

## Chapter 15

# Quantitative judgment

Quantitative judgment is the evaluation of cases on the basis of a set of evidence and with respect to a set of criteria. Some judgments involve assigning numbers: for example, assigning grades to students' essays, salaries to employees, or sales quotas to sales people. Quantitative judgments can involve ranking people or things: ranking entrants in a beauty contest, applicants for graduate school (to be accepted in order of rank as places become available), or applicants for a job. In some cases, the judgment is basically a matter of determining whether a person or thing is above or below some cutoff point. For example, is this patient sufficiently depressed so as to require hospitalization? Quantitative judgment, then, consists of *rating*, assigning numbers or grades; *ranking*, putting things in order on some dimension; and *classifying*, which in this chapter will mean assigning something to one of two groups.

When we judge livestock, paintings, or automobiles with a view to purchasing, our judgment of each item provides input to a decision that will select one or more of these items. We can think of quantitative judgment as part of a decision process in which each option is first evaluated separately, and the final decision is based on a comparison of the evaluations. (Other decisions might be made by looking only at differences among options.)

In most experiments on judgment, subjects are presented with evidence about attributes of each option and are asked to evaluate each option. For example, subjects might be given a student's test scores, grades, and disposable income and asked to judge the student's worthiness for a scholarship on a 100-point scale. Or the subjects might evaluate a car on the basis of its price, safety record, handling, and other features. Typically the subjects use their goals, so we can infer their utilities from their responses. Judgment tasks are thus another way of eliciting judgments of utility, and we shall discuss that application in this chapter. The study of judgment is closely related to Multiattribute Utility Theory (MAUT) (discussed in Chapter 14). In most judgment tasks that have been studied, the cases to be judged can be described in terms of a set of attributes, just as if MAUT were going to be applied.

Some judgment experiments have the same form but use goals other than the

Table 15.1: Measures used in regression example

| Abbreviation | Definition   |
|--------------|--|
| <i>GPA</i>   | High school grade point average on a scale from 0 to 4.0.  |
| <i>SAT</i>   | Total Scholastic Aptitude Test score on a scale up to 1600.  |
| <i>REC</i>   | A rating summarizing the letters of recommendation on a scale from 1 to 5. This rating is assigned by an admissions officer who reads the letters of recommendation and assigns the number. (Notice that here a judgment is used as input for another judgment.) |
| <i>ESS</i>   | A rating of the student's own essay on a 1 to 5 scale.   |

subjects' own. We can even ask the subjects to predict some objective criterion. This procedure may involve many of the same processes as evaluation, so we discuss it here too.

A major question in the study of judgment concerns the relative efficacy of unaided holistic judgment — in which the judge simply assigns a number to each case — and judgment aided by the use of calculations, like those done in MAUT. The answers to this question have strong implications for decision making in government, business, and the professions, and in any situation where important decisions are made about a number of cases described on the same dimensions.

### Multiple linear regression

Most of the literature on judgment has looked at situations in which each of a number of possibilities, such as applicants for college, is characterized by several numbers. Each number represents a value on some dimension, or *cue*, such as grades or quality of recommendations. The judge's task is to evaluate each possibility with respect to some goal (or goals), such as college grades as a criterion of success in college. Each dimension or cue has a high and low end with respect to the goal. For example, high test scores are assumed to be better than low test scores. In these situations a certain kind of normative model is assumed to apply, specifically, a statistical model called *multiple linear regression* (or just *regression*, for short).

Suppose that we are admissions officers at a college, and our task is to rate several applicants. For each applicant, we have the numerical ratings, scores and other information given in Table 15.1. Suppose also that we have a computer, and we want to discover a formula for predicting the applicant's *college* grade-point average, which we call *COL*.

We have data on a number of students who are now at the college. For each student, we know the four variables given in the table (*SAT*, the aptitude test score; *REC*, a rating of the letters of recommendation; *ESS*, a rating of the essay; and *GPA*, high school grades), and we know *COL*. What we want is a predictive index,

Table 15.2: Data and predictions (*PRE*) for regression example

| Student | <i>COL</i> | Predictors |            |            |            | <i>PRE</i> | Error  |
|---------|------------|------------|------------|------------|------------|------------|--------|
|         |            | <i>SAT</i> | <i>REC</i> | <i>ESS</i> | <i>GPA</i> |            |        |
| 1       | 3.8        | 1500       | 4.0        | 4.0        | 4.0        | 3.910      | 0.110  |
| 2       | 3.6        | 1310       | 4.0        | 3.0        | 3.6        | 2.902      | -0.698 |
| 3       | 3.5        | 1300       | 5.0        | 3.0        | 3.9        | 3.560      | 0.060  |
| 4       | 3.2        | 1280       | 3.0        | 5.0        | 3.7        | 3.428      | 0.228  |
| 5       | 3.0        | 1260       | 4.0        | 4.0        | 3.5        | 2.921      | -0.079 |
| 6       | 2.8        | 1210       | 3.0        | 4.0        | 3.4        | 2.631      | -0.169 |
| 7       | 2.5        | 1320       | 5.0        | 3.0        | 3.5        | 2.807      | 0.307  |
| 8       | 2.2        | 1220       | 4.0        | 3.0        | 3.2        | 2.129      | -0.071 |
| 9       | 2.0        | 1200       | 2.0        | 5.0        | 3.0        | 1.997      | -0.003 |
| 10      | 1.5        | 1170       | 3.0        | 2.0        | 3.2        | 1.811      | 0.311  |

