PERSONNEL PSYCHOLOGY
2000, 53

MAKING DECISIONS FROM AN INTERVIEW: EXPERT MEASUREMENT AND MECHANICAL COMBINATION

YOAV GANZACH
Faculty of Management
Tel Aviv University

AVRAHAM N. KLUGER
Department of Management
The Hebrew University of Jerusalem

NIMROD KLAYMAN
Faculty of Management
Tel Aviv University

One way of making decisions on the basis of qualitative impressions is to identify a number of relevant dimensions, translate the impressions into quantitative ratings on each of the dimensions, and integrate the ratings using a mechanical combination scheme. The paper compares the output of this method to global (clinical) judgment. The basis for the comparison is a large database that includes both information collected in a structured interview and a relevant criterion. The results clearly suggest that mechanical combination outperforms clinical judgment, but also that the combination of both schemes produces the highest accuracy.

One of the strongest findings in behavioral decision making is that when the decision inputs, or cues, are in quantitative form, their mechanical combination (e.g., a weighted average of the cues, where the weights reflect cue importance) outperforms global (clinical) expert judgment in predicting real-life criteria (cf. Meehl, 1986; Sawyer, 1966). For instance, a weighted average of 11 Minnesota Multiphase Personality Inventory (MMPI) scales (based on weights derived from a regression model) outperforms the global judgments that are based on these scales in predicting the likelihood of psychosis among mental patients (Goldberg, 1965).

However, quite often the information available to the decision maker is qualitative rather than quantitative. In particular, this is typical of

To a large extent, the ideas presented in this paper were developed over the course of many years of research and practice by the psychologists of the Behavioral Science Department of the Israeli Defense Forces. We are grateful for what we learned from them. Financial support for this research was provided by the Israel Institute of Business Research, by the Israel Science Foundation, and by the Reccanati Fund.

Correspondence and requests for reprints should be addressed to Yoav Ganzach, Faculty of Management, Tel Aviv University, Tel Aviv, Israel 69978; yoavgn@post.tau.ac.il.

an interview situation, in which most of the information consists of impressions formed by the interviewer. Can the advantages of mechanical combination of cues be retained in such a situation?

Einhorn (1972) suggested a positive answer to this question. He proposed "expert measurement and mechanical combination" as a method that incorporates the need to base decisions on qualitative impressions with the advantages of mechanical combination of cues. In this method, the dimensions that are relevant to the decision (e.g., impressions about the motivation or conscientiousness of an interviewee) are identified, and are used by the experts to translate their qualitative impressions into quantitative ratings. These ratings are then combined into a *predicted score*—a single number to be used to predict the criterion. The most important combination methods are reviewed below, and their accuracy is then examined on the basis of a large, real-life database.

Optimal Weights Combination Versus Global Judgment

Optimal weights combination is a combination method in which the predicted score is a weighted average of the dimensions, where the weights are derived from a regression (OLS) model of the criterion on the dimensions.

It is often argued that optimal weights combination of expert measurement has a higher predictive accuracy than expert global judgment. This argument is based on the view that people are efficient in selecting the information that is important for making predictions, but are not efficient in integrating this information (e.g., Dawes, 1988). However, the evidence in support of this argument is usually indirect, based primarily on findings showing that human information integration is flawed with inconsistencies, misweighing and heuristic thinking (cf. Kahneman, Slovic, & Tversky, 1988; Nisbett & Ross, 1990). Direct evidence of the superiority of optimal weights combination of expert measurement over global judgment is rare. In fact, the only rigorous study to examine this issue was reported 25 years ago by Einhorn (1972). In his study, three pathologists evaluated 193 biopsy slides of Hodgkins disease patients. For each slide, they first assessed the relative amounts of nine histological dimensions, and then made a global judgment of the severity of the disease. The results of the study showed that optimal weights combination was more accurate than global judgment in predicting the criterion (the patient's survival time). The cross-validated multiple correlation between the dimensions and the criterion was higher than the correlation between the global judgment and the criterion. Whereas the former correlation was about .3, and the latter was about zero.

However, the evidence of Einhorn's (1972) study is not strong enough to conclude that optimal weights combination is generally more accurate than global judgment. First, from an empirical point of view, the low correlation between the global judgment and the criterion in Einhorn's study is problematic. Even if optimal weights combination is superior to global judgment, the absolute lack of accuracy of the global judgment in Einhorn's study is unlikely to be representative of many expert prediction tasks. Second, from a theoretical point of view, there are a number of reasons why global judgment may outperform optimal weighting. It is possible that relevant information associated with missing dimensions—dimensions not rated by the judge—is incorporated into the global judgment; such information cannot be incorporated into optimal weights combination of expert measurement. It is also possible that information relevant to the criterion is associated with certain unique details, which cannot be captured by any dimension (e.g., information about an individual's broken leg is important in predicting whether he will go to the movies, yet it cannot be captured by any general dimensions underlying preference towards going to the movies. See Meehl, 1954). Furthermore, even if all the relevant information is captured by the measured dimensions, global judgment may outperform optimal weights combination because it can take into account complex (configural, nonlinear) relationships which are not typically taken into account by a mechanical combination. Finally, studies of behaviorally anchored rating scales indicate that global judgment may be more reliable, and therefore more valid, than component judgment (Heneman, 1988; but see Fay & Latham, 1982), perhaps because the cognitive processes involved in forming a global judgment require less effort than those involved in forming component judgments (e.g., Lingle & Ostrom, 1979).

