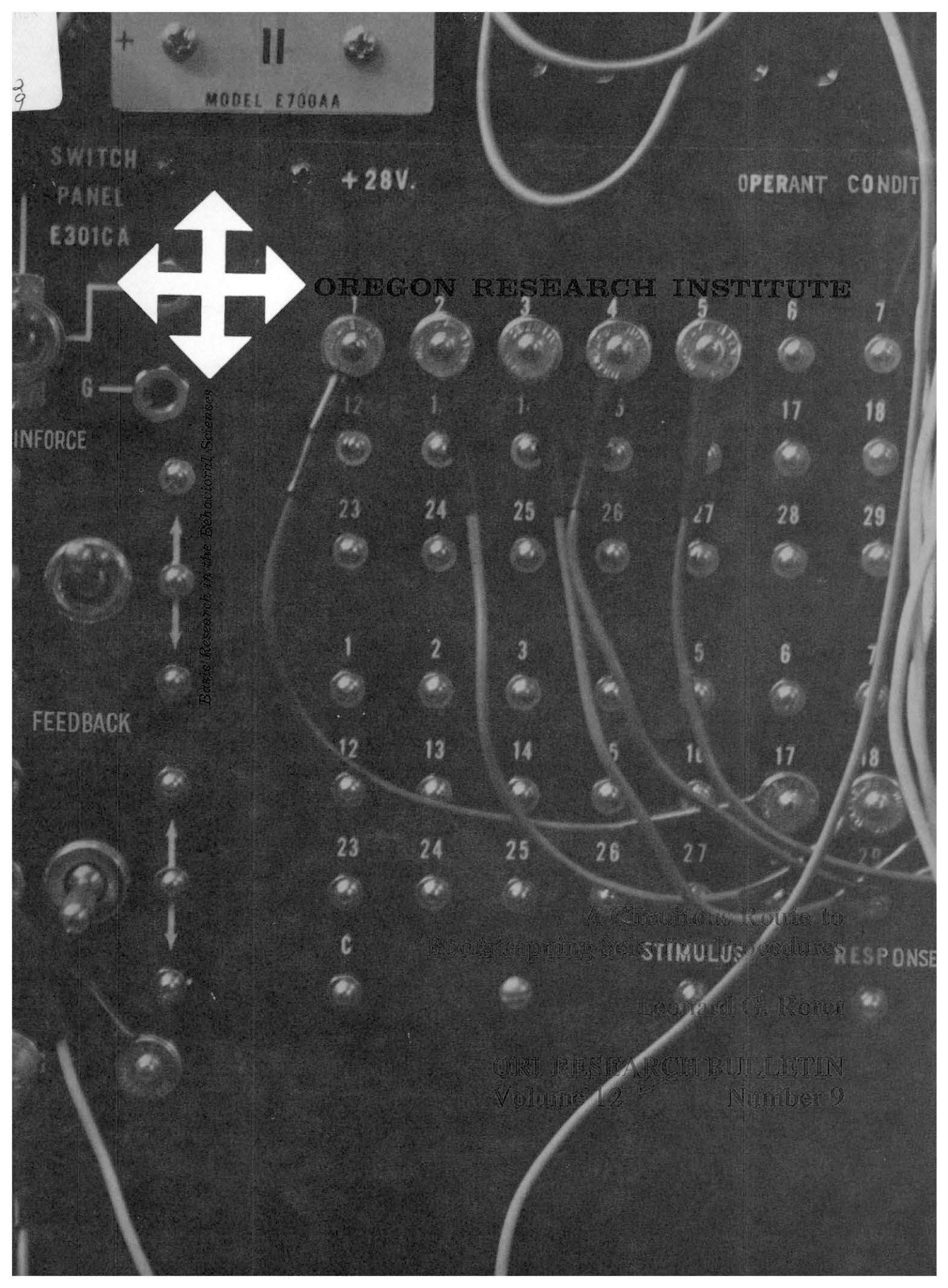

A Cautionary Route to Bootstrapping Selection Procedures

Leonard C. Rorer

ORI RESEARCH BULLETIN
Volume 12 Number 9

A CIRCUITOUS ROUTE TO BOOTSTRAPPING SELECTION PROCEDURES[1]

Leonard G. Rorer[2]

Oregon Research Institute[3]

In considering personality measurement in medical education, we are interested in measures that can be used for some purpose, such as selection, placement, counseling, or program design. If measures are to be useful in this sense, then we must be able to use them to predict something with greater accuracy than we would be able to achieve without them. There are at least two ways to go about improving the accuracy of one's predictions. The first is to improve the measures on the basis of which the predictions are made. Earlier in this conference Jackson ( ), Stein ( ), Fiske ( ), and Sechrest ( ) have already suggested ways in which this might be done. The second is to improve the way in which predictions are made on the basis of those measures which are available. It is this second possibility that I am going to discuss. The general logic of what I will say applies to any of the uses that might be made of personality measurements, but, because it is easier to discuss one procedure at a time, I will talk only about selection.