*PRE*, a measure that comes as close as possible to predicting *COL* for the new applicants from the four other variables. Table 15.2 shows, for 10 of the college students, the four variables, the college grades (*COL*), and the predictive index (*PRE*). The last column is the error in each prediction, the difference between *PRE* and *COL*. *PRE* was calculated using a computer program for multiple linear regression (a technique described in detail in most modern statistics texts). The numbers in the first five columns were typed into the computer. The regression program was told to assume that *COL* is a linear function of the four other variables — that is, to assume that the following equation is true:

$$COL = a \cdot SAT + b \cdot REC + c \cdot ESS + d \cdot GPA + e + error$$

The error is a different number for each student, but each of the other coefficients — *a*, *b*, *c*, *d*, and *e* — is the same for all students. The coefficients *a* through *d* may be seen as weights; they indicate how much each of the variables (*SAT*, and so forth) affects *COL*. The coefficient *e* is a constant that is added or subtracted so that the mean predicted *COL* comes out right. The computer figures out the values of *a*, *b*, *c*, *d*, and *e* so as to make the error as small as possible. (Usually the computer does this by minimizing the mean of the squares of the error values.) Once we know the values of the coefficients, *PRE* is found by using the same equation but without the error:

$$PRE = a \cdot SAT + b \cdot REC + c \cdot ESS + d \cdot GPA + e$$

*PRE* would be as close as we can come to predicting *COL* with this kind of formula, if all we knew were the other four variables. Using the data given in Table 15.2,

the values of  $a$  through  $e$ , respectively, are 0.000175 (for  $SAT$ ); 0.092 (for  $REC$ ); 0.217 (for  $ESS$ ); 1.893 (for  $GPA$ ); and  $-5.161$  (the constant  $e$  added at the end). Therefore, the equation for  $PRE$  is as follows:

$$PRE = 0.000175 \cdot SAT + 0.092 \cdot REC + 0.217 \cdot ESS + 1.893 \cdot GPA - 5.161$$

Notice that the constant ( $-5.161$ ) would not be needed, if all we wanted to do was to compare students with one another. What we are trying to do, though, is to compare the four variables with each other, to determine their relative importance. The four values  $a$  through  $d$  represent the relative importance of each of the four variables.<sup>1</sup>

Notice that  $SAT$  is unimportant in the formula. Notice also, however, that  $SAT$  does correlate with  $COL$ : The student with the highest  $COL$  got the highest  $SAT$ , and the two students with the lowest  $COL$  got the two lowest  $SAT$ 's. How could this happen? The answer is that  $SAT$  correlates with  $GPA$  in high school, and  $GPA$  in high school also correlates with  $COL$ . The reason that  $SAT$  correlates with  $COL$  appears to be that it correlates with  $GPA$ . Here,  $SAT$  seems to measure something like the ability to get good grades, but it does not measure this as well as the grades themselves. If we did not know high school  $GPA$ , then the  $SAT$  would be a useful predictor of  $COL$ , in this example.

This example is, of course, overly simple. In deciding whom to admit to college, there are other predictors to consider aside from these four factors, and there are other things to predict besides college grades. Many of these variables can be expressed as numbers, but when we apply a numerical formula of this sort we may always run across unusual cases that require us either to make exceptions or add a variable to the formula for the benefit of a single case. Would it make sense to include a measure of every applicant's criminal record, when probably only a few applicants have any record at all?

Moreover, the idea of a formula might be too simple. The basic idea of the model — that everything is multiplied by a weight representing its importance and all of the values are then added together — might be wrong. One way in which the model could be wrong is that there might be an *interaction* between two variables. This means that the *importance* (or weight) of one variable depends on the *value* of the other. For example, perhaps we should weigh  $REC$  more when a student does poorly on the Scholastic Aptitude Test. This would amount to accepting students who did well either on that test or in  $REC$  (or both), no matter how badly they might have done on one of the measures. Thus, when one of these two measures was high, the other would not matter.

