The Robust Beauty of Improper Linear Models in Decision Making

ROBYN M. DAWES *University of Oregon*

ABSTRACT: *Proper linear models are those in which predictor variables are given weights in such a way that the resulting linear composite optimally predicts some criterion of interest; examples of proper linear models are standard regression analysis, discriminant function analysis, and ridge regression analysis. Research summarized in Paul Meehl's book on clinical versus statistical prediction—and a plethora of research stimulated in part by that book—all indicates that when a numerical criterion variable (e.g., graduate grade point average) is to be predicted from numerical predictor variables, proper linear models outperform clinical intuition. Improper linear models are those in which the weights of the predictor variables are obtained by some nonoptimal method; for example, they may be obtained on the basis of intuition, derived from simulating a clinical judge's predictions, or set to be equal. This article presents evidence that even such improper linear models are superior to clinical intuition when predicting a numerical criterion from numerical predictors. In fact, unit (i.e., equal) weighting is quite robust for making such predictions. The article discusses, in some detail, the application of unit weights to decide what bullet the Denver Police Department should use. Finally, the article considers commonly raised technical, psychological, and ethical resistances to using linear models to make important social decisions and presents arguments that could weaken these resistances.*

Paul Meehl's (1954) book *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* appeared 25 years ago. It reviewed studies indicating that the prediction of numerical criterion variables of psychological interest (e.g., faculty ratings of graduate students who had just obtained a PhD) from numerical predictor variables (e.g., scores on the Graduate Record Examination, grade point averages, ratings of letters of recommendation) is better done by a proper linear model than by the clinical intuition of people presumably skilled in such prediction. The point of this article is to review evidence that even improper linear models may be superior to clinical predictions.

A *proper linear model* is one in which the weights given to the predictor variables are chosen in such a way as to optimize the relationship between the prediction and the criterion. Simple regression analysis is the most common example of a proper linear model; the predictor variables are weighted in such a way as to maximize the correlation between the subsequent weighted composite and the actual criterion. Discriminant function analysis is another example of a proper linear model; weights are given to the predictor variables in such a way that the resulting linear composites maximize the discrepancy between two or more groups. Ridge regression analysis, another example (Darlington, 1978; Marquardt & Snee, 1975), attempts to assign weights in such a way that the linear composites correlate maximally with the criterion of interest in a new set of data.

Thus, there are many types of proper linear models and they have been used in a variety of contexts. One example (Dawes, 1971) was presented in this Journal; it involved the prediction of faculty ratings of graduate students. All gradu-

Work on this article was started at the University of Oregon and Decision Research, Inc., Eugene, Oregon; it was completed while I was a James McKeen Cattell Sabbatical Fellow at the Psychology Department at the University of Michigan and at the Research Center for Group Dynamics at the Institute for Social Research there. I thank all these institutions for their assistance, and I especially thank my friends at them who helped.

This article is based in part on invited talks given at the American Psychological Association (August 1977), the University of Washington (February 1978), the Aachen Technological Institute (June 1978), the University of Groeningen (June 1978), the University of Amsterdam (June 1978), the Institute for Social Research at the University of Michigan (September 1978), Miami University, Oxford, Ohio (November 1978), and the University of Chicago School of Business (January 1979). I received valuable feedback from most of the audiences.

Requests for reprints should be sent to Robyn M. Dawes, Department of Psychology, University of Oregon, Eugene, Oregon 97403.

ate students at the University of Oregon's Psychology Department who had been admitted between the fall of 1964 and the fall of 1967—and who had not dropped out of the program for nonacademic reasons (e.g., psychosis or marriage)—were rated by the faculty in the spring of 1969; faculty members rated only students whom they felt comfortable rating. The following rating scale was used: 5, outstanding; 4, above average; 3, average; 2, below average; 1, dropped out of the program in academic difficulty. Such overall ratings constitute a psychologically interesting criterion because the subjective impressions of faculty members are the main determinants of the job (if any) a student obtains after leaving graduate school. A total of 111 students were in the sample; the number of faculty members rating each of these students ranged from 1 to 20, with the mean number being 5.67 and the median being 5. The ratings were reliable. (To determine the reliability, the ratings were subjected to a one-way analysis of variance in which each student being rated was regarded as a treatment. The resulting between-treatments variance ratio (η^2) was .67, and it was significant beyond the .001 level.) These faculty ratings were predicted from a proper linear model based on the student's Graduate Record Examination (GRE) score, the student's undergraduate grade point average (GPA), and a measure of the selectivity of the student's undergraduate institution.[1] The cross-validated multiple correlation between the faculty ratings and predictor variables was .38. Congruent with Meehl's results, the correlation of these latter faculty ratings with the average rating of the people on the admissions committee who selected the students was .19;[2] that is, it accounted for one fourth as much variance. This example is typical of those found in psychological research in this area in that (a) the correlation with the model's predictions is higher than the correlation with clinical prediction, but (b) both correlations are low. These characteristics often lead psychologists to interpret the findings as meaning that while the low correlation of the model indicates that linear modeling is deficient as a method, the even lower correlation of the judges indicates only that the wrong judges were used.

An *improper linear model* is one in which the weights are chosen by some nonoptimal method. They may be chosen to be equal, they may be chosen on the basis of the intuition of the person making the prediction, or they may be chosen at random. Nevertheless, improper models may have great utility. When, for example, the standardized GREs, GPAs, and selectivity indices in the previous example were weighted equally, the resulting linear composite correlated .48 with later faculty rating. Not only is the correlation of this linear composite higher than that with the clinical judgment of the admissions committee (.19), it is also higher than that obtained upon cross-validating the weights obtained from half the sample.

