3

A General Framework for Judgment

Psychology has forgotten that it is a science of organism–environment relationships, and has become a science of the organism. This is somewhat reminiscent of the position taken by those inflatedly masculine medieval theologians who granted a soul to men but denied it to women.

—Egon Brunswik

3.1 A Conceptual Framework for Judgment and Prediction

"The causes of the disaster are not due to faulty organization, but to misfortune in all risks which had to be undertaken. . . . We took risks, we knew we took them; things have come out against us, and therefore we have no cause for complaint, but bow to the will of Providence, determined still to do our best to the last." These were the last recorded words of British explorer Robert Scott, who lost the race to the South Pole and then perished from starvation and exhaustion only 11 miles from his return supply depot. Scott's eloquent message describes himself and his men as heroes defeated by the implacable, enigmatic natural world. But history has not been kind to Scott, and most commentators now attribute Scott's failure to repeated episodes of poor judgment as much as to unpredictable adverse events during his trek to and from the South Pole (Diamond, 1989; Huntford, 1999). It seems that Scott made many bad judgments, for example, about where to locate his supply base; about the endurance of his men, pack animals, and machines; and about numerous other details of his expedition.

This chapter is an introduction to the psychology of judgment, the human ability to infer, estimate, and predict the character of unknown events. Our judgment faculties are subject to certain systematic flaws, perhaps the most prominent of which is simple overconfidence.

The human mind has been designed by nature to go beyond the information given by our senses, and to go further beyond "the given" than does the nervous system of any other organism on this planet. Even the apparently effortless perception of a three-dimensional physical scene involves inferences that are mathematically impossible if based on only the information given to our retinas (Attneave, 1954; Pinker, 1997). Nonetheless, evolution has endowed us with a cognitive system that has the right assumptions built into it to do an excellent job of navigating through our three-dimensional environment without bumping into major landmarks. Our visual system is so good at making these unconscious inferences that it is impossible for us to figure out how we make them by examining our conscious experience. In some unusual cases of brain damage, a phenomenon called *blindsight* reveals that we are still able to make these judgments even when, due to damage to our primary visual cortex, we have no conscious awareness of the perceptual experience itself. This chapter is about the process of judgment, including a broad range of accomplishments, from the intuitive visual cognition involved in anticipating the path of a fly ball to the deliberate inferences of a physician trying to find out what is wrong with a patient's kidney.

For the moment, we will focus on the psychology of judgment processes where the goal of the judgment is to infer the nature of some condition that does or could exist in the external world (and ignore issues concerning judgments of internal mental events associated with evaluated consequences and personal values). Within psychology, a conceptual framework has been developed to deal with our judgments and expectations concerning events and outcomes of possible courses of action. The framework and its associated terminology may seem a little antiquated today, but the basic concepts still provide an excellent organizational scheme to summarize judgments made under *irreducible uncertainty,* meaning uncertainty that cannot be eliminated before a decision about what action to take must be made.

The framework is called the *Lens Model,* and it was invented by an Austrian-American psychologist named Egon Brunswik (Hammond &

Stewart, 2001). The model gets its name from the notion that we cannot make direct contact with the objects and events in the world outside our sense organs; we only perceive them indirectly through a "lens" of information that mediates between the external objects and our internal perceptions (Pepper, 1942). The framework is divided into two halves, one representing the psychological events inside the mind of the person making a judgment and the other representing events and relationships in the "real world" in which the person is situated. The framework forces us to recognize that a complete theory of judgment must include a representation of the environment in which the behavior occurs. We refer to it as a *framework*, because it is not a theory that describes the details of the judgment process; rather, it places the parts of the judgment situation into a conceptual template that is useful by itself and can be subjected to further theoretical analysis.

Let's take an example judgment and work our way through the conceptual diagram (Figure 3.1) for the Lens Model. Suppose we are trying to estimate the biological age of a man encountered on the street. (Judgments of the gender,

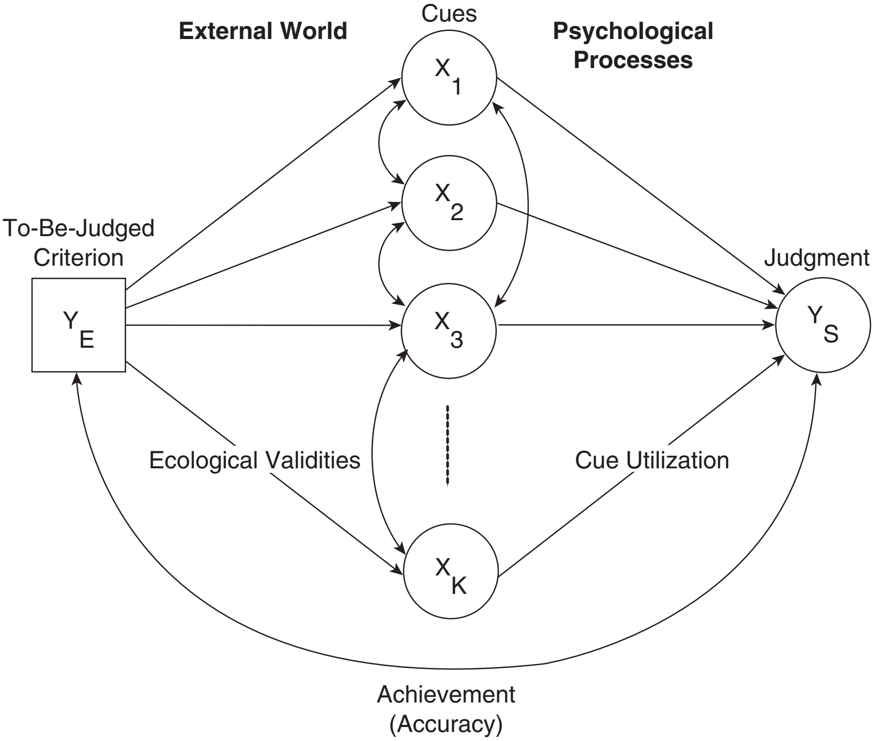

Figure 3.1 Lens Model conceptual framework for the global judgment process

age, and ethnicity of other people are usually automatic.) The Lens Model frames this judgment as a process through which we, the judges, are trying to "see" a true state of the world (the person's age) through a proximal lens of items of information called *cues* that are available to us. In the case of an age judgment, we probably observe and rely on cues such as characteristics of the man's hair (Is it gray? Is he balding?), his skin quality (Wrinkled or smooth?), his body (How fit does he appear to be? Does he exhibit the gait and posture of a youthful or an aged man?), his clothes (Is he dressed like an older person or younger?), his voice (Is it childish, adolescent, harsh, faint?), and other signals that might support inferences about his age. Note that for an intuitive judgment (like age), even the person making the judgment will not be able to provide a report of the cues he or she is relying on.