Another way in which the model could be wrong is that some variables might not have a simply linear effect. The importance of a variable might be different for different parts of its range. For example, the difference between a  $SAT$  of 1100 and one of 1200 might be much more important than the difference between a  $SAT$  of

<sup>1</sup>A better measure of importance would take into account the amount of variation on each variable. In this case,  $SAT$ 's would have even less weight, because they vary by hundreds of points instead of just a few.

Table 15.3: Data for regression example, with judgments ( $JUD$ )

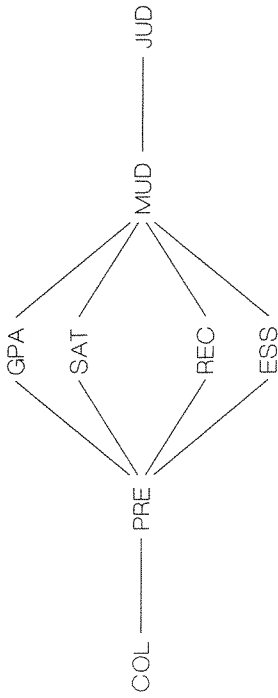
| Student | COL | Predictors |     |     |     | GPA   | PRE | Error |
|---------|-----|------------|-----|-----|-----|-------|-----|-------|
|         |     | SAT        | REC | ESS | JUD |       |     |       |
| 1       | 3.8 | 1500       | 4.0 | 4.0 | 4.0 | 3.910 | 4.0 |       |
| 2       | 3.6 | 1310       | 4.0 | 3.0 | 3.6 | 2.902 | 3.1 |       |
| 3       | 3.5 | 1300       | 5.0 | 3.0 | 3.9 | 3.560 | 3.8 |       |
| 4       | 3.2 | 1280       | 3.0 | 5.0 | 3.7 | 3.428 | 3.4 |       |
| 5       | 3.0 | 1260       | 4.0 | 4.0 | 3.5 | 2.921 | 3.0 |       |
| 6       | 2.8 | 1210       | 3.0 | 4.0 | 3.4 | 2.631 | 2.7 |       |
| 7       | 2.5 | 1320       | 5.0 | 3.0 | 3.5 | 2.807 | 3.0 |       |
| 8       | 2.2 | 1220       | 4.0 | 3.0 | 3.2 | 2.129 | 2.3 |       |
| 9       | 2.0 | 1200       | 2.0 | 5.0 | 3.0 | 1.997 | 1.9 |       |
| 10      | 1.5 | 1170       | 3.0 | 2.0 | 3.2 | 1.811 | 2.1 |       |

1300 and one of 1400. The effect of the  $SAT$  would then be *curvilinear* rather than linear. (We might want to use for our calculations something such as the square root of  $SAT$  rather than  $SAT$  itself.) We shall return to this question.

## The lens model

Suppose we asked an admissions official to predict  $COL$  (college grades) without the benefit of the formula. We could then obtain a list of judgments that we could place beside the true values for comparison, as shown in Table 15.3. We could then ask several questions about these judgments. For example, we could ask how close they come to the true values, or whether the judgments themselves could be predicted from the four main variables, and so on.

One useful way to think about this kind of situation is the *lens model*, based on the work of Brunswik (1952), Hammond (1955), and others. This term results from the sort of diagram shown here, which is supposed to look something like light rays being focused by a lens. Each line in the diagram represents a relationship, usually a correlation. (Some possible lines of correlation are left out of my diagram.) Each variable has a particular role.  $COL$  is the *criterion*, the thing to be predicted.  $JUD$  is the *judgment* provided by the judge.  $PRE$  is the value of  $COL$  predicted from the regression formula.



The new idea in the diagram is *MUD* (the Model of the *JUD*ge). This is what we get if we try to predict *JUD* from the four main variables, just as we originally tried to predict *COL* from these variables. *MUD* is a *model of the judge*, just as *PRE* is a model of the criterion. In this example, *MUD* is based on the following equation:

$$MUD = 0 \cdot SAT + 0.1 \cdot REC + 0.1 \cdot ESS + 2.0 \cdot GPA - 4.8$$

Compared to *PRE*, the judge depends a little too much on *GPA* (2.0 for the judge versus 1.893 for *PRE*) and not enough on *ESS* (0.1 versus 0.217). This leads the judge, for example, to predict too high for student 10, whose *ESS* rating was very low. On the whole, however, the judge does well, almost as well as the formula for *PRE*. (Because that formula was chosen to minimize error, the judge cannot possibly do as well on evaluating these students' potential, unless the judge uses the same formula.)

Notice that the correlation between *MUD* and *JUD* is perfect (1.00) in this case. *MUD* is identical to *JUD*. This is not the usual case. It seems that our judge was calculating, rather than making an intuitive judgment, but the judge was not using quite the best formula for the calculations.