An example of an improper model that might be of somewhat more interest—at least to the general public—was motivated by a physician who was on a panel with me concerning predictive systems. Afterward, at the bar with his wife and me, he said that my paper might be of some interest to my colleagues, but success in graduate school in psychology was not of much general interest: "Could you, for example, use one of your improper linear models to predict how well my wife and I get along together?" he asked. I realized that I could—or might. At that time, the Psychology Department at the University of Oregon was engaged in sex research, most of which was behavioristically oriented. So the subjects of this research monitored when they made love, when they had fights, when they had social engagements (e.g., with in-laws), and so on. These subjects also made subjective ratings about how happy they were in their marital or coupled situation. I immediately thought of an improper linear model to predict self-ratings of marital happiness: rate of lovemaking minus rate of fighting. My colleague John Howard had collected just such data on couples when he was an undergraduate at the University of Missouri—Kansas City, where he worked with Alexander (1971). After establishing the intercouple reliability of judgments of lovemaking and fighting, Alexander had one partner from each of 42 couples monitor these events. She allowed us to analyze her data, with the following results: "In the thirty happily married

[1] This index was based on Cass and Birnbaum's (1968) rating of selectivity given at the end of their book *Comparative Guide to American Colleges*. The verbal categories of selectivity were given numerical values according to the following rule: most selective, 6; highly selective, 5; very selective (+), 4; very selective, 3; selective, 2; not mentioned, 1.

[2] Unfortunately, only 23 of the 111 students could be used in this comparison because the rating scale the admissions committee used changed slightly from year to year.

couples (as reported by the monitoring partner) only two argued more often than they had intercourse. All twelve of the unhappily married couples argued more often" (Howard & Dawes, 1976, p. 478). We then replicated this finding at the University of Oregon, where 27 monitors rated happiness on a 7-point scale, from "very unhappy" to "very happy," with a neutral midpoint. The correlation of rate of lovemaking minus rate of arguments with these ratings of marital happiness was .40 ($p < .05$); neither variable alone was significant. The findings were replicated in Missouri by D. D. Edwards and Edwards (1977) and in Texas by Thornton (1977), who found a correlation of .81 ($p < .01$) between the sex–argument difference and self-rating of marital happiness among 28 new couples. (The reason for this much higher correlation might be that Thornton obtained the ratings of marital happiness after, rather than before, the subjects monitored their lovemaking and fighting; in fact, one subject decided to get a divorce after realizing that she was fighting more than loving; Thornton, Note 1.) The conclusion is that if we love more than we hate, we are happy; if we hate more than we love, we are miserable. This conclusion is not very profound, psychologically or statistically. The point is that this very crude improper linear model predicts a very important variable: judgments about marital happiness.

The bulk (in fact, all) of the literature since the publication of Meehl's (1954) book supports his generalization about proper models versus intuitive clinical judgment. Sawyer (1966) reviewed a plethora of these studies, and some of these studies were quite extensive (cf. Goldberg, 1965). Some 10 years after his book was published, Meehl (1965) was able to conclude, however, that there was only a single example showing clinical judgment to be superior, and this conclusion was immediately disputed by Goldberg (1968) on the grounds that even the one example did not show such superiority. Holt (1970) criticized details of several studies, and he even suggested that prediction as opposed to understanding may not be a very important part of clinical judgment. But a search of the literature fails to reveal any studies in which clinical judgment has been shown to be superior to statistical prediction when both are based on the same codable input variables. And though most nonpositivists would agree that understanding is. not synonymous with prediction, few would agree that it doesn't entail some ability to predict.

Why? Because people—especially the experts in a field—are much better at selecting and coding information than they are at integrating it.

But people *are* important. The statistical model may integrate the information in an optimal manner, but it is always the individual (judge, clinician, subjects) who chooses variables. Moreover, it is the human judge who knows the directional relationship between the predictor variables and the criterion of interest, or who can code the variables in such a way that they have clear directional relationships. And it is in precisely the situation where the predictor variables are good and where they have a conditionally monotone relationship with the criterion that proper linear models work well.[8]

The linear model cannot replace the expert in deciding such things as "what to look for," but it is precisely this knowledge of what to look for in reaching the decision that is the special expertise people have. Even in as complicated a judgment as making a chess move, it is the ability to code the board in an appropriate way to "see" the proper moves that distinguishes the grand master from the expert from the novice (deGroot, 1965; Simon & Chase, 1973). It is not in the ability to integrate information that people excel (Slovic, Note 2). Again, the chess grand master considers no more moves than does the expert; he just knows which ones to look at. The distinction between knowing what to look for and the ability to integrate information is perhaps best illustrated in a study by Einhorn (1972). Expert doctors coded biopsies of patients with Hodgkin's disease and then made an overall rating of the severity of the process. The overall rating did not predict survival time of the 193 patients, all of whom

[8] Relationships are conditionally monotone when variables can be scaled in such a way that higher values on each predict higher values on the criterion. This condition is the combination of two more fundamental measurement conditions: (a) independence (the relationship between each variable and the criterion is independent of the values on the remaining variables) and (b) monotonicity (the ordinal relationship is one that is monotone). (See Krantz, 1972; Krantz, Luce, Suppes, & Tversky, 1971). The true relationships need not be linear for linear models to work; they must merely be approximated by linear models. It is not true that "in order to compute a correlation coefficient between two variables the relationship between them must be linear" (advice found in one introductory statistics text). In the first place, it is always possible to compute something.

died. (The correlations of rating with survival time were all virtually 0, some in the wrong direction.) The variables that the doctors coded did, however, predict survival time when they were used in a multiple regression model.

In summary, proper linear models work for a very simple reason. People are good at picking out the right predictor variables and at coding them in such a way that they have a conditionally monotone relationship with the criterion. People are bad at integrating information from diverse and incomparable sources. Proper linear models are good at such integration when the predictions have a conditionally monotone relationship to the criterion.