The left side of the Lens Model diagram summarizes the relationships between the true, to-be-judged state of the world, called the *criterion* (the man's age), and the cues that may point to that state of the world. In the case of the age judgments, physical anthropological studies would address the relationships on the left-hand side of the diagram: What are the true relationships between biological age and the visible cues or signs it produces? Those relationships are often conceptualized as causal—the criterion state or outcome causes the cue, or maybe the criterion state produces or moderates the cue values where the relationships are not directly causal. In the middle of the diagram is the "lens" of cues that connect the judgment to the criterion or goal of the judgment. The vertical arrows connecting the cues (X_1, X_2, . . .) represent the interdependencies or intercorrelations that usually exist between cues in most judgments. The right-hand side of the lens diagram is the psychological judgment process part of the framework. It refers to the inferences that a person makes to integrate information conveyed by the cues so as to form an estimate, prediction, or judgment of the value of the criterion. The overarching path in the diagram (labeled "achievement") represents the judge's ability to estimate the to-be-judged criterion accurately.

Using a statistical model to "capture" a person's *internal* psychological "judgment policy" (the right-hand side of the Lens Model) may seem odd to readers who are familiar with the common practice of modeling relationships between variables in the *external* world (the left-hand side of the Lens Model). To clarify the application of statistical modeling to analyze an internal psychological process, let us walk through a concrete example.

Several years ago, some students thought it would be interesting to capture one of the author's (Hastie's) judgment policy as he evaluated applicants to graduate studies in his PhD program. Every year about 125 written applications were received and he read all of them and assessed each applicant's qualifications for admission to the program. For purposes of the study, his students reviewed the contents of each application package and assigned quantitative scores to each of the 28 most obvious “cues” that Hastie might be relying on to make his judgments. Some of this information was already quantitative (e.g., the applicant’s age, test scores, and grade point average), but much of it had to be “coded” into numbers by the students. For example, “college quality” was coded on a 4-point scale based on a popular national rating service and “warmth” of the recommendation letters was rated subjectively by the student researchers (with high inter-rater agreement).

Then Hastie reviewed the applications from two years of the admissions process and made a rating on a 10-point scale of “admissibility.” The students took that matrix containing 28 items of information on 245 applications plus Hastie’s 1-10 rating of admissibility and conducted a statistical analysis to determine the best-fitting linear model to predict Hastie’s ratings from the 28 cues (see Freedman, Pisani, and Purves, 2007 or another good introduction to statistical modeling for the details of these analyses). Essentially, this approach provides a rough estimate of the average impact of the different items of information on Hastie’s judgments across the 245 cases he judged. With due caution, we can say the statistical model is a summary of his policy for making admission judgments (the right-hand side of the Lens Model). In this case the equation was:

$$\begin{aligned}\text{ADMISSIBILITY RATING} = \rightarrow &+ 0.012(\text{VERBAL GRE TEST SCORE})\\ &+ 0.015(\text{QUANTITATIVE GRE TEST SCORE})\\ &+ 0.250(\text{‘WARMTH’ OF RECOMMENDATIONS})\\ &+ 0.410(\text{COLLEGE QUALITY})\\ &-13.280\end{aligned}$$

What does this equation tell us about Hastie’s judgment habits? First, he is reliably using only four items of information—two test scores, recommendation letters, and undergraduate college quality. Second, it’s obvious he relies heavily on standardized aptitude test scores. The most remarkable result is how well the model does overall in predicting his behavior. The correlation between the model’s predictions and his actual ratings was .90. To put that correlation in perspective, Hastie made repeated judgments on 50 cases, two weeks after he made the original judgments of all 245 cases. The reliability, measured by the correlation between his first round of judgments and the second round, was .88. In other words, the model was capturing every scrap of reliable predictive habits in his admissibility ratings!

Although the model does an excellent job of describing Hastie's behavior, it does not necessarily tell us how we should conduct admissions evaluations. To do that we would need an analysis of the cue-criterion relationships in the environment, on the left-hand side of the Lens Model (see Dawes, 1971, for such an analysis of graduate admissions).

3.2 Research With the Lens Model Framework

The Lens Model was invented by psychologists for use in research, so it can be interpreted as a blueprint for a method to analyze judgment processes. (Cooksey, 1996, provides a good introduction to the methodology and reviews results from this research paradigm.) Once a judgment has been selected for study, the first step for the researcher is to identify and measure the cues on which the judge relies. This is often a laborious task requiring several rounds of measurement and testing before all of the effective cues have been discovered. Obviously, this task is especially difficult for intuitive judgment processes, where the judge can't tell the researcher what cues are relied on (by the judge) to make the judgment. Often, this situation arises in important decisions made by experts. It is often very difficult for a physician, an engineer, or a financial analyst to "unpack" his or her highly practiced, automatic judgment process and to explain "how it's done." In the case of the age judgment, we would probably start out with our own intuitions, maybe consult with other people about how they make the judgment, maybe do a little research in the anthropometric literature on actuarial facts about human aging (a good first guess is that a human judge will adaptively use the scientifically correct cues to make any judgment), and come up with an initial set of candidate cues. Then we would conduct a study of the age judgment, and keep open the possibility that the initial cue set might need to be enhanced to include additional cues that are used by people to make the judgment.

The second step in the analysis is the creation of a model of the events on the left side of the diagram. Often, a linear regression model can be used to summarize the criterion–cue relationships in terms of the many correlations between the criterion and each of the cues that are related to it and might be used by a judge to infer the criterion (see a good elementary statistics text for an introduction to linear equations, e.g., Freedman, Pisani, Purves, & Adhikari, 1991, or Anderson, 2001). In this analysis, the correlation coefficient (or a related statistic) is used to summarize the strength of the relation between the criterion and a cue (the *ecological validity* of the cue) and between the cue and the judgment (the *cue utilization coefficient* or, more informally, the psychological impact of the cue on the judgment). Sometimes the modeler recognizes that the linear model is a simplified abbreviation of those "external environment" dynamics, although in many domains, linear equations provide a surprisingly complete summary of the environment. Our experienced world is dominated by approximately linear relationships.