Global judgment versus optimal weights combination in a structured interview. There is substantial similarity between Einhorn's (1972) experiment and a structured interview. In both, the decision maker has to base his or her decision on qualitative impressions of target stimuli (i.e., biopsy results, interviewee's behavior). In both, it is possible to identify a number of dimensions that are relevant to the criterion, and in both it is possible to ask the decision maker, after examining the stimulus, to provide quantitative assessments of these dimensions.

In fact, the importance of identifying and quantifying relevant dimensions has already been recognized in the interview literature in general and in the employment interview literature in particular. Structured interviews often involve standard questions that attempt to evaluate the interviewee along predetermined relevant dimensions, and often involve the quantification of these dimensions by the interviewers using rating

scales (e.g., Huffcutt & Arthur, 1994). However, it is not clear what exactly are the features of a structured interview that lead to a superior accuracy (see Huffcutt & Arthur, 1994). Thus, one purpose of this paper is to compare the accuracy of optimal weights combination in a structured interview to the accuracy of global judgment.

Optimal Weights and Global Judgment Combination

In the previous section, optimal weights combination of expert measurement was discussed as an alternative to global judgment. However, the two methods can be used together. The expert's global judgment can be (mechanically) combined with the dimensions, using, for example, regression weights of a model that includes both the dimensions and the global judgment. This method (to be labeled optimal weights + global judgment combination) may be more accurate than a simple optimal weights combination because global judgment can add predictive accuracy to the dimensions, accuracy associated with configural relationships, missing dimensions, and "broken leg" rules (see the discussion above).

Einhorn (1972) examined whether adding the global judgment to the dimension ratings improves accuracy, but—because of the small sample size—was unable to draw any firm conclusions. For two of the judges, adding the global judgment decreased the (cross-validated) accuracy, but for one judge it increased accuracy. The current study overcomes this limitation by using an unusually large sample of interviewers and interviewees.

Expert Measurement and Mechanical Combination When Criterion Information is Unavailable

As described so far, expert measurement and mechanical combination does require criterion information. However, in many real-life situations, there is scant, if any, information about the criterion to generate credible weights for a mechanical combination of expert measurement. There are two methods that could be used in these situations. One is *unit weight combination*, in which each dimension has the same weight as the other dimensions. In the other method—to be labeled, following Goldberg (1970), *model of man combination*—one first regresses the *global judgment* on the dimensions, and then uses the weights obtained from this regression to mechanically combine the dimensions (this method is often called "bootstrapping"). Neither method requires criterion information; however, whereas unit weight combination requires only an identification of the relevant dimensions, the model of man combination

requires actual judgments—the larger the number of these judgments, the more credible the weights. Furthermore, whereas the accuracy of the unit weight combination does not depend on judges' ability to correctly integrate dimension information, the accuracy of the model of man combination does depend on this ability. If judges do integrate dimension information correctly, but not if they do not, the model of man combination should have a higher accuracy than unit weight combination. In this case, the model of man combination often has a higher accuracy than the global judgment. Indeed, a number of researchers found that bootstrapping the global judgments leads to more accurate predictions than using the global judgments themselves (e.g., Camerer, 1981; Dawes, 1976).

Expert Measurement, Global Judgment, and the Fine Tuning of Predicted Scores

An important advantage of expert measurement may be that it is more finely tuned than global judgment. Because judges' discriminative ability is limited (e.g., Lissitz & Green, 1975; Ramsey, 1973), rating scales in general, and the rating scales used in structured interviews in particular, include only a small number of categories. This can result in the predicted score of expert measurement and mechanical combination being more finely tuned than the global judgment. Consider, for example, a case in which both the rating of the global judgment and the rating of the dimensions is performed on a 5-point rating scale. In this case, the predicted score obtained from combining the dimensions is a continuous measure, having many possible levels, but the global judgment is more discrete, having only five possible levels. This produces two advantages to expert measurement and mechanical combination over global judgment. First, it increases overall accuracy, by decreasing the coarseness of the measure (see Russell & Bobko, 1992, for a related discussion). Second, it permits the division of the interviewed population into smaller, more homogenous, groups; if the rating scale of the global judgment is a 5-point scale, it is possible to divide the population into only five groups. On the other hand, expert measurement and mechanical combination allows the division of the population into smaller groups such as deciles or percentiles. From a practical point of view, this division is desirable. For example, it is desirable in a situation in which an expensive treatment is considered cost-effective only for a small extreme group.

