions may be used on opposite sides of the lens. With respect to the normative question as to whether judges should be configural, these results point to the fact that one must specify exactly what one means by being configural since two different functions may be configural but nonmatching. Illustrating this is the fact that the conjunctive model, which is best for predicting the global judgment, is quite poor for predicting the survival time criterion. These results also bear on the recent study by Goldberg (1970), dealing with the use of "models of man" vs man himself. The results obtained here show that without an outside criterion, one would have no way of knowing if the function rule one was using to model the judge was the same function that best combined the information for dealing with the distal stimulus. By assuming that a linear model can be used in the right hand side of the lens, Goldberg has implicitly assumed that the linear model is also the most appropriate model for combining the information to predict Y_e in the left side of the lens. Although it would be desirable to say that one doesn't need a criterion to "boost" one by one's own bootstraps, the use of a linear model (i.e., the clinical combining rule was closer to $g(X_i)$ than the man himself. This would occur if the "clinical" combining rule was closer to the function that best predicted Y_e on the basis of X_i than the linear model (i.e., the clinical combining rule was closer to $g(X_i)$ than the linear model).

Multiple Judges

Up to this point this study has dealt with the components of each doctor's judgment in order to predict the survival time criterion. However, it is certainly possible that each judge is only an expert in a *subset* of the judgmental components. As Pankoff (1967) has stated,

> Another problem in deciding how to best use an expert's judgments is the determination of his proper bailiwick or scope of specialized knowledge. For example, some foreign policy experts may be much better than others at assessing the chances of war with particular countries and certain psychoanalysts may specialize in cases of schizophrenia. It would be a mistake not to recognize such limitations as they apply to the decision in question [p. 36].

The use of different judges for different components can be conceptualized in a different manner. If one thinks of the various judges as "methods" and the components of the judgment as "traits," then one has a multi-trait, multi-method situation where the methods and traits interact (in the statistical sense of the term) with regard to predicting the criterion variable in question. In other words, the use of multiple judges may allow one to take advantage of any interaction between the components and the judges so that these interactions can be used to maximize the predictability of the criterion. The potential usefulness of this idea depends to a large extent on having differences between the judges with respect to their judgments of the components. In this study this was unfortunately not the case (although fortunate for the patients), since examination of the multi-trait, multi-method matrix showed a high degree of convergent validity. The average inter-rater reliability, averaging over the nine components, was $r = .583$ ($p < .01$). In the cases where there are larger differences between judges, the method advocated here should produce better results than those obtained here.

The use of the judges by components interaction was dealt with in the following manner: All the signs, from all the judges, were put into the computer as input for a mechanical combination, using the three models as before for predicting the survival time criterion. In one case, the global judgments were left out and in another case the global judgments were included. A step-wise regression procedure was used so that the machine could pick out those variables that best predict the criterion. In this situation, the superiority of the machine to the man is evidenced since this task is beyond the computational skills of the man. The step-wise procedure was used since Darlington (1968) has suggested that this procedure be used when one wants to select a subset of predictors from a larger set. The first ten steps of the program were used so that the number of variables would be the same as in the individual analyses.

TABLE 5
RESULTS FOR MULTIPLE JUDGES ON INITIAL FIT AND CROSS-VALIDATION

	Components						Components + global judgment					
	Linear		Conj.		Disj.		Linear		Conj.		Disj.	
	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2	R	R^2
Initial fit	.453	.205	.510	.260	.549	.301	.521	.271	.510	.260	.560	.314
Cross-validation	.202	.041	.210	.044	.363*	.132	.287	.083	.180	.032	.396*	.157

Note: Initial fit on $n = 193$. Cross-validation done on $n = 100$ for initial fit and $n = 93$ for cross-validated sample.
* $p < .01$ for cross-validated sample.

The results for using the multiple judges, both for initial fit and on cross-validation, are shown in Table 5.

While the results on initial fit are higher than any individual judge it can be seen that there is considerable shrinkage for the multiple judge technique as seen in the cross-validation results. However, if one looks at the disjunctive model global judgment column, it can be seen that the use of multiple judges does do better than any individual judge. Although the increase in explained variance is not large, any increase may contain utility (Cronbrach & Gleser, 1965). It is also the case that because of the high degree of convergent validity between our judges that these results are most assuredly conservative and the method should work better where convergent validity is not as high.