We can use the lens model to answer a great variety of questions about judgment in particular situations. Most of these judgment situations are quite similar to the one in the example just given. A judge is asked to predict some numerical criterion, such as stock prices, success in school or work, livestock quality, or *COL*, from a set of numerical predictors. The predictors (or the criterion) themselves sometimes represent summaries of other judgments. There is reason to think that each of the predictors might be related to the criterion in a simple way — that is, by either a positive or a negative correlation. Let us consider some of the questions about such judgments, and their answers.

*Which does better, the judge or the best-fitting model of the data (PRE)?* In practically every study in which this question has been asked, the answer is that the model does better over a set of judgments. When possible, it is always better to use a formula than an individual human judgment (Meehl, 1954; Goldberg, 1970; Dawes, 1971, 1979; Dawes and Corrigan, 1974; Camerer, 1981). This has been found in studies of the judgment of how psychotic people are (from personality-test profiles), graduate student success, success of people in various jobs, and future stock prices.

*Which does better, the judge (JUD) or the model of the judge (MUD)?* Again, in practically every study, the answer is that the model of the judge does better than the judge. Suppose we have a judgment task such as predicting college grades, and we have a judge who claims to be able to do it. Consider two ways we could proceed. The first is to have the judge make judgments of every case. The second is to have the judge make judgments of a number of cases, enough so that we can find a formula that predicts *the judge's judgments*. Then we would tell the judge to go home, and we would use the formula for all the cases, including those the judge already judged. The fact is that the second method is better than the first, even for the cases that the judge already judged. (Note that we are using the formula for predicting the *judgments*, not the formula for predicting what the judge is trying to predict. The latter formula, when we can find it, works even better at predicting the criterion.)

This is a rather surprising result. Why does it happen? Basically, it happens because the judge cannot be consistent with his own policy (unless, like our judge in the example, he calculates on his own). He is unreliable, in that he is likely to judge the same case differently on different occasions (unless he recognizes the case the second time). As Goldberg (1970, p. 423) puts it,

The clinician is not a machine. While he possesses his full share of human learning and hypothesis-generating skills, he lacks the machine's reliability. He "has his days": Boredom, fatigue, illness, situational and interpersonal distractions all plague him, with the result that his repeated judgments of the exact same stimulus configurations are not identical. He is subject to all those human frailties which lower the reliability of his judgments below unity. And if the judge's reliability is less than unity, there must be error in his judgments — error which can serve no other purpose but to attenuate his accuracy. If we could remove some of this human unreliability by eliminating the random error in his judgments, we should thereby increase the validity of the resulting predictions.

We can, however, think of reasons why the judge might still be better than the model of himself. In particular, there might be interactions or curvilinear relations (nonlinearities) that are not represented in the linear models we have been considering. It might really be true that *SAT*s are more important in the middle range than the higher range, for example. If this were true, the judge would be able to take it into account, and the linear model of the judge would not.

The judge might also have additional information not included in the linear model, such as observations made during a personal interview with the applicant. This gives the judge an unfair advantage over the model, however, since it is possible for the judge to assign numerical ratings to that information and include it in the model.

In the great variety of situations examined so far, interactions and nonlinearities have never been as important as the error to which Goldberg refers. There are

sometimes nonlinearities and interactions in the true situation, and there are sometimes nonlinearities and interactions in the judge's judgments, but the nonlinearities and interactions expressed in his actual judgments have little to do with the nonlinearities and interactions actually present (Camerer, 1981). (Later in this chapter, however, we shall consider a case in which subjects make use of interactions among features when deciding whether a stimulus is a member of a category.)

In general, then, it is better to use the formula than the judge, because the formula eliminates the error. This method is called *bootstrapping*. We pull ourselves up by our bootstraps. We improve our judgment by modeling ourselves. A good example of the successful use of bootstrapping methods is a formula devised by Gustafson, Tianen, and Greist (1981), which predicts suicide attempts in psychiatric patients on the basis of variables that can be assessed in clinical interviews (such as extent of suicide plans and degree of isolation from other people). The predictions are more accurate than those made holistically by the judges.

When we use bootstrapping to make decisions, we estimate weights from holistic decisions, rather than from systematic considerations of tradeoffs; we calculate the weights so as to best explain a person's holistic ratings.<sup>2</sup> Although bootstrapping does better than holistic judgments, it is still derived from such judgments. We noted in Chapter 12 that less important dimensions will be underweighted in holistic judgments. Bootstrapping does solve the problem of random error, but it does not remove the effects of this underweighting. Studies have found that in a bootstrapping procedure the dimensions considered less important in fact receive less weight, relative to other dimensions, than in a MAUT procedure (von Winterfeldt and Edwards, sec. 10.4). In addition, bootstrapping leads to greater disagreement among judges than MAUT.