Consider, for example, the problem of comparing one graduate applicant with GRE scores of 750 and an undergraduate GPA of 3.3 with another with GRE scores of 680 and an undergraduate GPA of 3.7. Most judges would agree that these indicators of aptitude and previous accomplishment should be combined in some compensatory fashion, but the question is how to compensate. Many judges attempting this feat have little knowledge of the distributional characteristics of GREs and GPAs, and most have no knowledge of studies indicating their validity as predictors of graduate success. Moreover, these numbers are inherently incomparable without such knowledge, GREs running from 500 to 800 for viable applicants, and GPAs from 3.0 to 4.0. Is it any wonder that a statistical weighting scheme does better than a human judge in these circumstances?

Suppose now that it is not possible to construct a proper linear model in some situation. One reason we may not be able to do so is that our sample size is inadequate. In multiple regression, for example, *b* weights are notoriously unstable; the ratio of observations to predictors should be as high as 15 or 20 to 1 before *b* weights, which are the optimal weights, do better on cross-validation than do simple unit weights. Schmidt (1971), Goldberg (1972), and Claudy (1972) have demonstrated this need empirically through computer simulation, and Einhorn and Hogarth (1975) and Srinivisan (Note 3) have attacked the problem analytically. The general solution depends on a number of parameters such as the multiple correlation in the population and the covariance pattern between predictor variables. But the applied implication is clear. Standard regression analysis cannot be used in situations where there is not a "decent" ratio of observations to predictors.

Another situation in which proper linear models cannot be used is that in which there are no measurable criterion variables. We might, nevertheless, have some idea about what the important predictor variables would be and the direction they would bear to the criterion *if* we were able to measure the criterion. For example, when deciding which students to admit to graduate school, we would like to predict some future long-term variable that might be termed "professional self-actualization." We have some idea what we mean by this concept, but no good, precise definition as yet. (Even if we had one, it would be impossible to conduct the study using records from current students, because that variable could not be assessed until at least 20 years after the students had completed their doctoral work.) We do, however, know that in all probability this criterion is positively related to intelligence, to past accomplishments, and to ability to snow one's colleagues. In our applicant's files, GRE scores assess the first variable; undergraduate GPA, the second; and letters of recommendation, the third. Might we not, then, wish to form some sort of linear combination of these variables in order to assess our applicants' potentials? Given that we cannot perform a standard regression analysis, is there nothing to do other than fall back on unaided intuitive integration of these variables when we assess our applicants?

One possible way of building an improper linear model is through the use of *bootstrapping* (Dawes & Corrigan, 1974; Goldberg, 1970). The process is to build a proper linear model of an expert's judgments about an outcome criterion and then to use that linear model in place of the judge. That such linear models can be accurate in predicting experts' judgments has been pointed out in the psychological literature by Hammond (1955) and Hoffman (1960). (This work was anticipated by 32 years by the late Henry Wallace, Vice-President under Roosevelt, in a 1923 agricultural article suggesting the use of linear models to analyze "what is on the corn judge's mind.") In his influential article, Hoffman termed the use of linear models a *paramorphic* representation of judges, by which he meant that the judges' psychological processes did not involve computing an implicit or explicit weighted average of input variables, but that it could be simulated by such a weighting. Paramorphic representations have been extremely successful (for reviews see Dawes & Corrigan, 1974; Slovic & Lichtenstein, 1971) in contexts in which predictor variables have conditionally monotone relationships to criterion variables.

The bootstrapping models make use of the weights derived from the judges; because these weights are not derived from the relationship between the predictor and criterion variables themselves, the resulting linear models are improper. Yet these paramorphic representations consistently do better than the judges from which they were derived (at least when the evaluation of goodness is in terms of the correlation between predicted and actual values).

Bootstrapping has turned out to be pervasive. For example, in a study conducted by Wiggins and Kohen (1971), psychology graduate students at the University of Illinois were presented with 10 background, aptitude, and personality measures describing other (real) Illinois graduate students in psychology and were asked to predict these students' first-year graduate GPAs. Linear models of every one of the University of Illinois judges did a better job than did the judges themselves in predicting actual grade point averages. This result was replicated in a study conducted in conjunction with Wiggins, Gregory, and Diller (cited in Dawes & Corrigan, 1974). Goldberg (1970) demonstrated it for 26 of 29 clinical psychology judges predicting psychiatric diagnosis of neurosis or psychosis from Minnesota Multiphasic Personality Inventory (MMPI) profiles, and Dawes (1971) found it in the evaluation of graduate applicants at the University of Oregon. The one published exception to the success of bootstrapping of which I am aware was a study conducted by Libby (1976). He asked 16 loan officers from relatively small banks (located in Champaign-Urbana, Illinois, with assets between $3 million and $56 million) and 27 loan officers from large banks (located in Philadelphia, with assets between $.6 billion and $4.4 billion) to judge which 30 of 60 firms would go bankrupt within three years after their financial statements. The loan officers requested five financial ratios on which to base their judgments (e.g., the ratio of present assets to total assets). On the average, the loan officers correctly categorized 44.4 businesses (74%) as either solvent or future bankruptcies, but on the average, the paramorphic representations of the loan officers could correctly classify only 43.3 (72%). This difference turned out to be statistically significant, and Libby concluded that he had an example of a situation where bootstrapping did not work—perhaps because his judges were highly skilled experts attempting to predict a highly reliable criterion. Goldberg (1976), however, noted that many of the ratios had highly skewed distributions, and he reanalyzed Libby's data, normalizing the ratios before building models of the loan officers. Libby found 77% of his officers to be superior to their paramorphic representations, but Goldberg, using his rescaled predictor variables, found the opposite; 72% of the models were superior to the judges from whom they were derived.[4]

Why does bootstrapping work? Bowman (1963), Goldberg (1970), and Dawes (1971) all maintained that its success arises from the fact that a linear model distills underlying policy (in the implicit weights) from otherwise variable behavior (e.g., judgments affected by context effects or extraneous variables).