Another potential advantage for the use of multiple judges would be in situations where judges from different areas of specialization might be used. For example, in predicting job success, one might want to obtain the components of judgment from a test specialist, a clinical psychologist, and an interviewer. The best components (in terms of predicting the criterion) from each specialist could then be used to set up an optimal prediction equation. If there were differential validity with respect to the components, the computer would be able to pick out those variables that best predict the criterion. This extension of the idea of multiple judges might also imply that "decision teams" be set up when dealing with particular criterion variables.

Further Results

Although the results for this study show the superiority for using components rather than the global judgment as well as the potential usefulness for using multiple judges, the correlations obtained here are low (of course, the judgment as to what constitutes a "low" correlation may differ for different judges). Some discussion of this is necessary with regard to this study. Firstly, in both the case of predicting the survival time criterion as well as predicting the global judgment, the predictor variables were themselves judgments. This means that there was a certain degree of unreliability in the predictors which would attenuate the obtained correlations. In the case where the global judgment was used as a criterion, there was unreliability in both the predictors and the criterion. Table 6 gives the intra-rater reliability for the nine components and the judgmental criterion for the four judges.

When one "corrects" for attenuation, the correlations are increased. This is shown for the disjunctive model in predicting the survival time criterion and the conjunctive model for predicting the global judgment in Table 7.

TABLE 6
INTRA-RATER RELIABILITY FOR SIGNS AND GLOBAL JUDGMENT

Signs	Judge 1	Judge 2	Judge 3	Judge 4	Mean
1	.43	.71	.40	.67	.57
2	.82	.86	.72	.69	.78
3	.86	.90	.95	.95	.92
4	.84	.84	.55	.53	.72
5	.47	.53	.57	.50	.52
6	.49	.64	.55	.44	.54
7	.85	.76	.73	.71	.77
8	.22	.17	.16	.46	.25
9	.59	.83	.59	.67	.69
Global judgment	.69	.46	.71	.90	.73
Mean of all signs for judge	.68	.74	.64	.67	

Note: These reliabilities were obtained from a sample of $n = 26$ repeat slides. These reliabilities are therefore test re-test. All means were computed using Fisher's z transformation.

The second reason for the low correlations with regard to the survival time criterion has already been mentioned. This is the fact that the criterion is contaminated and the amount of variance due to factors not connected with the information presented would tend to lower any correlation having to do with this measure. A third factor is that this problem was a real-life decision problem and not a highly controlled

TABLE 7
CORRECTION FOR ATTENUATION FOR PREDICTING BOTH SURVIVAL TIME AND THE GLOBAL JUDGMENT

Judges	$r_{obtained}$	r^2	$r^a_{corrected}$	r^2
	Survival time criterion (disjunctive model, components only)			
1	.380	.145	.463	.214
2	.293	.086	.340	.115
3	.229	.052	.286	.082
4	.350	.122	.426	.181
	Global judgment criterion (conjunctive model)			
Judges	$r_{obtained}$	r^2	$r^b_{corrected}$	r^2
1	.447	.200	.657	.432
2	.073	.005	.130	.017
3	.445	.198	.660	.436
4	.332	.110	.430	.185

[a] The formula for the correction is given by: $r_{corrected} = r_{obt.}/(r_{xx})^{1/2}$.
[b] The formula for the correction is given by: $r_{corrected} = r_{obt.}/[r_{xx}r_{yy}]^{1/2}$.

laboratory study where the cues are presented to the subject in an abstract form and the subject makes some judgment. The cue in this study represent a *perceptual achievement* in their own right so that defining what information is being used is not clear (see, e.g., Garner, 1970, for a discussion of the stimulus in information processing). It may also be that the highly controlled study where very high correlations are not uncommon have raised our aspiration level with respect to what we expect in other decision problems of a more complex nature. A fifth factor is that this task is highly complex and difficult. This may be forcing the decision makers into "using" models other than the ones used here for approximating what might be going on. It may also be that more random responding is going on in situations where the difficulty of the task is so great.