The third step in research shifts over to the right-hand side of the diagram and involves inventing and testing models of the psychological process of cue utilization: How do people use the cues to make inferences about the criterion state? Here again, researchers have often found the linear statistical model to be a good description. The usual research tactic is to collect a sample of to-be-judged stimuli—for example, a sample of videotapes of men of various ages to present to an experimental subject for judgments of the age of each stimulus person. The judge's cue utilization habits are "captured" in an algebraic equation that relates the judgment to a weighted sum of the cue values. (Note that this analysis depends on the researcher's ability to measure the cue values on psychologically meaningful numerical scales.) Here the research literature is clear; the most general principle to describe cue utilization processes is the linear equation. For an amazing range of everyday and expert judgments, people seem to infer the implications of cue information as if it is measured on numerical scales, weight it, and add it up.

Imagine sitting in a doctor's office watching her diagnose patients. Each patient comes in, has an interview with the doctor, provides the history of a medical problem, and describes some symptoms. Usually, laboratory tests are made, and maybe some X-rays (or other "scans") are taken. Then, after reviewing all this material, the doctor makes a diagnostic decision about what is wrong with the patient. Consider recording these events for a few weeks to have a good sample of the cues (patient's history, symptoms, and test results) and diagnoses for this judgment task. Or transfer the same scenario to a busy college admissions office. Consider admissions officers reading applications—reviewing *objective* measures of achievement, like test scores and high school grades, and more *subjective* material such as letters of reference, lists of extracurricular activities, and a personal essay—and then making judgments about the admissibility of many applicants. Again, you observe until you have a sizable sample of cases (cues) and judgments.

The Lens Model approach analyzes the judgment by calculating an algebraic model to provide a summary of the weights placed on the cue values for each case so as to predict the judge's (physician's, admissions officer's) judgments. The weights are based on the correlation coefficients summarizing the linear dependency of the judgment of each cue; with everything else equal, the higher the correlation is, the greater the weight will be. The model can be extended to include nonlinear relationships (e.g., a U-shaped functional form with high judgments associated with extreme values on the cue dimension—for example, where both extremely thin and obese patients are at high risk of injury, while those of average body weight are at low risk; or perhaps an admissions officer who likes applicants who either participate in many extracurricular activities or have specialized in one activity, but does not like "average," 2–3 activity participators). The model can also represent *configural* relationships where the judgment depends on combinations of cues (e.g., high levels of a particular hormone in the blood are bad news for female patients, but uninformative for male patients; see discussion below of "interaction effects" in intervariable relationships). But again, the simple linear model is surprisingly successful in many applications. We say "surprisingly," because many judges claim that their mental processes are much more complex than the linear summary equation would suggest—although empirically, the equation does a remarkably good job of "capturing" their judgment habits.

If we had criterion values for our sample of judgments, we could also calculate a summary model for the left-hand side of the Lens Model diagram. In many applications to actual judgment tasks, however, it is difficult to obtain criterion values. In medical contexts, it is too time-consuming for a physician to track the history of patients to obtain final opinions about their presenting condition or outcomes of treatment; similarly, in the academic context, we have no access to values representing success in a college for students who were not admitted. But we are often interested in the psychology of the judgments, the right side of the lens diagram, not the complete environment-behavior system encompassed by the full framework.

Hundreds of studies have been conducted of judgments ranging from medical diagnosis to highway safety, from financial stock values to livestock quality (Brehmer & Joyce, 1988). There is great variety in patterns of results across judgment domains (i.e., weather forecasting is different from internal medicine, which is different from college admissions, which is different from livestock pricing) and across judges. (There are big individual differences in the weights placed on different types of informational cues—and there are some, but only a few, truly remarkably expert judges, while there are many so-called experts who are no better than complete novices; see, for example, Sherden, 1998.) At the risk of overgeneralization, here are some conclusions about typical judgment habits that are true of both amateur and expert judgment:

1. Judges (even experts) tend to rely on relatively few cues (3–5). There are some exceptions to this generalization, for example, in very expert judgments of weather conditions and livestock quality. We believe that judgments are sensitive to more cue information in these exceptional domains because training for judgment involves immediate, precise feedback to the people learning to make the judgments (unlike, for example, training in medical diagnosis,

admissions decisions, or financial forecasting, where feedback is usually delayed and often never available to the person learning to judge).

2. Few judgment policies exhibit nonlinearity; most are additive and linear—again, contrary to many judges' own beliefs about their judgment processes.
3. Judges lack insight into their policies—they are unable to estimate their own relative "cue utilization weights" accurately—especially when they are expert and highly experienced.
4. Many studies (e.g., students' judgments of physical attractiveness, professors' graduate school admissions judgments, radiologists' judgments of tumor malignancy) reveal large individual differences in types of policies (patterns of cue utilization weights) across judges and low interjudge agreement on the judgments themselves. In important domains like medical diagnosis, this conclusion is disturbing, because we would like our medical experts to agree with one another (and with biological theory) when they make diagnoses and prescriptions. At a minimum, interjudge disagreements tell us someone is wrong, and undermine our confidence in all judgments.
5. When associated, but non-diagnostic, irrelevant information is presented to judges, they become more confident in the accuracy of their judgments, although true accuracy does not increase.

The picture of the expert, painted in broad brushstrokes by this research, is unflattering. However, the important message is that before we draw any conclusions about a judge's performance (whether it is the automatic acceptance of claims of wisdom and accuracy or the blanket assumption that all judges are inept), we need to take a careful look at that performance—and we should be prepared for surprises. Vaunted experts with extensive credentials and impressive demeanors may be no better than college sophomores at their specialty judgments, but then there are some true experts who are really worth heeding or hiring.

3.3 Capturing Judgment in Statistical Models

Historically, some of the earliest psychological research on judgment addressed the question of whether trained experts' predictions were better than statistically derived, weighted averages of the relevant predictors. Employing multiple regression analyses within the Lens Model framework in Figure 3.1, we can ask the following question: Which is better, a linear statistical model summarizing the left-hand side of the Lens Model diagram or the human judgment on the right-hand side of the diagram? This question has been studied extensively by psychologists and other behavioral scientists interested in predicting outcomes such as college success, parole violation, psychiatric diagnosis, medical diagnosis, investment values, and business success and failure. In the early studies, the information on which clinical experts based their predictions was the same as that used to construct linear models. Typically, this information consisted of test scores or biographical facts, but some studies included observer ratings of specific attributes as well. All of these variables could easily be represented by (coded as) numbers having positive or negative relationships to the criterion outcome to be predicted. (Higher test scores and grade point averages predict better performance in subsequent academic work; a higher leukocyte count predicts greater severity of Hodgkin's disease; more gray hair and more wrinkles predict more biological years, etc.)