Selection procedures are usually divided into two broad categories, characterized as clinical decision making, on the one hand, and either statistical or actuarial decision making on the other hand. Since this distinction has already been discussed at this conference, I will not define it further. However, I would like to suggest a change in terminology, from clinical decision making to subjective decision making. The term subjective is not only descriptively more accurate, but, in addition, it avoids the implication that nonactuarial decision making is something that is done only by clinicians.

The standard actuarial prediction procedure for the selection situation may be described briefly as follows. Given: (a) a criterion, which may be a composite of several variables (in this case measures of success as a medical student or as a physician); (b) a set of variables which on a priori grounds seem likely to be useful in predicting individuals' ranks on the criterion (e.g., MCAT scores, undergraduate GPA, and interview ratings); and (c) a set of individuals for whom scores are available on both the predictor variables and the criterion; then statistical models may be used to assign weights to each of the predictors so that the resulting function will provide the highest accuracy possible for that model. These weights, once obtained, may be used to predict the probable success of future applicants. Once the applicants have been ordered, selection begins with the individual with the highest predicted probability of success and continues until the desired number of entrants has been obtained. The most commonly-used model is the linear regression procedure. This procedure provides the best linear composite of the predictors in the least squares sense. Because of its simplicity, it is usually thought of as providing an estimate of the minimum level of predictive accuracy which one can expect in the situation. It therefore forms a kind of baseline against which to evaluate various alternatives which are thought to offer hope of greater validity.

The alternatives may be either actuarial or subjective, but in either case the aim is to identify and utilize valid, nonlinear relationships. The rationale is simply that many relationships may, in fact, be nonlinear, and, if this is the case, then identification of the actual nonlinear relationship would, of course, result in improved prediction. In the actuarial domain, most of the work on this problem has been in the form of a search for

moderator variables. If the predictive validity of one variable varies as a function of a second variable, then the second variable is termed a moderator variable (Saunders, 19), and the greatest predictive accuracy in that situation will be achieved by some nonlinear function of the two variables. Even among published studies, the proportion of cases in which nonlinear models have provided significant incremental accuracy relative to linear models on cross-validation is probably little greater than would be expected on a chance basis, given the level of statistical significance employed. In addition there must be many unsuccessful attempts which have never been reported.

The a priori rationale for the use of human decision makers is expressed in a number of different ways, all of which mean substantially the same thing. It has been said that humans are able to use information nonlinearly, or nonadditively, or configurally; or that they can recognize patterned relationships or other cues that are too rare to achieve significance in a regression function. Note that, given the outcome of the actuarial studies, the argument for the human judge involves, in addition to the assumption that nonlinear relationships exist, the additional assumption that such relationships can be recognized and utilized subjectively, even when it is not possible to do so actuarily. There is now a considerable literature in which actuarial and subjective predictions have been compared for a wide variety of situations. It is one of the most consistent literatures to be found in psychology. These studies have repeatedly shown actuarial procedures to be superior to subjective ones. Furthermore, they have shown this to be the case even when the human decision maker has been provided with the actuarial prediction as part of the information on the basis of which to formulate his, presumably better, prediction (Sawyer, 1966).

For most of us, this is a counterintuitive result; so, we question it. Even if studies show that others cannot, we still feel that we can take special circumstances into account in individual cases, and, hence, make better decisions than the computer. How are we to account for this discrepancy between a consistent set of experimental findings and a widely-held belief? Are our subjective impressions really as much in error as the experimental results suggest, or is there something "artifactual" about the findings or the way in which they have been presented? At Oregon Research Institute (ORI) we have carried out a series of studies which are relevant to this question.

In contrast to those studies which have focused on either the absolute or comparative validity of subjective decision making, the aim of the ORI judgment studies has been to understand the way in which judges use cues in probabilistic decision-making tasks, such as medical diagnoses or graduate school admissions. In other words, we have tried to find out what it is that the decision maker does, in contrast to finding out how well he does it, though, naturally, we keep track of that, too. In the usual experimental situation, we give the judge the data (e.g., patient records, student personnel files, etc.) on the basis of which the decision is to be made, and then we simply note the decisions he makes for each case. Given a series of such decisions, it is possible to draw inferences concerning the way in which the judge used the cues to reach his decisions. In some studies in which simulated data are used, we know how the decision ought to be made; in other cases, where we are studying real problems, we often do not.

The first report of this research was a now classic paper by Hoffman (1960). What Hoffman did, and what many others have done since, was to treat the cues as predictor variables and the decisions as criteria, and use multiple-

regression procedures to predict the latter from the former. He dubbed the resulting function a "paramorphic representation" of a judge's decision strategy, and suggested that the β weights from the regression equation could be used as a first approximation for estimating the relative importance of the different cues in accounting for a judge's responses. There are problems involved in interpreting β weights in terms of the importance a judge has attached to a cue, especially if the cues are not orthogonal, so Hoffman and others have attempted to develop estimates that can more reasonably be interpreted as reflecting the relative importance which the judge attached to one cue as opposed to another. However, these problems need not concern us here.