DISCUSSION

The results found here and the methods advocated bring up some very important issues. The first is that the use of expert information or judgment can be a very useful method for getting input for a mechanical combination process. The use of the components of the global judgment, as well as the global judgment when combined with the components, may open the way to a much more expanded use of clinical measurement. It would seem that in cases where "objective" measures are not available, one has to use expert opinion or judgment. In these cases, our results argue for the quantification of the components of the judgments as well as the global judgments themselves. It is expected that the small cost involved in getting the components will be greatly outweighed by the improvement in prediction that one can obtain. In situations where both objective and expert information are available, some sort of compromise might be reached with respect to using both kinds of data in order to optimally predict some criterion.

It was stated at the outset that the methods advocated here would be general to any situation where one has a quantified criterion and judgmental (as well as objective) information. While this is obviously a broad statement, a few examples might be in order to see how one might use the techniques in situations other than a medical diagnosis task. Let us first deal with the selection of graduate students by a graduate committee of some sort made up of "experts" (professors, for example). Each judge has to evaluate the applicant, not only in some global manner (such as accept or reject), but the components of the judgment also have to be quantified. This might involve setting up rating scales for variables such as: motivation, intensity of goal attainment, rating of the essay "Why I want to go to graduate school," ratings of the quality

of the undergraduate school, evaluation of the letters of recommendation, etc. These variables could be combined with the more "objective" measures (such as GRE scores, grade-point average, etc.) but the important point here is that the components of judgment be obtained from *each judge*. One could now use the components as the predictors, and if one were interested in predicting grade-point average, for example, one could then find the best combination of variables (using an a priori mechanical combining rule) so that one could maximize the prediction of the grade-point criterion. This "experiment" could even be done on "older" cases so that one might be able to get the prediction equation without having to go through a new group of applicants. It might be found, for example, that certain judges are very good at judging motivation while others are much better at judging the essay question. If this were the case one would be taking advantage of this differential ability with regard to the different components.

Let us take a different example. Suppose that there are three different theories of job satisfaction and each theory has a different way of measuring the various components of satisfaction. In this case the theories serve as the "methods" while the components serve as the "traits." If one has a criterion measure whose variance is theoretically related to the components then one can use the relationship between the various components and the criterion as evidence about the efficacy of the theory. In this example, one might use the criterion of absenteeism as reflecting variance that should be due to satisfaction. One could now design a study where one obtained measures for the various components for each *theory*, for each subject in the study. It might be found that different components from different theoretical frameworks might best predict the criterion in question. If this were the case, this might argue for a reformulation or rebuilding of a new theoretical framework that might take into account those variables that best predicted the criterion. This method might be quite useful in the building of theory on the basis of empirical findings.

A possibility for extending the techniques discussed here involves the possible use of a "Bayesian" approach to prediction as advocated by Pankoff and Roberts (1968). They have stated,

> Whatever the method by which predictions and decisions are made, reliance upon clinical judgment is inescapable. It is often possible to incorporate clinical judgment into a statistical analysis that blends judgment with sample evidence . . . . The clinician's prior distribution for the parameters of this regression model determines the initial weighting of these judgments. The prior distribution, and therefore the weighting, is modified by the analysis of data in accordance with Bayes' theorem. The result is a posterior distribution for the parameters. The posterior distribution in turn implies

> a predictive distribution, which yields prediction for new cases. The predictive distribution thus reflects both clinical judgments and past data. For this reason we speak of a Bayesian synthesis of clinical and statistical prediction [pp. 772–773].

The blending of judgment and sample evidence might begin with the use of prior probability distributions for the judgmental components. This would seem to be a good strategy since experts should have non-diffuse priors with regard to the judgmental components. Moreover, as sample evidence modifies the priors, the judge could continue to add his judgments to the new information so that this would be similar to what Sawyer has labeled a "mechanical synthesis" (1966). It would therefore seem that a "Bayesian" approach should be very useful for predicting the criterion in question.

While there may be advantages associated with the use of components and multiple judges, there may also be costs. It may be that the cost of having more than one judge might outweigh the advantages gained in prediction. If the judge's time is highly valuable, it might not be economically feasible to have many judges perform the same task in order to pick and choose the best parts of each. This might be partially offset by the fact that the judge would only have to make his judgments on those components that are needed by the machine for the optimal combination. One might actually do a cost analysis of the relative increase in prediction as opposed to the increased cost of using several judges in any situation. One should always be aware that one is interested in increasing utility and that cost should be included in the analysis (Cronbach & Gleser, 1965).