In 1954, Paul Meehl published a highly influential book in which he reviewed approximately 20 such studies comparing the clinical judgments of people (expert psychologists and psychiatrists in his study) with the linear statistical model based only on relationships in the empirical data on the events of interest (the left side of the Lens Model). *In all studies evaluated, the statistical method provided more accurate predictions (or the two methods tied).* Approximately 10 years later (1966), Jack Sawyer reviewed 45 studies comparing clinical and statistical prediction. Again, there was *not a single study* in which clinical global judgment was superior to the statistical prediction (termed "mechanical combination" by Sawyer). Unlike Meehl, Sawyer did not limit his review to studies in which the clinical judge's information was identical to that on which the statistical prediction was based; he even included two studies in which the clinical judge had access to *more* information (an interview with each person being judged) but still did *worse.* (In one of these, the performance of 37,500 sailors in World War II in U.S. Navy basic training was better predicted from past grades or test scores, alone or in combination, than from the ratings of judges who both interviewed the sailors and had access to the test and grade information used in the model.)

The near-total lack of validity of the *unstructured* interview as a predictive technique had been documented and discussed by E. Lowell Kelly in 1954 (see, more recently, Hunter & Hunter, 1984, and Wiesner & Cronshaw, 1988). There is no evidence that such interviews yield important information beyond that of past behavior—except whether the interviewer likes the interviewee, which is important in some contexts. (Some of our students maintain it is necessary to interview people to avoid admitting "nerds" to graduate study, but they cannot explain how they would spot one, or even what they mean by the term.)

A representative study of psychodiagnosis was reported by Lewis Goldberg (1968), a professor of psychology who was influential in the early history of the use of linear models to analyze judgment. Goldberg asked experienced clinical diagnosticians to distinguish between neurosis and psychosis on the basis of personality test scores (a decision that has important implications for treatment and for insurance coverage in psychotherapeutic practice). He constructed a simple linear decision rule (add the patient's scores on three scales together and subtract the scores on two other scales; if the result exceeds "45," diagnose the patient as psychotic). Starting with a new sample of patient cases and using the patients' discharge diagnoses as the to-be-predicted criterion value, "Goldberg's rule" achieved an accuracy rate of approximately 70%. The human judges, in comparison, performed at rates from slightly above chance (50%) to 67% correct. Not even the best human judge was better than the mechanical adding-and-subtracting rule.

Another study of clinical versus statistical prediction was conducted by Hillel Einhorn (1972). He studied predictions of the longevity of patients with Hodgkin's disease during an era when the disease was invariably fatal (prior to the 1970s). (Einhorn had a personal interest in the subject matter as he had just been diagnosed with the condition, which eventually took his life in 1987.) A world expert on Hodgkin's disease and two assistants rated nine characteristics of biopsies (cues) taken from patients and then made a global rating of the "overall severity" of the disease process for each patient. Upon the patients' deaths, Einhorn correlated the global ratings with their longevity. While a rating of overall severity is not precisely the same as a prediction of time until death, it should predict that. (At least, the world expert thought it would.) Einhorn found that it does not. In fact, the slight trend was in the wrong direction: higher severity ratings were associated with longer survival time. In contrast, a multiple regression analysis, based on the nine biopsy characteristics scaled by the doctors, was statistically reliable and significantly more accurate than the physicians' severity ratings.

Another striking example comes from a study by Robert Libby (1976). He asked 43 bank loan officers (some senior, in banks with assets up to $4 billion) to predict which 30 of 60 firms would go bankrupt within 3 years of a financial report. The loan officers requested and were provided with various financial ratios (cues)—for example, the ratio of liquid assets to total assets—in order to make their predictions. Their individual judgments were 75% correct, but a regression analysis based on the financial ratios themselves was 82% accurate. In fact, the ratio of assets to liabilities *alone* predicted 80% correctly.

The practical lesson from these studies is that in many judgment situations, we should ask the experts what cues to use, but let a mechanical model combine the information from those cues to make the judgment. The finding that linear combination is superior to global judgment is general; it has been replicated in diverse contexts. Not in psychology, but in some medical and business contexts, global judgment has been found to be superior; in those particular contexts, the people making the global judgments had access to "inside information" not available to the statistical model. A fair comparison would insist that both human experts and the models would have identical information. In at least one context, once this extra information was included in the statistical model, its predictions again became superior (in predicting 24-hour survival on an intensive care unit; see Knaus & Wagner, 1989). Meehl updated his classic review several times, and in 1996, he and a colleague concluded the following: "Empirical comparisons of the accuracy of the two methods (136 studies over a wide range of predictions) show that the mechanical method is almost invariably equal to or superior to the clinical method" (Grove & Meehl, 1996, p. 293).

3.4 How Do Statistical Models Beat Human Judgment?

Why is it that linear models predict better than clinical experts? We can explain this finding by hypothesizing a mathematical principle, a principle of "nature," and a psychological principle.

The mathematical principle is that both monotone relationships of individual variables and monotone ("ordinal") interactions are well approximated by linear models. Such interactions are illustrated in Figure 3.2. Two factors "interact" when their combined impact is greater than the sum of their separate impacts, but they do not interact in the sense that the *direction* in which one variable is related to the outcome is dependent upon the magnitude of the other variable. It is not, for example, true of monotone interactions that high-highs are similar to low-lows, but that high-highs (or low-lows) are much higher (or lower) than would be predicted by a separate analysis of each variable. If high-highs are similar to low-lows, the interaction is termed *crossed,* illustrated in Figure 3.2.