Overview of Research Questions

A structured interview in which responses are scored along predetermined dimensions provides an ideal environment for examining the

accuracy of expert measurement and mechanical combination, because expert measurement is a built-in feature of this task. Thus, one purpose of the current study is to examine the highly recommended, yet little researched, method of expert measurement and mechanical combination on the basis of real-life data which are routinely collected in the placement process of a large organization. Another purpose of the study is to compare various methods of expert measurement and mechanical combination, and to compare these methods with global judgment, in order to provide the users of interviews with ideas about designing structured interviews, and about using the information they elicit.

Within this context, the analyses below are organized around four issues. First, the accuracy of optimal weights combination is compared to the accuracy of global judgment. Second, the accuracy of these two methods is compared to the accuracy of optimal weights + global judgment combination. Third, the accuracy of unit weights combination and the model of man combination is examined. Finally, the effect of fine tuning on accuracy and on ability to divide the population into smaller, more homogenous, groups is evaluated.

A general framework for studying these questions is the Brunswikian idea of the lens model, in which accuracy is dissected by estimating both models of the judgment and models of the criterion. Indeed, in a review of the literature, Guion (1991) suggested this framework as a general framework for studying selection and placement problems. But Guion's review also indicates that only two studies applied this framework to empirical data, one by Dougherty, Ebert, and Callender (1986) and the other by Zedeck, Tziner, and Middlestadt (1983). However, these studies were based on a small number of interviewers (3 and 10, respectively), small numbers of interviewees (120 and 412, respectively), lack of criterion information (such information was available for only 60 and 132 interviewees, respectively), a "soft" criterion measure (performance evaluation), and low interview accuracy (only one interviewer in the former study, and none of the interviewers in the latter study, exhibited accuracy which was significantly different from zero). Furthermore, because of the small sample size, these studies examined only the accuracy of models of the *judgment*, but did not examine the accuracy of models of the *criterion*. The current study is based on a large database, with sufficient interview accuracy and a "hard" behavioral criterion. This allows firmer conclusions to be drawn about the accuracy of the various methods of making decisions in a structured interview.

Finally, the literature about the validity of assessment centers is also relevant to the issues discussed here. A number of studies compared the consensus overall evaluations of assessors to evaluations derived from mechanical combinations. In some of these studies the inputs for the

mechanical combination were the overall judgments of the individual assessors (e.g., Pynes & Bernadin, 1992; Sackett & Wilson, 1982), and in others they were a variety of test scores and background (e.g., demographic) variables (Borman, 1982; Feltham, 1988; Mitchel, 1975; Tziner & Dolan, 1982). Note, however, that these assessment center studies did not involve the mechanical combination of dimension ratings, which is the focus of the current study. Furthermore, the sample sizes in these studies were insufficient. They were, therefore, inadequate for detection of significant differences between the accuracy of global judgments and the accuracy of mechanical combination; and they did not allow for appropriate cross validation, resulting in upwardly biased accuracy measures of mechanical combination.

Method

Participants. The participants were 26,197 male prospects for military service in the Israeli army who were subsequently drafted. Most of the participants were 18. Because military service in Israel is mandatory, this sample is fairly representative of the Israeli male population in this age group. There were 116 interviewers, and each interviewee was interviewed by one interviewer. The number of interviewees per interviewer ranged from 41 to 697 and the standard deviation was 132.

The interview. Each prospective conscript to the Israeli army routinely goes through an initial interview. This interview, roughly 20 minutes in length, is administered by specialists, who have undergone a 3-month training course. The results of the interview are an important input for the selection and placement decisions concerning the interviewees. The interview was launched about 30 years ago, and much effort has been invested by the military in its design and implementation. Its accuracy is constantly monitored, and based on accumulated experience, changes in its content and method of implementation are periodically introduced (a detailed description of the interview and its role in the army's placement system is given in Gal, 1986). Note, however, that the data analyzed here were derived from a single version of the interview, and were collected in a relatively short period of time.

Dimension ratings. Six traits were assessed in the interview: activity, pride in service, sociability, responsibility, independence, and promptness. These traits were rated on a 5-point scale, where 5 represents a high level of the trait and 1 a low level. Interviewers were provided with specific guidelines, documented in a detailed manual, on how to translate interviewees' verbal responses into numerical values. Note that among other guidelines in this manual, interviewers were specifically instructed to rate each dimension independently of the other dimensions.

Global judgment. In addition to rating the dimensions, the interviewers also made an overall evaluation of the expected success of the prospect in his military service. This global judgment was also given on a 5-point rating scale where 1 meant low probability of success and 5 meant a high probability of success. Interviewers were instructed to provide a rating of expected success that would reflect their general impression, and were specifically told that "it is possible that the expected success of an interviewee whose dimension ratings are high will be low, and vice versa." Nevertheless, it is likely that the global judgment was influenced, to some extent, by the prior dimension ratings. Thus the global judgment in our study is not entirely independent of the dimension ratings. Note, however, that this situation is typical for structured interviews, because in such interviews both dimension ratings and global judgments are given by the same interviewer.