Another way to reduce judgment error is to average the judgment of several different judges. When this is done, the errors of different judges tend to cancel each other out. Such *composite* judgments are indeed more accurate than judgments made by individuals. If enough judges are used, they may do as well as a formula based on the composites themselves. Goldberg (1970) examined the ability of 29 clinical psychologists to predict the diagnosis of mental patients ("psychotic" versus "neurotic") from test profiles. The best prediction was accomplished by the composite of the 29 judges. A model of the composite did a little worse, but not significantly so. Next best was a model of the average *individual judge*, and worst of all was the average individual judge alone. Although use of many judges for prediction may help, it is much more time-consuming than use of a single judge. The model of the group may be best of all, taking efficiency into account.

Formulas may do better than judges, even when there is no objective way at all to measure the criterion. For example, suppose we wanted to judge desirability of prospective graduate students, on the basis of *GREs* (Graduate Record Examination scores), college *G.P.As*, and ratings of letters and essays, for the purpose of admitting candidates to graduate school. We do not know, in this case, what "desirability"

<sup>2</sup>We also assume that the utility scales are linear.

means. It is not the same as grades in graduate school, for nebulous things such as quality of scholarship are important as well. What we do is to take a judge, or a committee of judges, and ask them to rate a number of prospective students. We then find a formula (*MUD*) that accounts for *their judgments* (*JUD*). We tell the judges to go home, and we use the formula.

Because of the findings that *MUD* is better than *JUD* at predicting an objective criterion, we can assume with some confidence that here, too, *MUD* would do better, if *there were an objective criterion*. The fact that there is not can hardly be expected to affect the superiority of *MUD* over *JUD*. We can be confident that we are doing better with the formula than with the formula derived from the judge's judgments, even when we cannot measure how much better we are doing. Another advantage of a method like this is that it may prevent complex judgments from being reduced to single numbers, such as use of grades alone as a measure of academic success or (in war) body counts of enemy soldiers as a measure of military success.

*Do we need the judge at all?* Dawes (1979) points out that in most situations of interest the variables we use to make predictions tend to be correlated with each other. For example, grades, *SATs*, and recommendations tend to be correlated; a student who does well on one measure tends to do well on others. In such cases, the precise weight we give to each variable does not make much difference. If we overweight one variable and underweight another, and if the two are correlated, our judgments (in relative terms) will not be very different. In our example, the judge did almost as well as the ideal model, even though he weighed things somewhat differently.

What this implies is that we can often do quite well if we simply know *which* variables to use (and in which direction). For example, in our case concerning admission, an expert might tell us that grades, recommendations, and essays were all important, but *SATs* were not. (It would not have produced such different results if we had used *SATs*.) Often, consulting an expert is not really necessary. Of course, there are cases, such as medical diagnosis, in which a nonexpert has no idea what is relevant. Would you know which measures to look at in order to determine whether someone has diabetes? In such cases, an expert can go a long way simply by knowing what is relevant, whether she weighs it correctly or not.

When we use the method of weighing all predictors equally, we must make sure that they are all expressed in the same sort of units. In our *COL* example, *REC*, *ESS*, and *GPA* have about the same range, but *SAT* has a much larger range. If we simply added up the predictors, *SAT* alone would determine the ranking, so the predictors would not be weighed "equally" at all. The way this problem is usually handled is to divide each predictor by the *standard deviation* of that predictor. The standard deviation is — like the range between the highest and lowest scores — a measure of the extent to which that predictor varies across all the cases (here, students).<sup>3</sup> In our *COL* example, the standard deviations of the predictors are *SAT*,

<sup>3</sup>The standard deviation is essentially the square root of the average difference between each score and the average score.

89; *REC*, 0.90; *ESS*, 0.92; and *GPA*, 0.30.

Dawes (1979) describes a number of cases in which such an equal-weighting method has been used with apparent success. In those cases in which the question can be asked, the equal-weighting model did better than the judge by himself. So great is the judge's unreliability that we can do better than the judge (in most cases) by knowing only which variables are relevant.

*Why are people so resistant to the making of decisions by formula?* Universities and colleges typically pay an admissions staff substantial amounts to read each application and weigh all of the variables. Faculty members do the same for applicants to graduate school and applicants for jobs. Business managers devote great time and effort to such matters as the setting of sales and production quotas. In all of these cases, there is a need for human judgment, to be sure. Somebody has to read the letters of recommendation and (at least) assign numbers to them, but there is no apparent need for the human judge when it comes to combining all of the various pieces into a single overall judgment. This is done better by formula.

Many people say that they are reluctant to use formulas because they want to be free to use a special policy for special cases. Among graduate-school applicants, for example, the foreigner whose poor command of English keeps test scores down, or the handicapped person who has overcome enormous odds just to meet minimal admission standards, and so on. This is not a good reason to reject formulas altogether. These cases do not have to be ignored when we use the formula. We can include a "special factors" variable in our formula, or omit tainted data (such as a verbal *SAT* score for foreigners), or we can simply scan through the pile of applicants looking for such exceptions and judge just these individually.