Belief in the efficacy of bootstrapping was based on the comparison of the validity of the linear model of the judge with the validity of his or her judgments themselves. This is only one of two logically possible comparisons. The other is the validity of the linear model of the judge versus the validity of linear models in general; that is, to demonstrate that bootstrapping works because the linear model catches the essence of the judge's valid expertise while eliminating unreliability, it is necessary to demonstrate that the weights obtained from an analysis of the judge's behavior are superior to those that might be obtained in other ways, for example, randomly. Because both the model of the judge and the model obtained randomly are perfectly reliable, a comparison of the random model with the judge's model permits an evaluation of the judge's underlying linear representation, or *policy*. If the random model does equally well, the judge would not be "following valid principles but following them poorly" (Dawes, 1971, p. 182), at least not principles any more valid than any others that weight variables in the appropriate direction:

Table 1 presents five studies summarized by Dawes and Corrigan (1974) in which validities

[4] It should be pointed out that a proper linear model does better than either loan officers or their paramorphic representations. Using the same task, Beaver (1966) and Deacon (1972) found that linear models predicted with about 78% accuracy on cross-validation. But I can't resist pointing out that the simplest possible improper model of them all does best. The ratio of assets to liabilities (!) correctly categorizes 48 (80%) of the cases studied by Libby.

TABLE 1

Correlations Between Predictions and Criterion Values

Example	Average validity of judge	Average validity of judge model	Average validity of random model	Validity of equal weighting model	Cross-validity of regression analysis	Validity of optimal linear model
Prediction of neurosis vs. psychosis	.28	.31	.30	.34	.46	.46
Illinois students' predictions of GPA	.33	.50	.51	.60	.57	.69
Oregon students' predictions of GPA	.37	.43	.51	.60	.57	.69
Prediction of later faculty ratings at Oregon	.19	.25	.39	.48	.38	.54
Yntema & Torgerson's (1961) experiment	.84	.89	.84	.97	—	.97

Note. GPA = grade point average.

(i.e., correlations) obtained by various methods were compared. In the first study, a pool of 861 psychiatric patients took the MMPI in various hospitals; they were later categorized as neurotic or psychotic on the basis of more extensive information. The MMPI profiles consist of 11 scores, each of which represents the degree to which the respondent answers questions in a manner similar to patients suffering from a well-defined form of psychopathology. A set of 11 scores is thus associated with each patient, and the problem is to predict whether a later diagnosis will be psychosis (coded 1) or neurosis (coded 0). Twenty-nine clinical psychologists "of varying experience and training" (Goldberg, 1970, p. 425) were asked to make this prediction on an 11-step forced-normal distribution. The second two studies concerned 90 first-year graduate students in the Psychology Department of the University of Illinois who were evaluated on 10 variables that are predictive of academic success. These variables included aptitude test scores, college GPA, various peer ratings (e.g., extraversion), and various self-ratings (e.g., conscientiousness). A first-year GPA was computed for all these students. The problem was to predict the GPA from the 10 variables. In the second study this prediction was made by 80 (other) graduate students at the University of Illinois (Wiggins & Kohen, 1971), and in the third study this prediction was made by 41 graduate students at the University of Oregon. The details of the fourth study have already been covered; it is the one concerned with the prediction of later faculty ratings at Oregon. The final study (Yntema & Torgerson, 1961) was one in which experimenters assigned values to ellipses presented to the subjects, on the basis of figures' size, eccentricity, and grayness. The formula used was $ij + kj + ik$, where i, j, and k refer to values on the three dimensions just mentioned. Subjects in this experiment were asked to estimate the value of each ellipse and were presented with outcome feedback at the end of each trial. The problem was to predict the true (i.e., experimenter-assigned) value of each ellipse on the basis of its size, eccentricity, and grayness. .

The first column of Table 1 presents the average validity of the judges in these studies, and the second presents the average validity of the paramorphic model of these judges. In all cases, bootstrapping worked. But then what Corrigan and I constructed were ***random linear models,*** that is, models in which weights were randomly chosen except for sign and were then applied to standardized variables.[5]

> The sign of each variable was determined on an a priori basis so that it would have a positive relationship to the criterion. Then a normal deviate was selected at random from a normal distribution with unit variance, and the absolute value of this deviate was used as a weight for the variable. Ten thousand such models were constructed for each example. (Dawes & Corrigan, 1974, p. 102)

On the average, these random linear models perform about as well as the paramorphic models of the judges; these averages are presented in the third column of the table. Equal-weighting models, presented in the fourth column, do even better. (There is a mathematical reason why equal-weighting models must outperform the average random model.[6]) Finally, the last two columns present

[5] Unfortunately, Dawes and Corrigan did not spell out in detail that these variables must first be standardized and that the result is a standardized dependent variable. Equal or random weighting of incomparable variables—for example, GRE score and GPA—without prior standardization would be nonsensical.

[6] Consider a set of standardized variables S_1, X_2, $.X_m$, each of which is positively correlated with a standardized variable Y. The correlation of the average of the Xs with the Y is equal to the correlation of the sum of the

the cross-validated validity of the standard regression model and the validity of the optimal linear model.