The criterion. The number of deficiencies—disciplinary transgressions, such as desertion or imprisonment—was used as a criterion. These transgressions were recorded during 3 years of compulsory military service. Because the distribution of this variable is very skewed—83% of the sample did not have any deficiency (with 7.2%, 4.3%, 2.2%, 1.3%, 0.8%, 0.4%, and 0.5% for 2 through 7 deficiencies, respectively)—it was treated in the analyses as a dichotomous variable by dividing the sample into participants who had 0 deficiencies and participants with 1+ (one or more) deficiencies. Note that our conclusions do not depend on the dichotomization of the dependent variable; similar results are obtained when it is treated as a continuous variable. Note also that in calculating the correlations below, we assigned the value of 1 to participants who had 0 deficiencies and the value of 0 to participants who had 1+ deficiencies, so that the relationships between the independent variables and the dependent variable would be positive. That is, after this transformation, the criterion is a measure of participants' success in the military.

Results

Method of Analysis

The simplest measure of the accuracy of each of the methods is the correlation between its predicted score and the criterion. For the expert judgment method, this measure of accuracy is simply the correlation between the global judgment of the interviewee and the criterion. For the methods that involve mechanical combination, this measure is the (double) *cross-validated* multiple correlation between the predicted score of the rated dimensions and the criterion. This correlation is obtained by splitting the total sample into two sets, calculating the regression weights

of each of the sets, using the weights to calculate a predicted score for each participant in the other set, correlating this predicted score with the criterion, and averaging the two correlations. Note that when the sample size is large, shrinkage is small, and the accuracy measures of the methods that involve mechanical combination practically equal the multiple correlation between the predictors and the criterion.

There are two possible approaches to the analysis. One is based on individual models. The regressions are performed within each interviewer, the interviewers' accuracies are averaged and the mean accuracies of the various methods are compared using a test for difference between means of dependent observations. The other approach is based on an aggregate analysis. The individual differences between interviewers are ignored, and the analysis is performed across all the interviewees, resulting in correlations which are compared using a test for dependent correlations. The advantage of the individual analysis is that individual differences in predictive ability are explicitly introduced into the analysis. However, because of the relatively small sample size available for each interviewer (an average of 225 per interviewer), the error in estimating the parameters of the individual models is large. The aggregate approach, on the other hand, ignores individual differences in accuracy, but is characterized by a very small estimation error. We have chosen to present the results primarily in terms of the aggregate analysis, because this represents the global point of view of a decision maker who has to decide among various methods of obtaining a predicted score, and because of the simplicity of presentation. However, a brief description of the individual analyses is presented as well.

Finally, because the assumption of normal distribution of the criterion is seriously violated, significance tests involving correlations with the criterion may be biased. Therefore, we used the following procedure to test the null hypothesis that the difference between the accuracy of any two prediction methods is equal to zero. We standardized the 26,197 predicted scores of each method and calculated, for each method, the mean predicted score for the group that had 1+ deficiencies and the group that had 0 deficiencies. For each method, the difference between the two means is a measure of the accuracy of the method, the gap between the differences of any pair of methods is a measure of the difference in accuracy between the two methods, and the t-test for the hypothesis that this gap is equal to zero is the appropriate test for the null hypothesis that the methods are equally accurate. A formal description of this procedure is provided in the Appendix.

Descriptive Statistics

Table 1 presents the means, standard deviations, and intercorrelations among the variables. One interesting feature of the data in this table is the low intercorrelations among the dimension ratings. These correlations reflect the guidelines of the interview which were described above, and suggest that the ratings do indeed measure distinct underlying dimensions rather than an overall impression of the prospect. Therefore, it *cannot* be argued that expert measurement and mechanical combination, which involves six measurements, is simply a more reliable, and therefore more valid, measure of the same construct that is measured by the global judgment (which involves only one measurement).

Another interesting feature of the data in Table 1 is the relatively low correlation between the criterion and the global judgment, as well as the low correlations between the criterion and the dimensions. These low correlations are relevant to the interpretation of the results reported below. If the correlation between the interview results and the criterion is low, then differences between the accuracy of the various methods for deriving a predicted score from the interview should not be high (Schmitt & Levine, 1978). Thus, because of the large sample size, we expect significant, but only moderate, differences between the predictive accuracies of the various methods. Note, however, that the correlations between the predictors and the criterion are conservative estimates of the true correlations, because restriction of range is likely to attenuate the interview accuracy, and because our database contains only interviewees that were subsequently drafted.

Overview of the Results

Table 2 presents the measures of accuracy of the various methods. Column 2 presents the correlations between predicted scores and the criterion. This correlation allows easy comparisons between the methods. Columns 3 and 4 present, respectively, the mean standardized predicted score of the group that had 0 deficiencies and the group that had 1+ deficiencies ($\overline{PS}0$ and $\overline{PS}1$, respectively). The difference between these means, $\Delta\overline{PS}$, is presented in column 5. Like the correlation in column 2, it is a measure of the method's accuracy. It is also the basis of the significance tests for the differences in accuracy between the methods (see Appendix).