Another explanation of the reluctance to use formulas is that the judgment procedure often has other goals besides simply making the best judgment. For example, when my own department admits graduate students, the admissions committee asks faculty members who are likely to get these students in their laboratories to make an overall evaluation of each such applicant. The admissions committee does this not because it believes, for example, that a physiological psychologist is a better judge of physiological applicants than the committee is. Rather, it wants to alert faculty members who are not on the committee to the fact that certain students may end up working in their laboratories. (Technically, a memorandum to this effect would serve the purpose, but it is well known that many of us do not read memos.)

Another objection to formulas is the argument that individuals cannot be (or should not be) reduced to a single number. When we put individuals in rank order for some purpose, however, we are already reducing them to a single scale. The only issue is whether we do it well or badly given all the goals we ought to have.

Many of the people who object to the use of formulas are unaware of a hidden cause for their belief — *overconfidence* in their own powers of judgment. Psychologists have found this overconfidence factor elsewhere, in probability judgments (Chapter 6) and in phenomena such as the overuse of interviews for evaluating applicants. Hiring and admissions personnel often feel that they need a personal interview with each of many applicants, as though the fifteen- or thirty-minute sample of be-

havior will stack up against the four years of data represented in a student's transcript and recommendations or the ten years in a résumé. Evidence suggests that interviews ordinarily add nothing to the validity of prediction based on all other data that are usually available (Schmitt, 1976). If you were choosing members of an Olympic team, would you set up a thirty-minute tryout or look at each person's track record (literally) over the last few years?<sup>4</sup> An interview, of course, is like a thirty-minute tryout.<sup>5</sup>

Overconfidence in judgment has been found in an experiment by Arkes, Dawes, and Christensen (1986), who asked college-student subjects to choose which of three players had won the Most Valuable Player Award of the National Baseball League in each of the years from 1940 to 1961 (excluding those years in which pitchers won the award). Subjects were given the batting average, the number of home runs hit, the number of runs batted in, and the position of the player's team, for each of the three players listed for each year. Subjects were told — correctly — that the position of the player's team was an extremely useful cue, which would lead to 14 correct choices out of 19 if it was the sole basis for the judgment. Subjects who knew more about baseball (as determined from their answers to a baseball quiz) were more confident in their ability to "beat" the simple rule they had been given concerning team position. As a result of this confidence, they used the position rule less often and were actually less accurate in choosing the award winner (9.4 correct out of 19) than were other subjects with more modest knowledge of baseball (11.4 correct). Neither group of subjects did as well as they could have done by taking the experimenter's advice and attending only to team position. Overconfidence can be exacerbated by expertise.

We are not always, or usually, good judges of our own ability as judges. As a result, we waste effort making judgments that could be made more accurately by formula. The use of a formula guarantees that errors will be made, but we hope, however vainly, that human judgments will be perfect. As Einhorn (1986) puts it, we need to "accept error to make less error."

When we, as applicants, are rejected for a job or a school by a formula, we feel that an error is likely. The formula could not possibly have seen our special attributes (even though the teachers who gave us the grades and wrote the letters of recommendation, whose ratings go into the formula, had every opportunity to do so). When we are rejected by a person, we are more likely to feel that the rejection was based on real knowledge, yet, as we have seen, people make more errors than formulas (if we use people's judgments of specific dimensions as inputs).

A warning is in order lest anyone get carried away with the idea of replacing human judges with computers or even hand calculators. Although many of the objections to formulas are (arguably) mere mystical rhetoric, one is worth heeding. This is that the use of formulas may encourage us to use only the predictors that are easily quantified. For example, in admitting students to graduate school, we might tend to rely too heavily on test scores and grades and not enough on recommendations or

<sup>4</sup> Apparently, some nations do use tryouts instead of track records.

<sup>5</sup> On the other hand, an interview, even by telephone, may provide important information not usually included in a résumé, such as how serious the candidate's interest in the position in question is.

research papers. If admissions committees do this, then students will overemphasize tests and grades too, as they prepare themselves for graduate school. The solution to this problem is to turn recommendations and research papers into numbers so that they can be included in a formula. Indeed, in most of the examples given so far in this chapter, the *input* to the judgment process has consisted of other judgments, for which human judges are necessary. In many of the situations in which formulas are useful, the numbers are already available. For example, college admissions offices make ratings on applicants' essays, and even on their overall "sparkle," although these numbers are not always included in a formula.

The research that I have been discussing does not suggest that formulas should replace people altogether. It suggests, rather, that formulas should replace people *at the single task of combining a set of piecemeal judgments into an overall summary score*.