What is important for the moment is that the model provides an estimate of the linear predictability of a judge's responses. It therefore provides a base from which to start the search for models that will more accurately represent the judgment process. In other words, when treated in this way, the prediction of subjective decisions is but a special case of the general actuarial prediction procedure, and the rationale for trying other models is the same as in any other situation: If judges do, in fact, make decisions on the basis of nonlinear procedures, then a model which properly reflects this fact will result in a higher level of predictive accuracy than the linear one. Presumably, it will give us a better idea of what the judge is actually doing, as well.

All kinds of alternative models have been tried: models with logarithmic and power terms, models with multiplicative or "interactive" terms, models with sequential or branching algorithms, etc. The result is no different in this area than in prediction in general. Ten years of research, carried out

primarily at the University of Colorado, Carnegie Institute of Technology, Purdue, and ORI, can be summarized quite simply: the simple linear model predicts judges' decisions about as well as any alternative model that has been tried. There are some exceptions, but considering the intensity of the search, the range of areas that have been studied, and the variety of techniques that have been employed, the exceptions are nothing to get excited about (Goldberg, 1968). Furthermore, in most cases the simple linear model predicts a judge's decisions more accurately than the judge himself can predict them.

Are we, then, to conclude that humans really are linear information processors, after all? Not necessarily. There is another possibility, one that would be consistent with the failure to find significant incremental nonlinear variance in other areas as well. It is possible that significant departures from linearity exist, but that our techniques are not sufficiently precise to detect that fact. As long as one studies human judges, or other problems in the real world where the underlying relationship is unknown, there is no way to tell which of the alternative explanations is more nearly the correct one. In order to get any leverage on the problem, one must apply the linear and nonlinear models to situations where the underlying relationship is known to be nonlinear. If one then finds differences between the linear and nonlinear models, it would suggest that the previous studies do indicate that people are indeed linear information processors. If, on the other hand, one still finds no differences, then it would show that the models are unable to detect such relationships even when they are known to exist, and it would follow that one could not conclude from previous studies that people are linear information processors.

This, then, is what we did. In order to know that we were studying relationships that were indeed nonlinear, we turned to a decision maker we could trust--a computer. The computer was programmed to make decisions in the ways it has been hypothesized human judges make decisions; that is; it was given a series of decision procedures of various kinds (e.g., linear, nonlinear, sequential, configural, simple, complex, etc.) and it made decisions on the basis of these procedures.

In our simulations we limited ourselves to three cues. In the present context you might think of the first as an MCAT score, the second as undergraduate grade point average (GPA), and the third as some kind of interview rating. In addition, we limited the values that a cue could take to four. While that may seem artificial, if you are thinking of something like the MCAT, for example, a second thought may convince you that it is not. The four categories might be: extremely high, high, adequate, and inadequate. You may divide the scores into somewhat different categories, and you may give them different labels, but I suspect that most people using the test, in fact, use some such categorization, and that it does not involve many more than four categories.

A representative sample of the decision procedures used is given in Table 1. The first six illustrate the classes of functions that have been

---

Insert Table 1 about here

---

suggested as alternatives to the linear representation of the judgment process. While the use of terms like "curvilinear" is not mathematically precise, the intended meaning should be clear from the table, and it is convenient to have some term arbitrarily designated to refer to each of the classes of functions.

The last five strategies are ones that might plausibly be employed by an admissions committee. They were, in fact, written to represent procedures that members of an admissions committee said that they used. They may seem somewhat more reasonable to you if they are translated into everyday language.

The first, α, represents the point of view reflected in the old saying, "A chain is no stronger than its weakest link." It means that you want to avoid an applicant who shows any area of weakness, no matter how strong he may appear in other areas. The second, β, focuses on signs of strength; the more of them there are, the better. Under this strategy, a single low indicator would not necessarily disqualify someone. Under the third strategy, γ, an individual who was outstanding on something, and not really low on anything, would be admitted; contrariwise, if an individual was very low on some indicator and had no sign of strength to compensate for it, he would be excluded. Individuals who were intermediate on all indicators, or who had a mixture of high and low indicators, would be classified in a middle, or indeterminate, category.

Both δ and ε are sequential procedures. The underlying rationale for both is that predictors vary in their validity and the decision should be made on the basis of the most valid predictor that gives a clear signal. Thus, one considers the cues one at a time, in order of their presumed validity, and makes a decision on the basis of the first one that is significantly high or low. For example, if you consider the MCAT the most valid predictor, you would consider it first. If it was significantly high, you would admit; if it was significantly low, you would reject; and if it was intermediate, you would go on to the second predictor, GPA. Again, if GPA was significantly high, you would admit, and so on. If you run through all of your predictors

and still haven't made a decision, the strategies differ with regard to what you do then. The δ strategy makes a decision on the basis of the total of all the scores, whereas the ε strategy assigns the applicant to an indeterminate category, which is to say, it leaves him hanging in limbo until you know whether you want more students or not.

These five admissions strategies, all of which are plausible, represent widely differing points of view with regard to the basis on which admissions decisions should be made. For example, consider an applicant who had all very high indicators except one, which was very low. This individual would be rejected by α, admitted by β, and placed in a middle or indeterminate category by γ; his fate under δ and ε would depend on which variables were high and which were low. Thus, in the case of discrepant scores, these strategies would result in quite different decisions. Naturally, if an individual's scores are consistent, especially if they are consistently either very high or very low, it is not going to make much difference what strategy one pursues. The decisions of widely varying strategies will all be the same.