Table 2 indicates that the methods that use criterion information are more accurate than the methods that do not use this information. Among the methods that use criterion information, the accuracy of optimal weights + global judgment combination clearly exceeds the accu-

TABLE 1

Means, Standard Derivations, and Intercorrelations Among the Variables

	M	*SD*	1	2	3	4	5	6	7
1. Criterion	0.83								
2. Global evaluation	2.27	0.86	0.230						
3. Sociability	2.88	0.68	0.067	0.548					
4. Independence	3.12	0.57	−0.020	0.221	0.336				
5. Activity	2.98	0.89	0.192	0.459	0.306	0.044			
6. Promptness	3.01	0.70	0.156	0.330	0.147	0.038	0.247		
7. Responsibility	2.90	0.80	0.237	0.415	0.119	−0.024	0.311	0.505	
8. Motivation	3.19	1.09	0.147	0.449	0.249	0.077	0.319	0.182	0.253

Note: All the correlations in the table are significantly different from zero at the .0001 level.

TABLE 2
The Accuracy of Various Methods

Method	Accuracy	$\overline{PS0}$	$\overline{PS1}$	$\Delta\overline{PS}$
Global judgment	.230	.103	−.515	.618
Optimal weights combination	.276	.124	−.619	.743
Optimal weights + global judgment	.297	.133	−.664	.797
Unit weight combination	.236	.106	−.528	.634
Model of man combination	.216	.096	−.482	.578

Note: $\overline{PS0}$ and $\overline{PS1}$ are the mean *standardized* predicted score of the group that had 0 deficiencies and the group that had 1+ deficiencies respectively. All the accuracies in this table are significantly different from zero at the .0001 level.

racy of the combination method that is based solely on optimal weights. Among the methods that do not use criterion information, the model of man combination is clearly less accurate than either unit weight combination or global judgment.

Global Judgment Versus Expert Measurement and Mechanical Combination

The results of Table 2 indicate that the accuracy of the optimal weights combination was significantly greater than that of the global judgment (this test, as the rest of the tests reported below, are t-tests of the null hypothesis that the difference between $\Delta\overline{PS}$ of the two methods is equal to zero with $\alpha = 0.0001$ and $df = 26{,}195$). The cross-validated multiple correlation between the dimensions and the criterion was .276, whereas the correlation between the global judgment and the criterion was only .230. These results replicate Einhorn's (1972) results in showing that optimal weights combination outperform global judgment. Note however, that the gap between the accuracy of the global judgment and the accuracy of the optimal weights combination in Einhorn's data was larger than in our data. To a large extent, this is due to the lack of accuracy of the global judgment in Einhorn's study, and the fact that in the current study, the accuracy of the global judgment—although significantly lower than that of the optimal weights combination—is substantial.

Adding the Global Judgment

Adding the global judgment to the dimension ratings increases accuracy. A combination rule that included the dimensions as well as the global judgment resulted in a significant improvement in accuracy over a combination rule that included only the dimensions. The cross-validated multiple correlation between the criterion and the predicted score of the

model that included both the dimensions and the global judgment was .297.

When Criterion Information is Unavailable

The two methods of expert measurement and mechanical combination that do not use criterion information perform rather poorly relative to the methods that use this information. The accuracy of unit weight combination is .236 and the accuracy of the model of man combination is .216 (Table 2 also presents the values of $\overline{PS0}$, $\overline{PS1}$ and $\Delta\overline{PS}$ of the two methods).

A more interesting finding is that the model of man combination, which uses interviewers' knowledge (i.e., their subjective weights), does worse than unit weight combination, which does not use interviewers' knowledge. This finding is rather surprising because the interviewers are well-trained and experienced, and because the ratio of observations to predictors is very large. (When this ratio is small, estimation errors involved in the model of man combination may detract from the advantage it may have over unit weight combination. See Dawes, 1979.) To give some insight into the processes that lead to this phenomenon, Table 3 contrasts the standardized coefficient of the criterion and the judgment models. Two biases in the interviewers global judgments are evident in this table. First, whereas sociability is not at all associated with the criterion, keeping other dimensions constant, it has by far the largest effect on the global judgment. Second, whereas independence is negatively associated with the criterion, it has—keeping other dimensions constant—a positive effect on the global judgment. These two discrepancies explain why, in our study, model of man combination is less accurate than either unit weight combination or—contrary to other findings (e.g., Goldberg, 1970; Dougherty et al., 1986)—global judgment. It appears that this combination method introduces interviewer biases into the calculation of the predicted score.