For example, a doctoral student of Dawes (Glass, 1967) subjected alcoholic and nonalcoholic prisoners to a benign or a stressful experience. He then had them spend 20 minutes in a waiting room before being interviewed by a psychologist about their experience. A nonalcoholic punch was available in the waiting room, and the behavior of interest was how much punch the prisoners consumed. The alcoholic and nonalcoholic prisoners drank virtually identical amounts after experiencing the benign situation. After the stressful situation, however, the alcoholic prisoners drank twice as much punch as the nonalcoholics did (see the middle two panels in Figure 3.2). Thus, a true

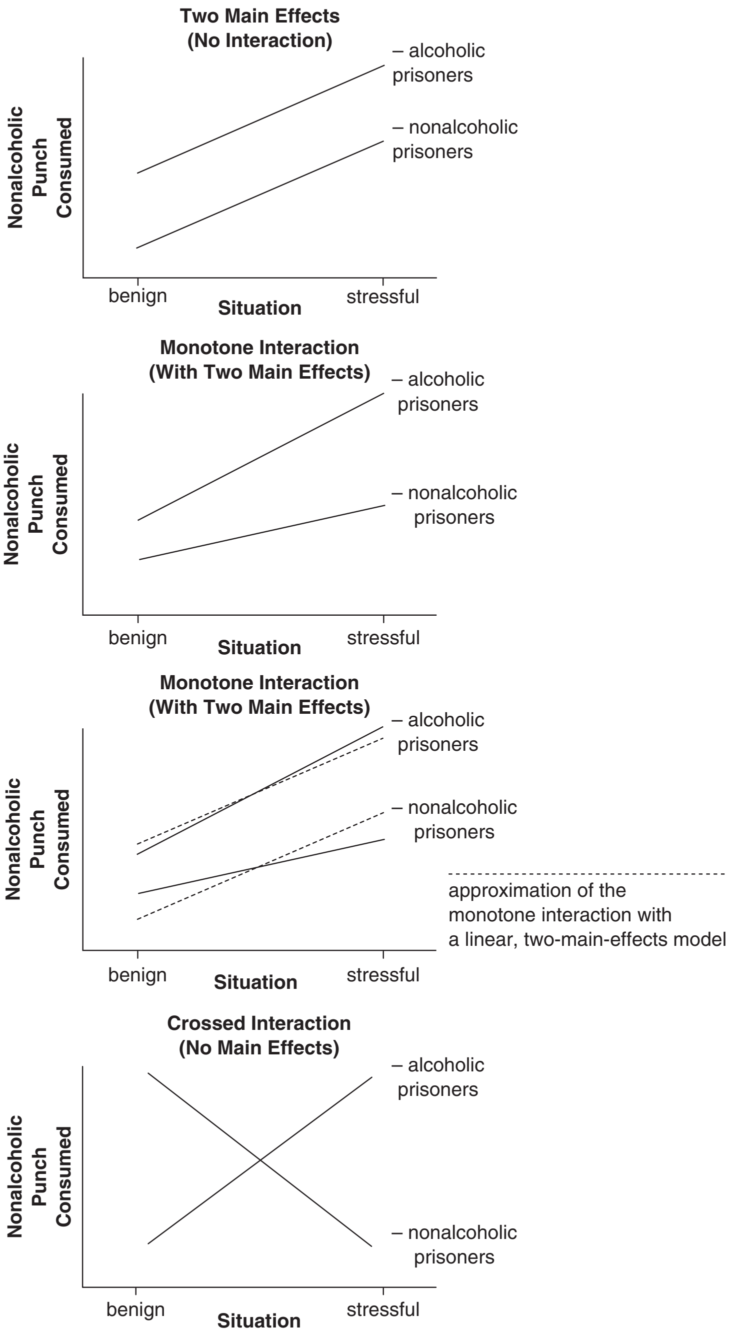

Figure 3.2 Examples of crossed and non-crossed (monotone) interaction effects

"monotone" interaction was found between stress and drinking behavior of diagnosed alcoholics: the amount of punch consumed could not be predicted by considering each factor independently; in the example, a distinctive prediction is made for the doubly potent alcoholism *plus* stress combination. However, the statistical analysis indicated that this interaction could be well approximated by the two independent main effects: One, alcoholics drank more punch, and two, all prisoners drank more punch after being stressed. A situation in which only main effects are present is truly linear.

To clarify our *mathematical principle*, consider the top panel in Figure 3.2; this depicts a pure main effects situation in which the two variables have simple, independent effects: Alcoholics drink more (no matter what), and prisoners in a stressful situation drink more (no matter what). A linear, weight-and-add model would fit these data perfectly. The bottom panel depicts the most complicated situation where we imagine a crossover interaction. In benign situations, alcoholics drink the least punch, but the pattern reverses in the stressful situation where alcoholics drink the most punch. No linear model can capture this pattern of effects, even approximately. However, true crossover patterns of causal relationships are very, very rare. And, as we just noted, the non-crossover relationships (which are much more prevalent) can be well approximated by linear relationships. (See the dotted lines in the lower "monotone interaction" panel of Figure 3.2. Also see any good introduction to statistics and data analysis for an exposition of the nature of interaction effects; e.g., Norman Anderson, 2001, is excellent, and Robert Abelson's insightful book, *Statistics as Principled Argument* [1995], contains an especially wise discussion of interactions and their interpretation in behavioral research.)

The *principle of nature* that partly explains the success of the linear statistical model is that most interactions that exist are, in fact, monotone. It is easy to hypothesize crossed interactions, but extraordinarily difficult to find them in everyday situations, especially in the areas of psychology and social interactions. Because the optimal amount of any variable does not usually depend upon the values of the others, what interactions there are tend to be monotone. Moreover, while a number of crossed interactions have been hypothesized in social interactions (e.g., authoritarian leadership is more effective in some types of situations, while libertarian leadership works better in others), they tend to be supported only by verbal claims and selective post hoc data analysis. In fact, interactions of *any* sort tend to be ephemeral, as was discovered by Goldberg (1972) in his analysis of how the "match" between teaching style and student characteristics predicts student success. Of 38 interactions he thought he had discovered in the first half of an extensive data set, only 24 "cross-validated" *in the* *right direction* in the second half (not significantly different from chance expectation of 19 cross-validations).

The *psychological principle* that might explain the predictive success of linear models is that people have a great deal of difficulty in attending to two or more noncomparable aspects of a stimulus or situation at once. (“Separable” and “incommensurate” are other technical labels for this relationship between stimulus dimensions.) Attention shifts from one cue to another and back again. For example, when Roger Shepard (1964) asked subjects to make similarity judgments between circles containing “spokes” at various angles (the stimuli looked like one-handed clock faces), the subjects attended to size of the circles *or* to angles of the spokes, but *not to both*. The experience of people evaluating academic applicants is similar. Often they anchor their judgment on a salient cue, such as a distinctively high or low grade point average or test score, and then adjust in light of less distinctive information in the applicant’s folder. Sometimes the format of the information will determine the salient anchor value, as when a bias is introduced by placing one type of information (e.g., test scores) in a prominent location, such as first in a list of applicant information. Other people consistently start by attending to one cue, for example, a favored test score, then to a second priority cue (perhaps grade point average [GPA]), and then to tertiary information that they believe is less important. But notice that although the rough-and-ready, anchor-and-adjust judgment strategy provides for cognitively efficient integration of a considerable amount of information in a manner analogous to a linear statistical model, it is not optimal. In reality, how *could* an admissions committee member rationally integrate test information and GPA information without knowing something about the distribution and predictability of each student within the applicant pool? The need for such comparisons is one reason that a purely statistical integration will be superior to a global judgment. The statistical model will use valid, independent information from as many cues as convey such information, will be “calibrated” to the ranges of values on all the variables available in the situation, and will do so relentlessly and consistently.