Having programmed the computer to act like 11 different decision makers, e.g., 11 different members of an admissions committee, we provided it with cases constructed by taking all possible combinations of the four levels of the three cues. In order that the computer would look more like an admissions committee, and less like a computer, we added error to its judgments. We did this systematically, so that we obtained judgments at different levels of reliability, ranging from 1.00, perfect reliability, to .50, which seemed like the minimum acceptable level for any procedure that anyone would bother working with. (Morse's ( ) report yesterday indicates that we are going to have to re-evaluate that assumption.)

After the computer had produced its judgments, we analyzed them exactly as though they has been produced by a human judge whose decision strategy we did not know. Now, the question was: If we had not known better, would we have concluded that these judges were behaving linearly? Or, to put it more precisely, would we have concluded that their judgments could have been linearly represented with adequate precision? Or, to be more pedantic still, would we have concluded that the incremental validities of the nonlinear models, relative to the linear one, were significant?

Three different kinds of analyses were run. First, we carried out analyses of variance, because these had been reported in a number of studies in the literature. The results for five errorless cases are presented in Tables 2 through 6. The values of ω^2 in the right-hand column of each table indicate

---

Insert Tables 2, 3, 4, 5, and 6 about here

---

the proportion of the variance attributable to that term or that line in the table. The results in Table 2 are for the simple linear strategy. In this case, the values of ω^2 are proportional to the squares of the weights of the predictors in the decision function. Now look at Table 3, which presents the results for the "totally configural" or "totally interactive" strategy, in which the judgments were actually made by multiplying the values of the cues together. Almost 80% of the variance was attributed to the main effects, the simple linear functions of x and y. In Table 4, note how little effect squaring the terms has. Only 1% of the variance was attributed to this departure from linearity. Table 5, on the other hand, provides an example of a case in which an interaction term did account for a substantial proportion of the variance relative to the main effects. Table 6 indicates that 84% of the variance in this seemingly complex strategy can be accounted for by the variation in x alone.

The analyses in Tables 2 through 6 are based on judgments of perfect reliability. Except as an artifact of sampling fluctuation, the interaction terms should be larger for this case than for those with less than perfect reliability. For errorless data of this kind, all nonzero terms are statistically significant. That, of course, does not mean that they are practically significant. For more realistic levels of error variance, most of the interaction terms presented here would no longer be statistically significant. In general, with the exception of the xy term in Table 5, it seems safe to say that a researcher presented with these analyses of variance would conclude that the interaction terms were not large enough to be considered of any practical importance.

In addition to the analyses of variance, two different multiple-regression analyses were run: first, the usual linear model, and, second, the general quadratic model (Hoffman, 1966). Terminology is difficult here, because in each case the functions are linear in the β weights, so that, in that sense, they are both examples of linear regression. However, in the second case, nonlinear and multiplicative functions of the original cues have been included in the model, as shown at the bottom of Table 7, which presents the multiple correlation coefficients (R) from these analyses. Results are given for the

---

Insert Table 7 about here

---

11 decision strategies, both for the case in which the decisions were made with perfect reliability ($r_{jj} = 1.00$) and for the case in which they were made with intermediate reliability ($r_{jj} = .60$).

The results for the perfect reliability case indicate the extent to which the underlying decision strategy is representable by either the linear or the

quadratic model. In other words, the results indicate the amount of precision that would be lost if one used the regression equation instead of the function itself. As you can see, in general, the loss is not great. With the exception of the sequential procedures, the simple linear function provides a very close approximation to the real function. The quadratic model, of course, can represent the configural or interactive judgment strategy perfectly, because it has the same product term as the decision function. But note that the linear function, which has no product term, reproduces decisions that were made multiplicatively with a correlation of .88. Relative to the question we set out to investigate, these results indicate vividly that it is difficult to infer nonlinearity, even when it exists. The approximation of the quadratic model to the real decision function, and of the linear model to the quadratic one, is so close that, in almost all situations, parsimony would require that we conclude that the departures from linearity were not practically significant, even if they should prove to be statistically so. Even statistically significant differences of this magnitude could not be considered real until they had been replicated, and it would require an unusual payoff matrix to justify the attempt at replication.

If you think about this a bit, it will certainly occur to you that there are strategies that are not linearly representable, and that there must be limits to the generality of these results. You are correct. But we have to distinguish between those strategies that are mathematically possible and those that are realistically plausible. The results that have been presented break down in the configural case if one cue reverses the significance of another. In the selection situation, for example, the results would not hold if it were the case that you looked for males with high MCAT scores and females

with low ones, or if you preferred high GPA for applicants from wealthy neighborhoods and low GPA for those from the ghetto. I do not mean simply that you would tolerate lower GPA in the latter case, but that for two applicants from the ghetto you would prefer the one with the lower GPA. While situations in which the value of one cue reverses that of another are theoretically possible, plausible examples are hard to conceptualize, and even harder to find. At ORI we have been actively looking for over 10 years, and the only good examples we have found so far involve stock market indicators. Heavy trading volume, for example, is supposed to be a good sign if it occurs in a rising market, but a bad sign if it occurs in a declining one (Slovic, 1969).