Controlling for Cognitive Ability

Because general cognitive ability was shown to be a very powerful predictor of performance (Gottfredson, 1986; Hunter & Hunter, 1984) and because our data contained information about the draftees general cognitive ability, we examined the accuracies of each of the methods controlling for cognitive ability. These accuracies were, respectively, .149, .209, .222, .161, and .142, for global judgment, optimal weights combination, optimal weights + global judgment, unit weight combination, and the model of man combination. Thus, the rank order of these accuracies

TABLE 3

Standard Regression Coefficient of the Global Judgment and the Criterion Models

	Criterion model	Judgment model
Sociability	0.001	0.375
	(0.007)	(0.005)
Independence	−0.028	0.072
	(0.006)	(0.005)
Activity	0.112	0.183
	(0.007)	(0.005)
Promptness	0.032	0.074
	(0.007)	(0.005)
Responsibility	0.168	0.221
	(0.007)	(0.005)
Motivation	0.065	0.222
	(0.006)	(0.005)

Note: Numbers in parentheses are standard errors of the estimates.

TABLE 4

The Average Accuracy of the Individual Interviewers for Each of the Methods

Method	Mean accuracy (unweighted)	*SD*	Range	Mean accuracy (weighted)
Global judgment	0.241	0.079	−0.04 to +0.48	0.240
Optimal weights combination	0.220	0.138	−0.33 to +0.56	0.247
Optimal weights + global judgment	0.236	0.138	−0.25 to +0.71	0.258
Unit weight combination	0.242	0.078	−0.04 to +0.47	0.243
Model of man combination	0.217	0.076	−0.21 to +0.43	0.221

is similar to the rank order of the uncontrolled accuracies, suggesting that controlling for cognitive ability does not alter the conclusions that can be drawn from the previous analyses.

Individual Interviewers Accuracy

Table 4 presents the average accuracy of the individual interviewers for each of the methods, as well as the standard deviations and ranges of these accuracies. In calculating these averages, regression models and accuracies are calculated for each interviewer, and the correlations of the various methods are averaged across interviewers.

It is clear from the results of Table 4 that, with regard to average individual accuracy, the methods that use criterion information are relatively *in*accurate. Optimal weights combination as well as optimal weights + global judgment combination are less accurate than either the global judgment or unit weight combination (the mean validities of the four

methods are, respectively, .220, .236, .241, and .242). The reason for this is most likely the instability in estimating dimensions' weights on the basis of the relatively small sample size of interviews available for each interviewer. For example, the correlation between sample size and interviewers' optimal weights' accuracy is .35. Note also that—consistent with this estimation instability explanation—the standard deviations of the individual accuracies of the methods that use criterion information, where accurate estimation of weights is particularly important, are much larger than the standard deviations of the other methods (see column 3 of Table 4).

One way to correct for this estimation instability in calculating true individual accuracies is to weight the accuracy of each interviewer by the number of interviews she conducted. The weighted means of these accuracies are given in the last column of Table 4. Indeed, in contrast to the pattern of the unweighted means, the pattern of these weighted means is similar to the pattern of the aggregate accuracies.

Finally, note that whereas the individual accuracies of the methods that use criterion information decrease *vis-à-vis* the aggregate accuracies, the individual accuracies of two of the methods that do not use this information—global judgment and unit weight combination—increase. This gain in accuracy is probably due to the fact that there are idiosyncratic differences between the interviewers in using the rating scales (see Dreher, Ash, & Hancock, 1988). The third method that does not use criterion information—the model of man combination—does not change much, most likely because it is also vulnerable to estimation instability.

Fine Tuning and Expert Measurement

We first show that expert measurement and mechanical combination permits the division of the interviewed population into smaller, more homogenous, groups. We do this within the context of the following selection problem. Suppose that the recruitment of prospects with 1+ deficiencies is undesirable, and the army desires to omit the maximum number of such prospects, without omitting too many 0-deficiency prospects. Using the global judgment, one would have to omit the bottom group of prospects who received a global judgment of 1. In this group the percentage of 1+ deficiencies is 31.1%. However, by eliminating this group, a large number of 0-deficiency prospects (68.9%) are eliminated as well, which may make the elimination undesirable.

On the other hand, using expert measurement and mechanical combination, it is possible to select smaller, more homogenous groups. For example, if the prospects are ranked by the predicted score obtained from optimal weights combination, and the bottom 5% is selected, the

percentage of 1+ deficiencies within this group is 65.3%, and only 35.7% 0-deficiency prospects will be lost by eliminating this group. If the predicted score of the model that includes the dimensions *and* the global judgment is used, the percentage of 1+ deficiencies in this bottom group is even higher, and even fewer 0 deficiency prospects are omitted.

Discussion

It is often suggested that in order to make better decisions, the decision maker should use the following procedure: (a) identify the relevant dimensions, (b) evaluate the alternatives on each of the dimensions, (c) estimate the weights of the dimensions, and (d) integrate the values of the dimensions to arrive at an overall judgment for each of the alternatives. This is essentially what Einhorn (1972) labeled expert measurement and mechanical combination. However, despite its popularity, there has been so far very little research on the accuracy of this decision procedure, perhaps because of the difficulty of obtaining criterion information against which expert measurement and mechanical combination can be examined. The major finding of the current paper supports the common belief that this procedure is more accurate than a "Gestalt" global judgment.