Given that monotone interactions can be well approximated by linear models (a statistical fact), it follows that because most interactions that exist in nature are monotone *and* because people have difficulty integrating information from noncomparable dimensions, linear models will outperform clinical judgment. The only way to avoid this broad conclusion is to claim that training makes experts superior to other people at integrating information (as opposed, for example, to knowing what information to look at). But there is no evidence that experts *think differently* from others. (Remember the example of chess grandmasters from Chapter 1:

Grandmasters did *not* possess special visual or intellectual skills, but they knew much more than novices about "where to look," and they had much more knowledge in long-term memory about specific chess board positions and what to do in each situation.)

A further, more speculative conjecture is that not only is the experienced world fairly linear, but our judgment habits are also adaptively linear. So, the linear models, which are so popular to describe the right-hand, cue utilization side of the Lens Model diagram, convey a correct image of the human mind (see, for example, Anderson, 1996; Brehmer & Joyce, 1988). The mind is in many essential respects a linear weighting and adding device. In fact, much of what we know about the neural networks in the physical brain suggests that a natural computation for such a "machine" is weighting and adding, exactly the fundamental processes that are well described by linear equations. We explore some of the nuances of this very general judgment habit in the next chapter.

3.5 Practical Implications of the Surprising Success of the Linear Model

There is an enormous and almost unequivocal research literature that implies expert judgments are rarely impressively accurate and virtually never better than a mechanical judgment rule. As Meehl (1986) put it, 40 years after his "disturbing little book" was published, "There is no controversy in social science which shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction as this one" (p. 373). The implication for practice seems clear: Whenever possible, human judges should be replaced by simple linear models. We put in the "whenever possible" qualification only because we also believe that some empirical tests should be done before any important decision is made in a new way. We do not advocate simply replacing all human judges without considering the specific circumstances of each judgment situation. There will always be special cases and changes in the nature of the task environment (perhaps a new diagnostic method is invented) that require oversight and adjustment. We do believe, however, that a substantial amount of time and other resources is squandered on expert judgments that could be made more equitably, more efficiently, and more accurately by the statistical models we humans construct than by we humans alone.

We advocate the broader use of actuarial, mechanical prediction methods. Research by one of the authors (Dawes, 1979) shows that it is not even necessary to use statistically optimal weights in linear models for them to outperform experts. For years, the nagging thought kept recurring: Maybe *any* linear model outperforms the experts. The possibility seemed absurd, but when a research assistant had some free time, Dawes asked him to go to several data sources and to construct linear models with weights "determined randomly except for sign." (It seemed reasonable that in any prediction context of interest, the direction in which each cue predicted the criterion would be known in advance.) After the first 100 such models outperformed human judges, Dawes constructed 20,000 such "random linear models"—10,000 by choosing coefficients at random from a normal distribution, and 10,000 by choosing coefficients at random from a rectangular distribution. Dawes used three data sets: (1) final diagnoses of neurosis versus psychosis of roughly 860 psychiatric inpatients, predicted from scores on the Minnesota Multiphasic Personality Inventory (the same set used by Goldberg in constructing his "add three, subtract two" rule); (2) first-year graduate school grade point averages of psychology students at the University of Illinois, predicted from 10 variables assessing academic aptitude prior to admission and personality characteristics assessed shortly thereafter; and (3) faculty ratings of performance of graduate students who had been at the University of Oregon for 2 to 5 years, predicted from undergraduate grade point averages, Graduate Record Examination (GRE) scores, and a measure of the selectivity of their undergraduate institutions. All three predictions had been made both by linear models and by human experts ranging from graduate students to eminent clinical psychologists. On the average, the random linear models accounted for 150% more variance between criteria and predictions than did the intuitive clinical evaluations of the trained judges. For mathematical reasons, *unit weighting* (that is, each variable is standardized and weighted +1 or −1 depending on direction) provided even better accountability, averaging 261% more variance. Unit or random linear models are termed *improper* because their coefficients (weights) are not based on statistical techniques that optimize prediction. The research indicates that such improper models are almost as good as proper ones. When it comes to the coefficients in a linear model, the signs on the coefficients are much more important than the specific numerical weights.

We would also point out that human judges relying on intuition are not very competent about adjusting for differences in the metrics of the scales that convey numerical information. If one type of information (e.g., test scores) is conveyed by numbers that range from 200 to 800 and another type (e.g., grades) is conveyed by numbers that range from 1 to 4, the human brain will be fooled into greater judgment adjustments based on the "larger quantities" on the first scale. The implication is that, when intuitive judgments are made, it's good practice to standardize the cue information scales. Another effective, though also "improper," approach is to fit a linear model to a large sample of a human judge's own judgments and then to use that model-of-the-judge instead of the original judge. This method is called *bootstrapping* (not to be confused with the "statistical bootstrap" introduced by Efron, 1988), and it almost invariably outperforms human experts, including the person who was used as the source of judgments for the original model. Again, there are several interpretations of the success of bootstrap models, including their reliability, imperturbability (the equations are not susceptible to bad moods or fatigue), and the fact that the abstracted judgment policy may better represent the human judge's true understanding of the process than either subjective reports or case-by-case explanations. But most of the success can probably be attributed to the remarkable robustness and power of (even improper) linear models that derive from their mathematical properties and their match to the underlying structure of the events in the to-be-judged environment.

3.6 Objections and Rebuttals

The conclusion that random or unit or "bootstrapped" weights outperform global judgments of trained experts is not a popular one with experts, or with people relying on them. All of these findings have had almost no effect on the *practice* of expert judgment. Meehl was elected president of the American Psychological Association at a young age, but the practical implications of his work were ignored by his fellow psychologists. States license psychologists, physicians, and psychiatrists to make (lucrative) global judgments of the form "It is my opinion that . . . ," in other words, to make judgments inferior to those that could be made by a layperson with a programmable calculator. People have transferred their confidence in their own global judgments to the global judgments of "experts," a confidence that is strong enough to dismiss an impressive body of research findings and to dominate predictions in our legal and medical systems.