In sum, within the range of strategies that a judge is likely to be using, it is the case that the strategy can be reproduced with a high degree of accuracy by the simple linear regression function. Furthermore, since the simulation had to do with patterned or configural relationships in general, the results are not limited to subjective decision making, but apply to any prediction situation in which patterned or configural relationships are alleged to exist. These results show why the search for moderator variables has been so fruitless, just as much as they show why human judges have been unable to improve upon actuarial prediction by means of their purported ability to recognize and use patterned or unusual relationships. Even if such relationships exist, their incorporation into the decision strategy, be it actuarial or subjective, is not going to result in much improvement over the linear regression function in either case.

Consider now the case in which the judge is not entirely reliable, but rather makes his decisions with intermediate reliability ($r_{jj} = .60$). That is,

if he judged the same cases on two different occasions, the correlation between his decisions on those two occasions would be .60. Since the data in Table 7 are based on a single replication, they exhibit the same kind of variability you would expect to see in analyses based on samples of real data. The values to which the coefficients in the derivation column converge can be calculated by inserting the results from the perfect reliability columns into the equation for correction for attenuation due to unreliability. In other words, except for sampling fluctuation, the derivation values are at the limits that could be expected, given the unreliability of the judgments and the extent to which the linear or quadratic function can approximate the actual decision strategy.

Those of you who have done much empirical work will find the cross-validation results startling, for they do not exhibit the kind of shrinkage to which you are accustomed. The explanation would seem to be that here, in contrast to many empirical studies, there is an underlying nonrandom functional relationship. In addition to the cross-validation results, the table also gives what I have called cross-function coefficients. These are the correlations between the estimates generated by the regression equation and the values of the function itself. As you can see, these are approximately at the limit of the linear representability of the functions. In other words, not only are the functions linearly representable, but, furthermore, the linear regression procedure will produce almost as good a representation on the basis of unreliable data as it will on the basis of perfectly reliable data. If there is an underlying relationship, error, in the sense that error is usually defined, namely as random variation, makes little difference. These results indicate that researchers who find that their initially promising

derivation results fail to cross-validate are entitled to make much stronger statements than are customarily made in the research literature concerning the lack of any functional relationship among the variables. Those who persist in their investigations in the face of such evidence merit the fate they are bound to receive.

Finally, look again at the cross-validation column in Table 7. Note that the simulation has produced a seemingly paradoxical result. The reliability of the data on which these analyses are based is only .60; yet, all but two of the cross-validation coefficients for the linear function exceed this figure, and only one fails to do so for the quadratic function. In other words, the values predicted by a model derived on the basis of one set of judgments correlate more highly with the second set of judgments than does the original set on the basis of which the model was derived. That is, these results indicate that if a judge were to provide decisions for a set of cases on two occasions, and if one were to predict his judgments on the second occasion using (a) the original set of judgments and (b) a model derived on the basis of the original judgments, (b) would, in general, predict more accurately than (a). Furthermore, this is the case even though the models being used were, in fact, nonlinear. The model's superiority stems from its close approximation to the underlying function (even though it was derived on the basis of unreliable data) and its perfect reliability (in contrast to the unreliable data on the basis of which it was derived). In more anthropomorphic terms, the model is able to figure out what the judge is trying to do, and then to do it with perfect reliability. Applying the model to the judge's decision strategy has increased the validity by an amount equal to that which would be expected on the basis of the correction for attenuation due to unreliability.

As can be seen by comparing the cross-validity and the cross-function results, the lower accuracy of the predictions for the intermediate reliability case relative to the perfect reliability case stems from the unreliability of the judgments, not from greater departures from the functional relationship.

All of this suggests that, if a judge has some correct notion of what ought to be done, but is unable do do it with complete reliability, then a model derived on the basis of his judgments ought to have greater validity than the judgments themselves; this, indeed, turns out to be the case. In contrast to the usual situation in psychology, in which the theory is invented after the fact to account for the results that have been obtained, this is an example of a case in which the result was first predicted on the basis of the simulations and theoretical derivations which I have described, and then confirmed by experimental investigation. At least three examples have now appeared in the literature (Dawes, 1971; Goldberg, 1970; Wiggins & Kohen, 1971). In addition, at least two previous studies reporting similar results have subsequently been identified (Dudycha & Naylor, 1966; Yntema & Torgerson, 1961).

Goldberg applied the principle to psychiatric diagnoses made by judges in a study reported by Meehl (1959) and Meehl and Dahlstrom (1960). While the incremental validities which he found were small, because the validity of the original diagnoses was low, the results clearly demonstrated the predicted phenomenon. In addition, he presented equations which indicate, in a different way than I have, the limits to the generality of the simulation results presented here. Anyone considering a study whose success depends on judges identifying valid nonlinear variance in a predictive situation should study Goldberg's equations carefully before starting. Wiggins and Kohen

reported more dramatic results than Goldberg's for an experiment in which graduate student subjects predicted first year GPAs for a sample of their peers.