Nevertheless, the difference in accuracy between expert measurement and mechanical combination and global judgment in the current study is not large. It is substantially smaller than the difference in Einhorn's (1972) study. Using our data it is also hard to examine whether the differences in accuracy between the methods have any practical importance, partly because the base rate for 0-deficiencies is high. However, the practical importance of expert measurement in enabling more finely tuned decisions is clear enough. Using this method, it is possible to identify extreme groups, that are quite different from the rest of the population with regard to the criterion. This is not possible with global judgment, more because of the low discrimination of this method than because its low accuracy.

Apart from demonstrating that, in a structured interview, expert measurement and mechanical combination is more accurate than global judgment, we find that combining the two methods increases accuracy over each of them separately. This result is consistent with recent findings reported by Blattberg and Hoch (1990), who showed that incorporating the global judgment with the dimension information results in a substantial increase in accuracy. However, these findings are not consistent with Einhorn's results that the addition of global judgment does not improve accuracy. Note that Einhorn's (1972) study is more similar in design to the current study than to Blattberg and Hoch's, because in

the latter study dimension values were not derived from expert measurement, but from objective measures.

There is considerable interest in the literature in the accuracy of structured interviews. Most of this research has focused on comparing the accuracy of structured interviews to the accuracy of unstructured interviews, concluding that, by and large, structured interviews have greater accuracy (e.g., Marcheese & Muchinsky, 1993; McDaniel, Whetzel, Schmidt, & Maurer, 1994; Wiesner & Cronshaw, 1988). Though the current research does not compare structured and unstructured interviews, it does suggest that dimension ratings, which are often collected in structured interviews, could be used to further enhance accuracy by estimating dimension weights and mechanically combining them. This idea is consistent with Huffcutt and Arthur's (1994) finding that increase in structure is associated with greater accuracy, because in highly structured interviews, dimension ratings are often collected, and used either implicitly (by influencing the interviewer's global evaluation) or explicitly (as an input to an actuarial formula).

Yet, interviews that emphasize the multidimensionality of the information are not necessarily superior to interviews that focus on a unidimensional evaluation. Unidimensional interviews, although *in*appropriate for situations such as adjustment to army life, in which the criteria are complex and multidetermined, may be appropriate for situations in which the criterion is simple. Thus, whereas for complex criteria, designing multidimensional interviews, and using this multidimensionality to calculate predicted scores may increase interview accuracy, for simple criteria the designing of such interviews may be unnecessary. When the criteria are simple, the focus should be on assessing a single relevant dimension, perhaps through numerous questions to achieve high reliability. Campion, Campion, and Hudson (1994) provide a good example of such a situation in a study that examined the accuracy of a unidimensional structured interview attempting to predict a relatively simple criterion (performance of machine operators). Nevertheless, in many work settings, the criteria of interest are complex enough to justify expert measurement of relevant dimensions accompanied by mechanical combination.

REFERENCES

Blattberg RC, Hoch SJ. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science, 36*, 887–899.

Borman WC. (1982). Validity of behavioral assessment for predicting military recruiter performance. *Journal of Applied Psychology, 72*, 463–474.

Camerer C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance, 27*, 411–422.

Campion MA, Campion JE, Hudson PI. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology, 79*, 998–1002.

Dawes RM. (1979). The robust beauty of improper models in decision making. *American Psychologist, 35*, 571–582.

Dawes RM. (1988). *Rational choice in an uncertain world*. New York: Harcort Brace Jovanovich.

Dougherty TW, Ebert RJ, Callender JC. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71*, 9–15.

Dreher GF, Ash RA, Hancock P. (1988). The role of the traditional research design in underestimating the validity of the employment interview. PERSONNEL PSYCHOLOGY, *41*, 315–327.

Einhorn HJ. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7*, 86–106.

Fay CH, Latham GP. (1982). Effects of training and rating scales on rating errors. PERSONNEL PSYCHOLOGY, *35*, 105–116.

Feltham R. (1988). Assessment centre decision making: Judgmental versus mechanical. *Journal of Occupational Psychology, 61*, 237–241.

Gal R. (1986). *A portrait of the Israeli soldier*. New York: Greenwood Press.

Goldberg LR. (1965). Diagnosticians versus diagnostic signs: The diagnosis of psychosis versus neurosis from the MMPI. *Psychological Monographs, 79*.

Goldberg LR. (1970). Man versus model of man: A rationale plus some evidence of improving clinical inference. *Psychological Bulletin, 73*, 422–432.

Gottfredson LS (Ed.). (1986). The *g* factor in employment. *Journal of Vocational Behavior, 29*, 311–319.

Guion RM. (1991). Recruitment, job choice, and posthire consequences: A call for new research directions. In Dunette MD, Hough LM (Eds.), *Handbook of industrial and organizational psychology* (Vol. 2, pp. 621–746). Palo Alto, CA: Consulting Psychologists Press.

Heneman RL. (1988). Traits behaviors and rater training: Some unexpected results. *Human Performance, 1*, 85–98.

Huffcutt AI, Arthur WJ. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry level jobs. *Journal of Applied Psychology, 79*, 184–190.