## The mechanism of judgment

### Do people really follow linear models?

Do judges really weigh cues numerically? The question is particularly acute when the task involves classifying stimuli into two categories (such as "neurotic" and "psychotic") on the basis of numerical predictors. A reasonable alternative strategy, it might seem, is to set cutoff points on each predictor. In order to be classified as psychotic, for example, a patient would have to exceed the cutoff point on a certain number of predictors (perhaps one, perhaps all). Although such a strategy would probably not be optimal, since it throws away information about how far from the cutoff point each patient is, we might use it anyway because it is easy. (A patient who is far above the cutoff point on one predictor and just below it on another is probably more likely to be psychotic than one who is just above the cutoff on both, yet the use of a strategy requiring both cutoff points to be exceeded would classify the former patient as neurotic and the latter as psychotic.)

Thinking-aloud protocols have been used to study judgment tasks (as well as other thinking tasks), and these results suggest that individual judges often use some sort of cutoff strategy (Payne, 1976; Einhorn, Kleinmuntz, and Kleinmuntz, 1979). Here are some fictional protocols (from Einhorn et al., p. 473) of a typical judge deciding whether a patient is psychologically "adjusted" or "maladjusted," on the basis of three numerical cues,  $x_1$ ,  $x_2$ , and  $x_3$ , which represent scores on various psychological tests:

*Case 1.* I'll look at  $x_1$  first — it's pretty high ... the  $x_2$  score is also high. This is a maladjusted profile.

*Case 2.* Let's see,  $x_1$  is high but ...  $x_2$  is low, better check  $x_3$  — it's low, too. I would say this is adjusted.

*Case 3.* This person's  $x_1$  score is low. Better check both  $x_2$  and  $x_3$  then — both pretty high. Ummm ... likely to be maladjusted.

*Case 4.*  $x_1$  is fairly low here ...  $x_2$  is quite high ... this is an interesting case ...  $x_3$  is very low ... I'd say adjusted.

*Case 5.*  $x_1$  is extremely high — this is maladjusted.

*Case 6.*  $x_1$  is an iffy score ... let's see  $x_2$  and  $x_3$ . Both are mildly indicative of pathology — I guess, that taking all three are pointing in the same direction, call this maladjusted.

Einhorn and his colleagues point out that these protocols can be described in terms of various rules involving cutoffs. For Case 1, it looks as though the judge is applying some sort of cutoff point to the  $x_1$  and  $x_2$  scores. If these are high enough, she does not look at  $x_3$ . In Case 2, it looks as though there is a rule instructing her to look at  $x_3$  when the other cues conflict. Case 4 is "interesting," because two cues that are usually correlated conflict. In Case 5, it looks as though the judge is applying another cutoff point to  $x_1$ ; if  $x_1$  is sufficiently high, she does not look at the other cues at all.

Einhorn and his colleagues point out that these protocols are also consistent with the use of a "compensatory" strategy more in line with the linear model itself. For example, suppose that the scores were all on a 1 to 10 scale, and the judge's linear model was the formula:

$$y = .6 \cdot x_1 + .2 \cdot x_2 + .2 \cdot x_3$$

The patient is called maladjusted if the score  $y$  is 5 or greater. The six cases could then be represented like this:

$$1. y = .6 \cdot 7 + .2 \cdot 6 = 5.4 \text{ (maladjusted)}$$

$$2. y = .6 \cdot 7 + .2 \cdot 1 + .2 \cdot 2 = 4.8 \text{ (adjusted)}$$

$$3. y = .6 \cdot 3 + .2 \cdot 8 + .2 \cdot 8 = 5.0 \text{ (maladjusted)}$$

$$4. y = .6 \cdot 4 + .2 \cdot 9 + .2 \cdot 1 = 4.4 \text{ (adjusted)}$$

$$5. y = .6 \cdot 10 = 6 \text{ (maladjusted)}$$

$$6. y = .6 \cdot 5 + .2 \cdot 6 + .2 \cdot 6 = 4.4 \text{ (adjusted)}$$

The reason for not looking at  $x_3$  in Case 1, then, is that the first two cues are sufficient to put the case above the cutoff point. Likewise,  $x_1$  alone does this in Case 5. The judge could be using a compensatory rule, based on a weighted combination of all three attributes, and yet could still appear to be using something like cutoff points applied to individual attributes.

In sum, the use of protocols here may be misleading if it leads us to think that judges are not using some sort of subjective weighing scheme much like the linear model itself. Thus the linear model may be a good description of what judges actually do, even though evidence from protocols appears initially to suggest that it is not. Of course, judges are not entirely consistent in the way they weigh various cues, and they may not be consistent in their protocols either. It is this inconsistency that makes the use of a formula attractive.