There are many reasons for the resistance to actuarial, statistical judgment models. First of all, they are an affront to the narcissism (and a threat to the income) of many experts. One common defense of expert judgment is to challenge the expertise of the experts making the global predictions in the particular studies. "Minnesota clinicians!" snorted a professor of psychology at the University of Michigan. Little did he know that most of the Minnesota clinicians in the study had obtained their PhDs at Michigan. "Had you used Dr. X," the dean of a prestigious medical school informed one of us, "his judgments would have correlated with longevity." In fact, "Dr. X" was the subject of Einhorn's study of Hodgkin's disease predictions.

Another objection is to maintain that the outcomes better predicted by linear models are all short-term and trivial (like dying, ending up in jail, or flunking out of school?). The claim is made that "truly important long-term outcomes" can be predicted better by global judgments. But as Jay Russo (personal communication) points out, this objection implies that the long-term future can be predicted better than the short-term future. Such prediction is possible for variables like death (as we'll all be dead 100 years from now) and rabies (after the incubation period), but those variables, which are very rare, are *not* of the type predicted in these studies. Moreover, as we come to understand processes (e.g., the existence of the rabies or the AIDS virus in the blood), "incubation period" becomes nothing more than a figure of speech, and longevity is more readily predicted than death.

A final objection is the one that says, "10,000 Frenchmen can't be wrong." Experts have been revered—and well paid—for years for their "It is my opinion that . . ." judgments. As James March has stated, however, such reverence may serve a *purely social function.* People and organizations have to make decisions, often between alternatives that are almost equally good or bad. What better way to justify such decisions than to consult an expert, and the more money he or she charges, the better. "We paid for the best possible medical advice," can be a palliative for a fatal operation (or a losing legal defense), just as throwing the *I Ching* can relieve someone from regretting a bad marriage or a bad career choice. An expert who constructs a linear model is not as impressive as one who gives advice in a "burst" of intuition derived from "years of experience." (One highly paid business expert we know constructs linear models in secret.) So we value the global judgment of experts independently of its validity.

But there is also a situational reason for doubting the inferiority of global, intuitive judgment. It has to do with the biased availability of feedback. When we construct a linear model in a prediction situation, we know exactly how poorly it predicts. In contrast, our feedback about our own intuitive judgments is flawed. Not only do we selectively remember our successes, we often have *no knowledge* of our failures—and any knowledge we do have may serve to "explain" them (away). Who knows what happens to rejected graduate school applicants? Professors have access only to accepted ones, and if the professors are doing a good job, the accepted ones will likewise do well—reinforcing the impression of the professors' good judgment. What happens to people misdiagnosed as "psychotic"? If they are lucky, they will disappear from the sight of the authorities diagnosing them; if not, they are likely to be placed in an environment where they may soon *become* psychotic. Finally, therapy patients who commit suicide were too sick to begin with—as is easily supported by an ex post perusal of their files.

The feedback problem is illustrated by the opening example presented in Malcolm Gladwell's best seller *Blink: The Power of Thinking Without Thinking* (2005). Gladwell relates the story of the Getty Museum's acquisition of a classic marble statue of a young male nude from 4th-century BCE Greece, known as a *kouros*. The provenance of the statue was uncertain, so the museum hired an expert to perform scientific tests to determine if the composition of the stone and its surface was consistent with similar authentic kouroi. The expert was satisfied, and the museum went ahead with the purchase. However, when it was placed on display, several art historians had negative gut reactions when they first glimpsed the statue. Angelos Delivorrias, director of a renowned museum in Athens, said he felt a wave of "intuitive revulsion." Thomas Hoving (1996), perhaps the most famous museum director in the world, immediately felt the statue looked too "fresh," and commented, "I had dug in Sicily, where we found bits and pieces of these things. They just don't come out looking like that" (p. 315). (But note that the conclusion that the statue was a carefully constructed forgery is still controversial; Goulandris Foundation & J. Paul Getty Museum, 1993.)

What can we conclude from this apparent triumph of intuitive judgment over systematic analysis? First, it is likely that this was truly a case where chemistry was not the best way to detect fakery. If it is a fake, the forgers did their homework when selecting marble materials and "aging" the statue's surfaces. But without a prospective study (like the ones conducted to assess the linear models), we don't know how often the experts whose intuitions were right in this one instance would be right on a representative sample of fakes. How often had they been fooled in the past? We don't even know how many other experts' intuitions were wrong for this particular statue. If 36 people have an intuitive feeling that the next roll of the dice will be snake eyes and are willing to bet even odds on that hunch, on the average 1 will win. That person is the one most likely to come to our attention; for one thing, the other 35 probably won't talk about it much.

Another instructive example is provided by a "Dear Abby" letter published in 1975:

> DEAR ABBY: While standing in a checkout line in a high-grade grocery store, I saw a woman directly in front of me frantically rummaging around in her purse, looking embarrassed. It seems her groceries had already been checked, and she was a dollar short. I felt sorry for her, so I handed her a dollar. She was very grateful, and insisted on writing my name and address on a loose piece of paper. She stuck it in her purse and said, "I promise I'll mail you a dollar tomorrow." Well, that was three weeks ago, and I still haven't heard from her! Abby, I think I'm a fairly good judge of character,

> and I just didn't peg her as the kind that would beat me out of a dollar. The small amount of money isn't important, but what it did to my faith in people is. I'd like your opinion.
>
> —SHY ONE BUCK

Note that Shy One Buck did not lose faith in her ability to predict future behavior on the basis of almost no information whatsoever; she lost her faith in people. Shy One Buck still believes she is a "good judge of character." It is just that other people are no damn good.

Hillel Einhorn and Robin Hogarth (1978) examined availability of post-judgment information and demonstrated how feedback *systematically* operates to make intuitive judgment appear valid. Consider the example of a waiter who decides he can judge whether people tip well from the way they dress. A judgment that some people are poor tippers leads to inferior service, which in turn leads to poor tips—thereby "validating" the waiter's judgment. (Not all prophecies are self-fulfilling—there must be a mechanism, and intuitive judgment often provides one. Intuition is also a possible mechanism for some *self-negating* prophecies, such as the feeling that one is invulnerable no matter how many risks one takes while driving.)