Essentially the same results were obtained when the weights were selected from a rectangular distribution. Why? Because linear models are robust over deviations from optimal weighting. In other words, the bootstrapping finding, at least in these studies, has simply been a reaffirmation of the earlier finding that proper linear models are superior to human judgments—the weights derived from the judges' behavior being sufficiently close to the optimal weights that the outputs of the models are highly similar. The solution to the problem of obtaining optimal weights is one that —in terms of von Winterfeldt and Edwards (Note 4)—has a "flat maximum." Weights that are near to optimal level produce almost the same output as do optimal beta weights. Because the expert judge knows at least something about the direction of the variables, his or her judgments yield weights that are nearly optimal (but note that in all cases equal weighting is superior to models based on judges' behavior).

The fact that different linear composites correlate highly with each other was first pointed out 40 years ago by Wilks (1938). He considered only situations in which there was positive correlation between predictors. This result seems to hold generally as long as these intercorrelations are not negative; for example, the correlation between $X + 2Y$ and $2X + Y$ is .80 when X and Y are uncorrelated. The ways in which outputs are relatively insensitive to changes in coefficients (provided changes in sign are not involved) have been investigated most recently by Green (1977), Wainer (1976), Wainer and Thissen (1976), W. M. Edwards (1978), and Gardiner and Edwards (1975).

Dawes and Corrigan (1974, p. 105) concluded that "the whole trick is to know what variables to look at and then know how to add." That principle is well illustrated in the following study, conducted since the Dawes and Corrigan article was published. In it, Hammond and Adelman (1976) both investigated and influenced the decision about what type of bullet should be used by the Denver City Police, a decision having much more obvious social impact than most of those discussed above. To quote Hammond and Adelman (1976):

> In 1974, the Denver Police Department (DPD), as well as other police departments throughout the country, decided to change its handgun ammunition. The principle reason offered by the police was that the conventional round-nosed bullet provided insufficient "stopping effectiveness" (that is, the ability to incapacitate and thus to prevent the person shot from firing back at a police officer or others). The DPD chief recommended (as did other police chiefs) the conventional bullet be replaced by a hollow-point bullet. Such bullets, it was contended, flattened on impact, thus decreasing penetration, increasing stopping effectiveness, and decreasing ricochet potential. The suggested change was challenged by the American Civil Liberties Union, minority groups, and others. Opponents of the change claimed that the new bullets were nothing more than outlawed "dum-dum" bullets, that they created far more injury than the round-nosed bullet, and should, therefore, be barred from use. As is customary, judgments on this matter were formed privately and then defended publicly with enthusiasm and tenacity, and the usual public hearings were held. Both sides turned to ballistics experts for scientific information and support. (p. 392)

The disputants focused on evaluating the merits of specific bullets—confounding the physical effect of the bullets with the implications for social policy; that is, rather than separating questions of what it is the bullet should accomplish (the social policy question) from questions concerning ballistic characteristics of specific bullets, advocates merely argued for one bullet or another. Thus, as Hammond and Adelman pointed out, social policymakers inadvertently adopted the role of (poor) ballistics experts, and vice versa. What Hammond and Adelman did was to discover the important policy dimensions from the policymakers, and then they had the ballistics experts rate the bullets with respect to these dimensions. These dimensions turned out to be stopping effectiveness (the probability that someone hit in the torso could not return fire), probability of serious injury, and probability of harm to by-

Xs with Y. The covariance of this sum with Y is equal to

$$\left(\frac{1}{n}\right) \sum_i y_i(x_{i1} + x_{i2} \ldots + x_{im})$$
$$= \left(\frac{1}{n}\right) \sum y_i x_{i1} + \left(\frac{1}{n}\right) \sum y_i x_{i2} \ldots + \left(\frac{1}{n}\right) \sum_i y_i x_{im}$$
$$= r_1 + r_2 \ldots + r_m \text{ (the sum of the correlations).}$$

The variance of y is 1, and the variance of the sum of the Xs is $M + M(M-1)\bar{r}$, where $\bar{r}$ is the average intercorrelation between the Xs. Hence, the correlation of the average of the Xs with Y is $(\Sigma r_i)/(M + M(M-1)\bar{r})^{\frac{1}{2}}$; this is greater than $(\Sigma r_i)/(M + M^2 - M)^{\frac{1}{2}} =$ average r_i. Because each of the random models is positively correlated with the criterion, the correlation of their average, which is the unit-weighted model, is higher than the average of the correlations.

standers. When the ballistics experts rated the bullets with respect to these dimensions, it turned out that the last two were almost perfectly confounded, but they were not perfectly confounded with the first. Bullets do not vary along a single dimension that confounds effectiveness with lethalness. The probability of serious injury or harm to bystanders is highly related to the penetration of the bullet, whereas the probability of the bullet's effectively stopping someone from returning fire is highly related to the width of the entry wound. Since policymakers could not agree about the weights given to the three dimensions, Hammond and Adelman suggested that they be weighted equally. Combining the equal weights with the (independent) judgments of the ballistics experts, Hammond and Adelman discovered a bullet that "has greater stopping effectiveness and is less apt to cause injury (and is less apt to threaten bystanders) than the standard bullet then in use by the DPD" (Hammond & Adelman, 1976, p. 395). The bullet was also less apt to cause injury than was the bullet previously recommended by the DPD. That bullet was "accepted by the City Council and all other parties concerned, and is now being used by the DPD" (Hammond & Adelman, 1976, p. 395).[7] Once again, "the whole trick is to decide what variables to look at and then know how to add" (Dawes & Corrigan, 1974, p. 105).