Dawes, who was at the time a member of an admissions committee for a graduate department, was able to produce significant increases in validity for the real judgments which were made by the members of his committee. His paper develops the rationale for the phenomenon, which he termed bootstrapping, in a straightforward manner from a different starting point than in this paper.

These papers are essential reading for anyone concerned with medical school admissions, diagnostic decision making, or any other form of probabilistic inference in which subjective procedures are traditionally used. In fact, it was in the hope that I would be able to arouse your curiosity sufficiently that you would read these papers that I chose to lead you on this circuitous route to bootstrapping.

Hopefully, at least some of the implications of these findings with respect to Dr. Haley's charge to the conference will be obvious. I would like to emphasize those that have to do with medical school admissions as a way of summarizing what I have said.

One of the underlying premises of this conference is that there may be personality measures that will contribute to the accuracy of the predictions that can be made of an individual's future performance. A number of potential predictors have been suggested by other participants. The first implication of the findings I have reported is that the evaluation of these suggested predictors can be carried out in a simple, straightforward way by trying them out in a multiple-regression analysis. Unfortunately, personality variables seldom result in appreciable incremental validity when incorporated into such

an analysis. Faced with disappointing results, researchers are wont to claim that their pet predictor would work if only more things were taken into account in a configural or interactive way. Examples of these kinds of interaction hypotheses have been presented at this conference by Snow ( ), and they are appealing to all of us because they seem to reflect a concern for the individual. But concern and good intentions are not enough. Even the published accounts of attempts to demonstrate such interactive relationships are overwhelmingly negative, and the additional number of unpublished ones must be substantial. Just to take one of Snow's examples, the most ambitious and comprehensive study that I know of that has been carried out to evaluate the interaction hypothesis with regard to matching students to teaching methods has been published in condensed form this year (Goldberg, 1972); but those of you who think that hope lies in this direction would do well to obtain a copy of the full report (Goldberg, 1969). I think the analyses reported here indicate that if the inclusion of a variable in a regression program does not result in significant incremental validity, that variable may confidently be discarded until evidence, not just speculation, to the contrary is provided.

Personality variables are intuitively appealing when one first considers them, because it seems obvious that individuals with different personalities are going to react differently in different situations. It is almost axiomatic that ability is not enough, but that one must have interest and motivation of the right kinds as well. But by the time an individual reaches the point of applying for medical school, there is available a host of performance data in which the effects of these personality, motivational, and interest variables are already reflected. To make plausible the notion that it is possible to improve on a prediction based on the individual's past performance (modified, perhaps, by some measure of ability and some consideration of the individual's

expressed interest and intentions), it is necessary to postulate the existence of personality variables which will, in the future, affect performance differently from the way in which they have affected performance in the past, or else to postulate the possibility of identifying personality changes which are going to take place in the future, or have taken place so recently that their effects on performance cannot yet be measured. Serious contemplation of the plausibility of these assumptions is not likely to increase your optimism with regard to the prospects for identifying heretofore undiscovered personality measures that will substantially improve the accuracy of the performance predictions that can be made on the basis of measures now available.

If hope for improvement does not lie in the identification of new predictors, then hope, if hope there be, must lie in making better use of those we have. If the implication of this research for the use of predictors were to be summed up as succinctly as possible, it would be that the greatest hope at the moment for improving the selection process lies in getting the human decision maker out of the decision process at the earliest possible moment. If the present admissions decisions have any validity at all, then regression equations derived on the basis of those decisions will almost certainly have more. In addition to improving the accuracy of prediction, this procedure has the added advantage of making explicit the variables on the basis of which the decision is being made. Since most of the judgment research indicates that people often have poor insight into what they are doing, this is an important contribution.

The charge is often made that an actuarial procedure of this kind tends to reify the selection process. The opposite is probably more nearly true. By making the policy explicit, it opens it to evaluation and to question,

thereby facilitating, rather than impeding, change. I have no doubt that an admissions policy can be changed by modifying the regression function on which it is based. I have no such faith that an admissions policy can be changed by modifying the attitudes or behavior of an admissions committee.

At the very least, each school ought to set up its own actuarial admissions program. A centralized operation based on the present coordinated admissions program would be even better. Presumably, the functions used to select student for different schools would not turn out to be the same. In other words, the differences in admissions policies would also become explicit. Applicants could then be counseled with regard to the schools to which they should apply, and the schools would know, for the first time, how their selection procedures compared to those of other schools.

Finally, if you are unwilling to let the regression equation make all of your decisions for you, you ought at least to be willing to let it do your initial screening. Dawes examined the results for the regression equation describing the actions of the committee of which he was a member, and found a score below which no applicant was ever admitted. He suggested that applicants scoring below this cutting point be eliminated immediately, without further consideration, since their elimination would not result in denying admission to anyone who might otherwise have been admitted. Using 10 minutes as the average amount of time that an admissions committee uses in discussing each case, and extrapolating his results to all graduate schools in the country, he estimated that the savings from this simple step alone would be $18,000,000 a year. I do not know what the comparable savings would be for medical schools, but they should be considerable, and it should not be difficult for someone familiar with the admissions programs to estimate them. The coordinated admissions program could be used to implement this procedure immediately, so

that no school would be sent an application from an applicant who had no chance of being admitted to that school. The savings from the implementation of this program could be used to finance the additional research needed to convince you that all of your decisions should be made in this way.