Hunter JE, Hunter RF. (1984). Validity and utility of alternative predictors of job behavior. *Psychological Bulletin, 96*, 72–98.

Kahneman D, Slovic P, Tversky A (Eds.). (1988). *Judgment under uncertainty: Heuristics and biases*. London: Cambridge University Press.

Lingle JH, Ostrom TM. (1979). Retrieval selectivity in memory-based impression judgments. *Journal of Personality and Social Psychology, 37*, 180–194.

Lissitz RW, Green SB. (1975). The effects of the numbers of scale points on reliability: A monte carlo approach. *Journal of Applied Psychology, 60*, 10–13.

Marcheese MC, Muchinsky PM. (1993). The validity of the employment interview: A meta analysis. *International Journal of Selection and Assessment, 1*, 18–26.

McDaniel MA, Whetzel DL, Schmidt FL, Maurer S. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 499–516.

Meehl P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota.

Meehl P. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50*, 370–375.

Mitchel JO. (1975). Assessment center validity: A longitudinal study. *Journal of Applied Psychology, 60*, 573–579.

Nisbett R, Ross L. (1990). *Human inference: Strategies and shortcomings of social judgment*. Engelwood Cliffs, NJ: Prentice Hall.

Pynes J, Bernardin HJ. (1992). Mechanical versus consensus derived assessment center ratings: A comparison of job performance validities. *Public Personnel Management, 21*, 17–28.

Ramsey JO. (1973). The effects of the number of categories in rating scales on precision of scale values. *Psychometrika, 38*, 152–164.

Russel CJ, Bobko P. (1992). Moderated regression analysis and Likert type scales: Too coarse for comfort. *Journal of Applied Psychology, 77*, 336–342.

Sackett PR, Wilson MA. (1982). Factors affecting consensus judgment processes in managerial assessment centers. *Journal of Applied Psychology, 67*, 728–736.

Sawyer J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66*, 178–200.

Schmitt N, Levine RL. (1978). Statistical and subjective weights: Some problems and proposals. *Organizational Behavior and Human Performance 20*, 15–30.

Tziner A, Dolan S. (1982). Validity of an assessment center for identifying future female officers in the military. *Journal of Applied Psychology, 67*, 728–736.

Wiesner W, Cronshaw S. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*, 275–290.

Zedeck S, Tziner A, Middlestadt SE. (1983). Individual validity and reliability: An individual analysis approach. PERSONNEL PSYCHOLOGY, *36*, 355–370.

APPENDIX

If $\overline{PS0_i}$ and $\overline{PS0_j}$ are the mean predicted scores of measures i and j in the group that had 0 deficiencies, and $\overline{PS1_i}$ and $\overline{PS1_j}$ are the mean predicted scores of measures i and j in the group that had 1+ deficiencies, then the difference in predictive power between the two measures could be examined by testing the null hypothesis that $\overline{G_{ij}}$ is equal to zero where $\overline{G_{ij}}$ is defined as:

$$\overline{G_{ij}} = (\overline{PS0_i} - \overline{PS1_i}) - (\overline{PS0_j} - \overline{PS1_j}) \quad [1]$$

Rearranging gives:

$$\overline{G_{ij}} = (\overline{PS0_i} - \overline{PS0_j}) - (\overline{PS1_i} - \overline{PS1_j}) \quad [2]$$

but:

$$\overline{PS0_i} = \frac{1}{N} \sum_{k \in N} PS_{ik} \quad [3]$$

$$\overline{PS0_j} = \frac{1}{N} \sum_{k \in N} PS_{jk} \quad [4]$$

$$\overline{PS1_i} = \frac{1}{M} \sum_{k \in M} PS_{ik} \quad [5]$$

$$\overline{PS1_j} = \frac{1}{M} \sum_{k \in M} PS_{jk} \quad [6]$$

where PS_{ik} and PS_{jk} are, respectively, the standardized predicted scores of measures i and j for participant k, and N and M are, respectively, the number of participants in the group that had 0 or 1+ deficiencies. Note that in [3] and [4] the summation is over the participants who had 0 deficiencies, whereas in [5] and [6] the summation is over the participants who had 1+ deficiencies. Note also that, using the definition of $\Delta\overline{PS_i}$ from Table 1, $\overline{G_{ij}} = \Delta\overline{PS_i} - \Delta\overline{PS_j}$.

Defining G_{ijk} as the gap between predicted score i and predicted score j for candidate k, $G_{ijk} = PS_{ik} - PS_{jk}$. That is, G_{ijk} is a random variable representing the difference between the two predicted scores.

Using this definition of G_{ijk} and substituting [3], [4], [5], and [6] into [2], $\overline{G_{ij}}$ can be expressed as

$$\overline{G_{ij}} = \frac{1}{N} \sum_{k \in N} G_{ijk} - \frac{1}{M} \sum_{k \in M} G_{ijk} \quad [7]$$

This suggests that the hypothesis that $\overline{G_{ij}}$ is different from zero can be tested using a t-test for the difference in G_{ijk} between the group that had 0 deficiency and the group that had 1+ deficiency.