So why don't people do it more often? I know of four universities (University of Illinois; New York University; University of Oregon; University of California, Santa Barbara—there may be more) that use a linear model for applicant selection, but even these use it as an initial screening device and substitute clinical judgment for the final selection of those above a cut score. Goldberg's (1965) actuarial formula for diagnosing neurosis or psychosis from MMPI profiles has proven superior to clinical judges attempting the same task (no one to my or Goldberg's knowledge has ever produced a judge who does better), yet my one experience with its use (at the Ann Arbor Veterans Administration Hospital) was that it was discontinued on the grounds that it made obvious errors (an interesting reason, discussed at length below). In 1970, I suggested that our fellowship committee at the University of Oregon apportion cutbacks of National Science Foundation and National Defense Education Act fellowships to departments on the basis of a quasi-linear point system based on explicitly defined indices, departmental merit, and need; I was told "you can't systemize human judgment." It was only six months later, after our committee realized the political and ethical impossibility of cutting back fellowships on the basis of intuitive judgment, that such a system was adopted. And so on.

In the past three years, I have written and talked about the utility (and in my view, ethical superiority) of using linear models in socially important decisions. Many of the same objections have been raised repeatedly by different readers and audiences. I would like to conclude this article by cataloging these objections and answering them.

Objections to Using Linear Models

These objections may be placed in three broad categories: technical, psychological, and ethical. Each category is discussed in turn.

TECHNICAL

The most common technical objection is to the use of the correlation coefficient; for example, Remus and Jenicke (1978) wrote:

> It is clear that Dawes and Corrigan's choice of the correlation coefficient to establish the utility of random and unit rules is inappropriate [*sic,* inappropriate for what?]. A criterion function is also needed in the experiments cited by Dawes and Corrigan. Surely there is a cost function for misclassifying neurotics and psychotics or refusing qualified students admissions to graduate school while admitting marginal students. (p. 221)

Consider the graduate admission problem first. Most schools have k slots and N applicants. The problem is to get the best k (who are in turn willing to accept the school) out of N. What better way is there than to have an appropriate rank? None. Remus and Jenicke write as if the problem were not one of comparative choice but of absolute choice. Most social choices, however, involve selecting the better or best from a set of alternatives: the students that will be better, the bullet that will be best, a possible airport site that will be superior, and so on. The correlation

[7] It should be pointed out that there were only eight bullets on the *Pareto frontier*; that is, there were only eight that were not inferior to some particular other bullet in both stopping effectiveness and probability of harm (or inferior on one of the variables and equal on the other). Consequently, any weighting rule whatsoever would have chosen one of these eight.

coefficient, because it reflects ranks so well, is clearly appropriate for evaluating such choices.

The neurosis–psychosis problem is more subtle and even less supportive of their argument. "Surely," they state, "there is a cost function," but they don't specify any candidates. The implication is clear: If they could find it, clinical judgment would be found to be superior to linear models. Why? In the absence of such a discovery on their part, the argument amounts to nothing at all. But this argument from a vacuum can be very compelling to people (for example, to losing generals and losing football coaches, who know that "surely" their plans would work "if"—when the plans are in fact doomed to failure no matter what).

A second related technical objection is to the comparison of average correlation coefficients of judges with those of linear models. Perhaps by averaging, the performance of some really outstanding judges is obscured. The data indicate otherwise. In the Goldberg (1970) study, for example, only 5 of 29 trained clinicians were better than the unit-weighted model, and none did better than the proper one. In the Wiggins and Kohen (1971) study, no judges were better than the unit-weighted model, and we replicated that effect at Oregon. In the Libby (1976) study, only 9 of 43 judges did better than the ratio of assets to liabilities at predicting bankruptcies (3 did equally well). While it is then conceded that clinicians should be able to predict diagnosis of neurosis or psychosis, that graduate students should be able to predict graduate success, and that bank loan officers should be able to predict bankruptcies, the possibility is raised that perhaps the experts used in the studies weren't the right ones. This again is arguing from a vacuum: If other experts were used, then the results would be different. And once again no such experts are produced, and once again the appropriate response is to ask for a reason why these hypothetical other people should be any different. (As one university vice-president told me, "Your research only proves that you used poor judges; we could surely do better by getting better judges"—apparently not from the psychology department.)

A final technical objection concerns the nature of the criterion variables. They are admittedly short-term and unprofound (e.g., GPAs, diagnoses); otherwise, most studies would be infeasible. The question is then raised of whether the findings would be different if a truly long-range important criterion were to be predicted. The answer is that of course the findings *could* be different, but we have no reason to suppose that they *would* be different. First, the distant future is in general less predictable than the immediate future, for the simple reason that more unforeseen, extraneous, or self-augmenting factors influence individual outcomes. (Note that we are not discussing aggregate outcomes, such as an unusually cold winter in the Midwest in general spread out over three months.) Since, then, clinical prediction is poorer than linear to begin with, the hypothesis would hold only if linear prediction got much worse over time than did clinical prediction. There is no a priori reason to believe that this differential deterioration in prediction would occur, and none has ever been suggested to me. There is certainly no evidence. Once again, the objection consists of an argument from a vacuum.

Particularly compelling is the fact that people who argue that different criteria or judges or variables or time frames would produce different results have had 25 years in which to produce examples, and they have failed to do so.

PSYCHOLOGICAL

One psychological resistance to using linear models lies in our selective memory about clinical prediction. Our belief in such prediction is reinforced by the availability (Tversky & Kahneman, 1974) of instances of successful clinical prediction—especially those that are exceptions to some formula: "I knew someone once with . . . who . . . ." (e.g., "I knew of someone with a tested IQ of only 130 who got an advanced degree in psychology.") As Nisbett, Borgida, Crandall, and Reed (1976) showed, such single instances often have greater impact on judgment than do much more valid statistical compilations based on many instances. (A good prophylactic for clinical psychologists basing resistance to actuarial prediction on such instances would be to keep careful records of their own predictions about their own patients—prospective records not subject to hindsight. Such records could make all instances of successful and unsuccessful prediction equally available for impact; in addition, they could serve for another clinical versus statistical study using the best possible judge—the clinician himself or herself.)