References

Dawes, R. M. A case study of graduate admissions: Application of three principles of human decision making. American Psychologist, 1971, 26, 180-188.

Dudycha, A. L., & Naylor, J. C. The effect of variations in the cue R-matrix upon the obtained policy equations of judges. Educational and Psychological Measurement, 1966, 26, 583-603.

Ghiselli, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. Educational and Psychological Measurement, 1960, 20, 675-684.

Goldberg, L. R. Simple models or simple processes: Some research on clinical judgments. American Psychologist, 1968, 23, 483-496.

Goldberg, L. R. Student personality characteristics and optimal college learning conditions. Oregon Research Institute Research Monograph, 1969, 9 (1).

Goldberg, L. R. Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. Psychological Bulletin, 1970, 73, 422-432.

Goldberg, L. R. Student personality characteristics and optimal college learning conditions: An extensive search for trait-by-treatment interaction effects. Instructional Science, 1972, in press.

Hoffman, P. J. The paramorphic representation of clinical judgment. Psychological Bulletin, 1960, 57, 116-131.

Hoffman, P. J. A general linear and quadratic regression analysis program. Behavioral Science, 1966, 11, 497.

Lykken, D. T. A method of actuarial pattern analysis. Psychological Bulletin, 1956, 53, 102-107.

Lykken, D. T., & Rose, R. Psychological prediction from actuarial tables. Journal of Clinical Psychology, 1963, 19, 139-151.

Meehl, P. E. A comparison of clinicians with five statistical methods of identifying psychotic MMPI profiles. Journal of Counseling Psychology, 1959, 6, 102-109.

Meehl, P. E., & Dahlstrom, W. G. Objective configural rules for discriminating psychotic from neurotic MMPI profiles. Journal of Consulting Psychology, 1960, 24, 375-387.

Saunders, D. R. Moderator variables in prediction. Educational and Psychological Measurement, 1956, 16, 209-222.

Sawyer, J. Measurement and prediction, clinical and statistical. Psychological Bulletin, 1966, 66, 178-200.

Slovic, P. Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. Journal of Applied Psychology, 1969, 53, 255-263.

Wiggins, N., & Kohen, E. S. Man versus model of man revisited: The forecasting of graduate school success. Journal of Personality and Social Psychology, 1971, 19, 100-106.

Untena, 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, Vol. HFE-2, No. 1, 20-26.

Footnotes

[1]This is a revised version of a paper presented at the "Conference on Personality Measurement in Medical Education," Des Plaines, Illinois, June 17-18, 1971.

[2]Paul J. Hoffman and Paul Slovic both participated in the design of this study, and they and Robyn M. Dawes consulted with regard to the interpretation of the results. Amos Tversky, Robyn M. Dawes, Lewis R. Goldberg, Richard R. Jones, Paul Hoffman, and Paul Slovic read an earlier version of this paper and all contributed suggestions that resulted in substantial improvements. This study was supported in part by Grant No. MH12972 and in part by Grant No. MH10822 from the National Institute of Mental Health, U.S. Public Health Service. Computing assistance was obtained from the Health Sciences Computing Facility, UCLA, sponsored by NIH Special Research Resources Grant RR-3.

[3]The present version of this paper was prepared while the author was a Fellow at the Netherlands Institute for Advanced Study, Wassenaar, The Netherlands.

Table I

Hypothetical Decision Strategies Which Were Simulated

Applicant was assigned a rating (j) based on three variables (x, y, and z,) which may be taken as x = score on the MCAT, y = undergraduate GPA, and z = an interview rating, for example. Scores were grouped into four categories (0, 1, 2, 3), for each variable.

Six Cases Representing Different Kinds of Functional Relationships

(a) Linear, but the cues are weighted differently, i.e., they are not treated as equally important.

$$j = x + 2y + 3z$$

(b) "Entirely curvilinear" use of 2 cues.

$$j = x^2 + y^2$$

(c) "Entirely configural" and uses only 2 cues. The significance of one cue depends on the value of the other, and vice versa.

$$j = xy$$

(d) Both curvilinear and configural use of 2 cues; no linear cue usage.

$$j = (x + y)^2 = x^2 + 2xy + y^2$$

(e) Sequential. The first cue determines which of the other two cues to use for the decision. Ghiselli's differential predictability variable (1960) for deciding which test to use exemplifies this strategy, which is also "entirely configural."

If x + 0, 1, then $j = y^2$; if x = 2, 3, then j = z.

(f) Categorical. This can be considered as either a complex sequential strategy, or as an "actuarial table" technique in which each cell is assigned a criterion value, as suggested, for example, by Lykken (1956) and Lykken and Rose (1963).

x	y	z	j
2,3	0,1	2,3	0
2,3	0,1,2,3	0,1	1
2,3	2,3	2,3	2
0,1	0,1,2,3	2,3	4
0,1	3	0,1	5
0,1	0,1,2	0,1	6

Continued on next page

Five Cases Representing Plausible Admissions Policies

α j = the lowest score

β j = the number of variables which = 2 or 3

γ j = 2 if any variable = 3 and none = 0
1 otherwise
0 if any variable = 0 and none = 3

δ j = 1 if x = 3, 0 if x = 0, y if x = 1 or 2;
1 if y = 3, 0 if y = 0, z if y = 1 or 2;
1 if z = 3, 0 if z = 0, Σxyz if z = 1 or 2;
1 if Σxyz = 5 or 6; 0 if Σxyz = 3 or 4.