Other problems with the use of multiple judges may be more social psychological; for example, the judges' motivation for performing the task might decrease if they cannot perform the whole task or if they feel that they are only incidental to the machine. These kinds of problems might be solved by informing the individual of the vital role he is playing in the prediction problem. While there may be many unsolved problems in using both components of judgment as well as multiple judges, it is only by doing research using these methods that we will become aware of these problems. However, it would seem that the potential utility for trying these methods is high since they offer an opportunity for increasing our ability for dealing with prediction problems in many diverse areas.

REFERENCES

BLENKNER, M. Predictive factors in the initial interview in family casework. *Social Service Review*, 1954, **28**, 65–73.

BROWN, T. The judgment of suicide lethality: A comparison of judgmental models obtained under contrived versus natural conditions. Unpublished doctoral dissertation, University of Oregon, 1970.

BRUNSWIK, E. *The conceptual framework of psychology.* Chicago: University of Chicago Press, 1952.

COOMBS, C. *A theory of data.* New York: John Wiley, 1964.

CRONBACH, L. J., & GLESER, G. C. *Psychological tests and personnel decisions.* (2nd ed.) Urbana: University of Illinois Press, 1965.

DARLINGTON, R. B. Multiple regression in psychological research and practice. *Psychological Bulletin,* 1968, **69,** 161–182.

DAWES, R. M. Social selection based on multi-dimensional criteria. *Journal of Abnormal and Social Psychology,* 1964, **68,** 104–109.

DUDYCHA, L. W., & NAYLOR, J. C. Characteristics of the human inference process in complex choice behavior situations. *Organizational Behavior and Human Performance,* 1966, **1,** 110–128.

EDWARDS, W. Dynamic decision theory and probabilistic information processing. *Human Factors,* 1962, **4,** 59–73.

EDWARDS, W., & PHILLIPS, L. D. Man as transducer for probabilities in Bayesian command and control systems. In G. L. Bryan and M. W. Shelley (Eds.), *Human judgments and optimality.* New York: Wiley, 1964.

EDWARDS, W., PHILLIPS, L. D., HAYS, W. L., & GOODMAN, B. C. Probabilistic information processing systems: Design and evaluation. *IEEE Transactions on Systems Science and Cybernetics,* 1968, Vol. SSC-4, 248–265.

EINHORN, H. J. The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin,* 1970, **73,** 221–230.

EINHORN, H. J. Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance,* 1971, **6,** 1–27.

GARNER, W. R. The stimulus in information processing. *American Psychologist,* 1970, **25,** 350–358.

GOLDBERG, L. R. Man vs. model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin,* 1970, **73,** 422–432.

GUSTAFSON, D. H. Evaluation of probabilistic information processing in medical decision making. *Organizational Behavior and Human Performance,* 1969, **4,** 20–34.

HAMMOND, K. R. Probabilistic functioning and the clinical method. *Psychological Review,* 1955, **62,** 255–262.

HURSCH, C. J., HAMMOND, K. R., & HURSCH, J. Some methodological issues in multiple-cue probability studies. *Psychological Review,* 1964, **71,** 42–60.

KLEINMUNTZ, B. The processing of clinical information by man and machine. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

LUSTED, L. B. *Introduction to medical decision making.* Springfield, Ill.: Charles C. Thomas, 1968.

MEEHL, P. E. *Clinical vs. statistical prediction.* Minneapolis: University of Minnesota Press, 1954.

PANKOFF, L. The quantification of judgment: A case study. Unpublished doctoral dissertation, University of Chicago, 1967.

PANKOFF, L., & ROBERTS, H. V. Bayesian synthesis of clinical and statistical prediction. *Psychological Bulletin,* 1968, **70,** 762–773.

PETERSON, C. R., & BEACH, L. R. Man as an intuitive statistician. *Psychological Bulletin,* 1967, **68,** 29–46.

SARBIN, T. R., TAFT, R., & BAILEY, D. E. *Clinical inference and cognitive theory.* New York: Holt, Rinehart and Winston, 1960.

SAWYER, J. Measurement *and* prediction: clinical *and* statistical. *Psychological Bulletin,* 1966, **66,** 178–200.

SLOVIC, P., & LICHTENSTEIN, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Eugene, Oregon: Oregon Research Institute, Monograph No. 1, Vol. **10,** 1970.

YNTEMA, D. B., & TORGERSON, W. S. Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics,* 1961, HFE **2** (1).

RECEIVED: January 4, 1971