Moreover, an illusion of good judgment may be reinforced due to selection (Einhorn & Hogarth, 1978) in those situations in which the prediction of a positive or negative outcome has a self-fulfilling effect. For example, admissions officers who judge that a candidate is particularly qualified for a graduate program may feel that their judgment is exonerated when that candidate does well, even though the candidate's success is in large part due to the positive effects of the program. (In contrast, a linear model of selection is evaluated by seeing how well it predicts performance *within* the set of applicants selected.) Or a waiter who believes that particular people at the table are poor tippers may be less attentive than usual and receive a smaller tip, thereby having his clinical judgment exonerated.[8]

A second psychological resistance to the use of linear models stems from their "proven" low validity. Here, there is an implicit (as opposed to explicit) argument from a vacuum because neither changes in evaluation procedures, nor in judges, nor in criteria, are proposed. Rather, the unstated assumption is that these criteria of psychological interest are in fact highly predictable, so it follows that if one method of prediction (a linear model) doesn't work too well, another might do better (reasonable), which is then translated into the belief that another *will* do better (which is not a reasonable inference)—once it is found. This resistance is best expressed by a dean considering the graduate admissions who wrote, "The correlation of the linear composite with future faculty ratings is only .4, whereas that of the admissions committee's judgment correlates .2. Twice nothing is nothing." In 1976, I answered as follows (Dawes, 1976, pp. 6–7):

> In response, I can only point out that 16% of the variance is better than 4% of the variance. To me, however, the fascinating part of this argument is the implicit assumption that that other 84% of the variance is predictable and that we can somehow predict it.
>
> Now what are we dealing with? We are dealing with personality and intellectual characteristics of [uniformly bright] people who are about 20 years old. . . . Why are we so convinced that this prediction can be made at all? Surely, it is not necessary to read *Ecclesiastes* every night to understand the role of chance. . . . Moreover, there are clearly positive feedback effects in professional development that exaggerate threshold phenomena. For example, once people are considered sufficiently "outstanding" that they are invited to outstanding institutions, they have outstanding colleagues with whom to interact —and excellence is exacerbated. This same problem occurs for those who do not quite reach such a threshold level. Not only do all these factors mitigate against successful long-range prediction, but studies of the success of such prediction are necessarily limited to those accepted, with the incumbent problems of restriction of range and a negative covariance structure between predictors (Dawes, 1975).

Finally, there are all sorts of nonintellectual factors in professional success that could not possibly be evaluated before admission to graduate school, for example, success at forming a satisfying or inspiring libidinal relationship, not yet evident genetic tendencies to drug or alcohol addiction, the misfortune to join a research group that "blows up," and so on, and so forth.

Intellectually, I find it somewhat remarkable that we are able to predict even 16% of the variance. But I believe that my own emotional response is indicative of those of my colleagues who simply assume that the future is more predictable. *I want it to be predictable, especially when the aspect of it that I want to predict is important to me.* This desire, I suggest, translates itself into an implicit assumption that the future is in fact highly predictable, and it would then logically follow that if something is not a very good predictor, something else might do better (although it is never correct to argue that it necessarily will).

Statistical prediction, because it includes the specification (usually a low correlation coefficient) of exactly how poorly we can predict, bluntly strikes us with the fact that life is not all that predictable. Unsystematic clinical prediction (or "postdiction"), in contrast, allows us the comforting illusion that life is in fact predictable and that we can predict it.

ETHICAL

When I was at the Los Angeles Renaissance Fair last summer, I overhead a young woman complain that it was "horribly unfair" that she had been rejected by the Psychology Department at the University of California, Santa Barbara, on the basis of mere numbers, without even an interview. "How can they possibly tell what I'm like?" The answer is that they can't. Nor could they with an interview (Kelly, 1954). Nevertheless, many people maintain that making a crucial social choice without an interview is dehumanizing. I think that the question of whether people are treated in a fair manner has more to do with the question of whether or not they have been dehumanized than does the question of whether the treatment is face to face. (Some of the worst doctors spend a great deal of time conversing with

[8] This example was provided by Einhorn (Note 5).

their patients, read no medical journals, order few or no tests, and grieve at the funerals.) A GPA represents 3½ years of behavior on the part of the applicant. (Surely, not all the professors are biased against his or her particular form of creativity.) The GRE is a more carefully devised test. Do we really believe that we can do a better or a fairer job by a 10-minute folder evaluation or a half-hour interview than is done by these two mere numbers? Such cognitive conceit (Dawes, 1976, p. 7) is unethical, especially given the fact of no evidence whatsoever indicating that we do a better job than does the linear equation. (And even making exceptions must be done with extreme care if it is to be ethical, for if we admit someone with a low linear score on the basis that he or she has some special talent, we are automatically rejecting someone with a higher score, who might well have had an equally impressive talent had we taken the trouble to evaluate it.)

No matter how much we would like to see this or that aspect of one or another of the studies reviewed in this article changed, no matter how psychologically uncompelling or distasteful we may find their results to be, no matter how ethically uncomfortable we may feel at "reducing people to mere numbers," the fact remains that our clients are people who deserve to be treated in the best manner possible. If that means—as it appears at present—that selection, diagnosis, and prognosis should be based on nothing more than the addition of a few numbers representing values on important attributes, so be it. To do otherwise is cheating the people we serve.