### Impression formation

Linear models can be applied even when the stimuli to be judged are not presented in the form of numbers. The *impression-formation* task invented by Asch (1946) is a good example. Asch was interested in the basic processes by which we form impressions of other people's personalities. His experimental procedure involved a very simple situation in which some of these processes could be studied. The subject was given a list of adjectives describing a particular person and was instructed to make some judgment about the person on the basis of these adjectives. For example, Asch told subjects in one group that a certain person was "intelligent, skillful, industrious, warm, determined, practical, cautious." Subjects were asked to make judgments about this person on other dimensions, such as those defined by the adjectives "generous versus ungenerous." (Try it.) Another group of subjects was given the description "intelligent, skillful, industrious, cold, determined, practical, cautious" and asked to make the same judgment. Notice that the two descriptions are identical, except for the words "warm" and "cold."

Asch (a Gestalt psychologist) found that the first group tended to make many more positive judgments than the second group. He suggested that the subjects formed impressions of a person as a whole. The parts worked together to create an overall impression, which, in turn, affected the meaning of the various parts. For example, when he asked subjects to give a synonym of the word "intelligent" in the two descriptions, subjects given the second description tended to give synonyms such as "calculating," whereas those in the first group gave synonyms such as "clever." The terms "warm" and "cold" seemed to be *central* to the whole description. They appeared to affect the meanings of the other terms (which Asch called *peripheral*).

Asch's account makes sense intuitively, but it has been questioned by many investigators (see Anderson, 1981, for a review). An alternative explanation (to Asch's) of the findings concerning synonyms (of words such as "intelligent," for example) is that the subjects found it difficult to follow the instructions to give a synonym only of the word to which they were told to attend. The word "calculating," for instance, is in fact an associate of the word "cold." Subjects might have associated to the word "cold" itself, or perhaps to both "cold" and "intelligent" in combination — but without the meaning of "intelligent" changing under the influence of the word "cold" (Anderson, 1981, p. 216).

The rest of Asch's results can be explained by a simple algebraic model, a variant of the linear regression model. Each adjective has a certain weight on each scale. For example, "warm" has a very high weight on the scale from "generous" to "ungenerous," and "cold" has a very low weight. (This is because we generally think that warm people are generous. The reason we think this is not relevant to the controversy at hand.) Anderson (1981) points out that a good account of all the results can be obtained by assuming that the subject *averages* the weights of the separate adjectives on the scale in question. If "generous" has a value of +10 and "ungenerous" has a value of -10, "warm" by itself might have a value of +6 and "cold" a value of -6. "Intelligent" might have a value of +1 by itself. A person described as warm and

intelligent would have an average of  $(6 + 1)/2$  or 3.5, which would come close to the subject's rating.

The averaging model applies to a great many judgments of the sort we have been discussing, judgments whose stimuli can be decomposed into separate attributes or features. For example, in one study (Anderson and Butzin, 1978), children were asked to play Santa Claus. Their task was to give a "fair share" of toys to other children, each of whom was described in terms of need (how many toys they already had) and their achievement (how many dishes they had washed for their mothers). The subjects were sensitive to both variables, and the number of toys allotted to each child was a linear combination of the two variables (each given a certain weight).

The averaging model is completely consistent with the linear regression model. Of course, "averaging" implies dividing by the number of cues or dimensions. As long as the number of cues is held constant, however, we do not know whether a subject is adding their weights or averaging them. The results obtained in connection with the averaging model therefore provide additional support for the regression model as a description of behavior, and vice versa.

### Averaging, adding, and number of cues

As noted, the term "averaging" implies that subjects add up the impression from each cue and then divide by the number of cues.<sup>6</sup> In the experiments described so far, the number of cues is held constant, so, for all we know, subjects might simply be *adding* the impressions from the separate cues. They might then divide (or multiply) by some number that would be the same regardless of the number of cues, merely in order to produce responses that fall within a reasonable range. To find out whether subjects are adding or averaging, we need to vary the number of cues.

The simplest way to do this is to compare the effect of one cue and two cues. When this is done, presentation of two equally positive cues usually leads to a higher rating than either term alone but less than the sum of the two ratings. It would seem that subjects do something in between adding and averaging. For example, if the personality rating (on a scale of "generosity" in which 0 represents neutrality) for "happy" alone were 3 and the rating for "friendly" alone were 3, the rating from the two together might be 4, which is higher than the average of 3 and 3 (which is 3) but less than their sum (which is 6).

To explain such findings (and others) Anderson assumes that subjects have some sort of starting point or initial impression (just from knowing that it was a *person* being described, perhaps) that had to be averaged in. Suppose that "happy" and "friendly" each has a weight of 6, and the initial impression has a value of 0. Given

<sup>6</sup>Actually, each *scale* (for example, friendliness, intelligence) is assumed to have a different *weight* for a given judgment. These weights correspond to the coefficients  $a$  through  $e$  in the regression example. The weight depends on the relevance of the scale (for example, friendliness) to the judgment being made (for example, generosity). The value of each cue (for example, the word "friendly") on its own scale (for example, friendliness) is thus multiplied by this weight before being added. Then this sum is divided by the sum of the weights themselves, rather than just by the number of cues.