ε j = 2 if x = 3, 0 if x = 0, y if x = 1 or 2;
2 if y = 3, 0 if y = 0, z if y = 1 or 2;
2 if z = 3, 0 if z = 0, 1 if z = 1 or 2.

Table 2

ANOVA for: $J = X + 2Y + 3Z$

Source	df	SS	MS	ω^2
X	3	80	27	.07
Y	3	320	107	.29
Z	3	720	240	.64
XY	9	0	0	
XZ	9	0	0	
YZ	9	0	0	
XYZ	27	0	0	
Total	63	1120		1.00

Table 3

ANOVA for: J = XY

Source	df	SS	MS	ω^2
X	3	180	60	.39
Y	3	180	60	.39
Z	3	0	0	
XY	9	100	11	.22
XZ	9	0	0	
YZ	9	0	0	
XYZ	27	0		
Total	63	460		1.00

Table 4

ANOVA for: $J = (X+Y)^2 = X^2 + 2XY + Y^2$

Source		df	SS	MS		ω^2
X		3	2944	981		.47
lin	1		2880		.46	
quad	1		64		.01	
cub	1		0			
Y		3	2944	981		.47
lin	1		2880		.46	
quad	1		64		.01	
cub	1		0			
Z		3	0	0		
XY		9	400	44		.06
lin x lin	1		400		.06	
residual	8		0		.00	
XZ		9	0			
YZ		9	0			
XYZ		27	0			
Total		63	6288			1.00

Table 5

ANOVA for Sequential Strategy:

If X = 0,1, then J = Y^2; df X = 2, 3, then J = Z

Source		df		SS	MS		ω^2
X		3		64	21		.13
lin	I		51			.10	
quad	I		0			.00	
cub	I		13			.03	
Y		3		196	65		.39
lin	I		180			.36	
quad	I		16			.03	
cub	I		0			.00	
Z		3		20	7		.04
lin	I		20			.04	
quad	I		0			.00	
cub	I		0			.00	
XY		9		196	22		.39
lin x lin	I		144			.29	
lin x quad	I		13			.03	
cub x lin	I		36			.07	
cub x quad	I		3			.01	
XZ		9		20	2		.04
lin x lin	I		16			.03	
cub x lin	I		4			.01	
YZ		9		0	0		
XYZ		27		0	0		
Total		63		496			.99

Table 6

ANOVA for:						
X	2,3	2,3	2,3	0,1	0,1	0,1
Y	0,1	0,1,2,3	2,3	0,1,2,3	3	0,1,2
Z	2,3	0,1	2,3	2,3	0,1	0,1
J	0	1	2	4	5	6

Source		df		SS	MS		ω^2
X		3		240	80		.84
lin	1		192			.68	
quad	1		0			.00	
cub	1		48			.17	
Y		3		3	1		.01
lin	1		1			.00	
quad	1		0			.00	
cub	1		1			.00	
Z		3		12	4		.04
lin	1		10			.04	
quad	1		0			.00	
cub	1		2			.01	
XY		9		7	1		.02
lin x lin	1		5			.02	
cub x lin	1		1			.00	
residual	7		1			.00	
XZ		9		12	1		.04
lin x lin	1		8			.03	
lin x cub	1		2			.01	
cub x lin	1		2			.01	
YZ		9		7	1		.02
lin x lin	1		5			.02	
lin x cub	1		1			.00	
residual	7		1			.00	
XYZ		27		3			.01
Total		63		284			.98

Table 7

Derivation, Cross Validation, and Cross Function Validity Coefficients for Linear and Quadratic Multiple-Regression Analyses

	r_{jj} = .60						r_{jj} = 1.00		
	1 Rep								
	Derivation		Cross-Validation		Cross-Function				
Function	L	Q	L	Q	L	Q	L	Q	
a	84	85	76	75	99	98	1.00	1.00	$x + 2y + 3z$
b	71	77	74	75	94	97	96	1.00	$x^2 + y^2$
c	70	85	68	76	88	97	88	1.00	xy
d	77	85	72	74	94	97	96	1.00	$(x + y)^2$
e	51	80	56	66	85	85	71	93	Sequential
f	61	69	65	65	83	83	85	88	Table
α	61	76	58	73	74	93	75	94	α
β	74	75	66	66	86	86	89	89	β
δ	65	69	67	63	85	80	85	85	δ
γ	71	74	62	59	79	77	81	81	γ
ε	75	76	64	63	81	80	82	82	ε

L: $\hat{j} = b_0 + b_1x + b_2y + b_3z$

Q: $\hat{j} = b_0 + b_1x + b_2y + b_3z + b_4x^2 + b_5y^2 + b_6z^2 + b_7xy + b_8xz + b_9yz$