REFERENCE NOTES

1. Thornton, B. Personal communication, 1977.
2. Slovic, P. Limitations of the mind of man: Implications for decision making in the nuclear age. In H. J. Otway (Ed.), *Risk vs. benefit: Solution or dream?* (Report LA 4860-MS). Los Alamos, N.M.: Los Alamos Scientific Laboratory, 1972. [Also available as *Oregon Research Institute Bulletin*, 1971, *11*(17).]
3. Srinivisan, V. *A theoretical comparison of the predictive power of the multiple regression and equal weighting procedures* (Research Paper No. 347). Stanford, Calif.: Stanford University, Graduate School of Business, February 1977.
4. von Winterfeldt, D., & Edwards, W. *Costs and payoffs in perceptual research.* Unpublished manuscript, University of Michigan, Engineering Psychology Laboratory, 1973.
5. Einhorn, H. J. Personal communication, January 1979.

REFERENCES

Alexander, S. A. H. *Sex, arguments, and social engagements in marital and premarital relations.* Unpublished master's thesis, University of Missouri—Kansas City, 1971.

Beaver, W. H. Financial ratios as predictors of failure. In *Empirical research in accounting: Selected studies.* Chicago: University of Chicago, Graduate School of Business, Institute of Professional Accounting, 1966.

Bowman, E. H. Consistency and optimality in managerial decision making. *Management Science*, 1963, *9*, 310–321.

Cass, J., & Birnbaum, M. *Comparative guide to American colleges.* New York: Harper & Row, 1968.

Claudy, J. G. A comparison of five variable weighting procedures. *Educational and Psychological Measurement*, 1972, *32*, 311–322.

Darlington, R. B. Reduced-variance regression. *Psychological Bulletin*, 1978, *85*, 1238–1255.

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

Dawes, R. M. Graduate admissions criteria and future success. *Science*, 1975, *187*, 721–723.

Dawes, R. M. Shallow psychology. In J. Carroll & J. Payne (Eds.), *Cognition and social behavior.* Hillsdale, N.J.: Erlbaum, 1976.

Dawes, R. M., & Corrigan, B. Linear models in decision making. *Psychological Bulletin*, 1974, *81*, 95–106.

Deacon, E. B. A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 1972, *10*, 167–179.

deGroot, A. D. *Het denken van den schaker* [*Thought and choice in chess*]. The Hague, The Netherlands: Mouton, 1965.

Edwards, D. D., & Edwards, J. S. Marriage: Direct and continuous measurement. *Bulletin of the Psychonomic Society*, 1977, *10*, 187–188.

Edwards, W. M. Technology for director dubious: Evaluation and decision in public contexts. In K. R. Hammond (Ed.), *Judgement and decision in public policy formation.* Boulder, Colo.: Westview Press, 1978.

Einhorn, H. J. Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, 1972, *7*, 86–106.

Einhorn, H. J., & Hogarth, R. M. Unit weighting schemas for decision making. *Organizational Behavior and Human Performance*, 1975, *13*, 171–192.

Einhorn, H. J., & Hogarth, R. M. Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, 1978, *85*, 395–416.

Gardiner, P. C., & Edwards, W. Public values: Multiattribute-utility measurement for social decision making. In M. F. Kaplan & S. Schwartz (Eds.), *Human judgment and decision processes.* New York: Academic Press, 1975.

Goldberg, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs*, 1965, *79*(9, Whole No. 602).

Goldberg, L. R. Seer over sign: The first "good" example? *Journal of Experimental Research in Personality*, 1968, *3*, 168–171.

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. Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs,* 1972, No. 72-2.

Goldberg, L. R. Man versus model of man: Just how conflicting is that evidence? *Organizational Behavior and Human Performance,* 1976, *16,* 13–22.

Green, B. F., Jr. Parameter sensitivity in multivariate methods. *Multivariate Behavioral Research,* 1977, *3,* 263.

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

Hammond, K. R., & Adelman, L. Science, values, and human judgment. *Science,* 1976, *194,* 389–396.

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

Holt, R. R. Yet another look at clinical and statistical prediction. *American Psychologist,* 1970, *25,* 337–339.

Howard, J. W., & Dawes, R. M. Linear prediction of marital happiness. *Personality and Social Psychology Bulletin,* 1976, *2,* 478–480.

Kelly, L. Evaluation of the interview as a selection technique. In *Proceedings of the 1953 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1954.

Krantz, D. H. Measurement structures and psychological laws. *Science,* 1972, *175,* 1427–1435.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. *Foundations of measurement* (Vol. 1). New York: Academic Press, 1971.

Libby, R. Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance,* 1976, *16,* 1–12.

Marquardt, D. W., & Snee, R. D. Ridge regression in practice. *American Statistician,* 1975, *29,* 3–19.

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

Meehl, P. E. Seer over sign: The first good example. *Journal of Experimental Research in Personality,* 1965, *1,* 27–32.

Nisbett, R. E., Borgida, E., Crandall, R., & Reed, H. Popular induction: Information is not necessarily normative. In J. Carrol & J. Payne (Eds.), *Cognition and social behavior.* Hillsdale, N.J.: Erlbaum, 1976.

Remus, W. E., & Jenicke, L. O. Unit and random linear models in decision making. *Multivariate Behavioral Research,* 1978, *13,* 215–221.

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

Schmidt, F. L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement,* 1971, *31,* 699–714.

Simon, H. A., & Chase, W. G. Skill in chess. *American Scientist,* 1973, *61,* 394–403.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance,* 1971, *6,* 649–744.

Thornton, B. Linear prediction of marital happiness: A replication. *Personality and Social Psychology Bulletin,* 1977, *3,* 674–676.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science,* 1974, *184,* 1124–1131.

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin,* 1976, *83,* 312–317.

Wainer, H., & Thissen, D. Three steps toward robust regression. *Psychometrika,* 1976, *41,* 9–34.

Wallace, H. A. What is in the corn judge's mind? *Journal of the American Society of Agronomy,* 1923, *15,* 300–304.

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

Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika,* 1938, *8,* 23–40.

Yntema, D. B., & Torgerson, W. S. Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics,* 1961, *2*(1), 20–26.