In contrast, the systematic predictions of linear models yield data on just how poorly they predict. For example, in Einhorn's (1972) study, only 18% of the variance in longevity of Hodgkin's disease patients is predicted by the best linear model (see Section 3.3 of this chapter), but that is in comparison to 0% for the world's foremost authority. Such results bring us abruptly to an unsettling conclusion: A lot of outcomes about which we care deeply are not very predictable. For example, it is not comforting to members of a graduate school admissions committee to know that only 23% of the variance in later faculty ratings of a student can be predicted by a unit weighting of the student's undergraduate GPA, his or her GRE score, and a measure of the student's undergraduate institution selectivity—but that is in comparison to 4% based on those committee members' global ratings of the applicant. We *want* to predict outcomes that are important to us. It is only rational to conclude that if one method (a linear model) does not predict well, something else may do better. What is not rational—in fact, it's irrational—is to conclude that this "something else" necessarily exists and, in the absence of any positive supporting evidence, that it's intuitive global judgment.

One important lesson of the many studies of human judgment is that outcomes are not all that predictable; there is a great deal of "irreducible uncertainty" in the external world, on the left-hand side of the Lens Model diagram (Figure 3.1). Academic success, for example, is influenced by whom one shares an office with as a graduate student, by which professors happen to have positions available for research assistants, by the relative strengths of those with whom one competes for a first job (as judged by the professors who happen to be appointed to the "search committee"), and so on (Bandura, 1982). Moreover, there are clearly self-amplifying features to an academic career. A "little bit of luck" may lead a new PhD to obtain a position in an outstanding university (or an MD in an outstanding hospital or a JD in an outstanding law firm), and the consequent quality of colleagues may then significantly reinforce whatever talents the individual brings to the job. (Conversely, a little bit of bad luck may saddle the new PhD with a nine-course teaching load, inadequate institutional resources for scholarly productivity, and "burnt out" colleagues. Not many people move from a patent office to a full professorship after publishing a three-page paper, as Albert Einstein did.)

People find linear models of judgment particularly distasteful in assessing other people. Is it important, for example, to interview students applying for graduate school? In a word, "No." What can an interviewer learn in a half-hour that is not present in the applicant's lengthy past record? As Len Rorer (personal communication to Dawes) points out, belief that one's own interviewing skills provide access to such information is grandiose overconfidence. Moreover, even if the interviewer thinks he or she has picked up some highly positive or negative quality in the interview, is it really fair to judge applicants *on the impression they make in a single interview conducted by one interviewer,* as opposed to a record of actual accomplishment (or failure) over a 4-year college career? A GPA is a "mere number," but it represents the combined opinions of some 50 or so professors over several years; some professors may be biased for or against particular students, but surely a combined impression based on actual work over time is fairer than one based on a brief interaction with a single person (who has biases and unreliabilities, too). Furthermore, GPAs predict better than interviews: Is it fair to judge someone on the basis of an impression that does not work?

A colleague in medical decision making tells of an investigation he was asked to make by the dean of a large and prestigious medical school to try to determine why it was unsuccessful in recruiting female students. The decision-making researcher studied the problem statistically "from the outside" and identified a major source of the problem: One of the older professors had cut back on his practice to devote time to interviewing applicants to the school. He assessed such characteristics as "emotional maturity," "seriousness of interest in medicine," and "neuroticism." Whenever he interviewed an unmarried female applicant, he tended to conclude that she was "immature." When he interviewed a married one, he tended to conclude that she was “not sufficiently interested in medicine,” and when he interviewed a divorced one, he tended to conclude that she was “neurotic.” Not many women received positive evaluations from this interviewer, although *of course* his judgments had *nothing* to do with gender (sarcasm intended).

3.7 The Role of Judgment in Choices and Decisions

We have restricted our focus in this chapter to the judgment of events and outcomes, but the implications also apply to the larger framework of decision and choice between alternate courses of action. Linear models often provide a valid description of the psychological processes of judgment and they are pretty good rough-and-ready statistical tools to predict events in the external world. But, they also provide an effective method to predict our own evaluations and preferences, events in the “internal,” subjective world. In a very real sense, making decisions requires us to predict what we will like in the future, often under conditions quite different from those at the time we must decide. Given that linear models predict better than intuitive judgment in situations where the accuracy of prediction can be checked, why not in situations where there is no clear criterion for truth as well? If we wish to make choices involving multiple factors, we would do well to construct *our own* (improper) linear models. This is, in essence, what Benjamin Franklin advised (discussed more fully in Chapter 10). His advice was to consider a course of action, to list the pros and cons, to weight them by apparent importance, and then to decide by adding up the weighted pros and cons to see which action had the highest total.

Thus, for practical advice about choosing, we rely on the robust beauty of even improper linear models. The philosophy presented in this chapter is based on the premise that “mere numbers” are in fact neither good nor bad. Just as numbers can be used to achieve either constructive or destructive goals in other contexts, they can be used for good or ill in decision making. Using them, however, requires us to overcome a view (*not* supported by the research) that the “mysteries of the human mind” allow us to reach superior conclusions without relying on deliberate, controlled thought processes. The mysteries are there, but not in this context. We are, all of us, overconfident in our abilities to judge. To do well by ourselves and to treat other persons fairly, we must overcome the attitude that leads us to reject adding numbers to make judgments, and we must experience no more shame when we do so than when we use numbers in determining how to construct a bridge that will not collapse.

References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum.

Anderson, N. H. (1996). *A functional theory of cognition.* Mahwah, NJ: Lawrence Erlbaum.

Anderson, N. H. (2001). *Empirical direction in design and analysis.* Mahwah, NJ: Lawrence Erlbaum.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61,* 183–193.

Bandura, A. (1982). The psychology of chance encounters and life paths. *American Psychologist, 37*(7), 747–755.

Brehmer, B., & Joyce, C. R. B. (1988). *Human judgment: The SJT view.* Amsterdam: North-Holland.

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications.* San Diego: Academic Press.

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

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571–582.

Diamond, J. (1989, April). The price of human folly. *Discover,* 73–77.

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? *Psychological Bulletin, 104,* 293–296.

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

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

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (4th ed.). New York: Norton.

Gladwell, M. (2005). *Blink: The power of thinking without thinking.* New York: Little, Brown.

Glass, L. B. (1967). *The generality of oral consumatory behavior of alcoholics under stress.* Unpublished doctoral dissertation, University of Michigan.

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23,* 483–496.

Goldberg, L. R. (1972). Student personality characteristics and optimal college learning conditions: An extensive search for trait-by-treatment interaction effects. *Instructional Science, 1,* 153–210.

Goulandris Foundation & J. Paul Getty Museum. (1993). *The Getty Kouros Colloquium: Athens, 25–27 May, 1992.* Athens: Kapon Editions.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik.* New York: Oxford University Press.

Hoving, T. (1996). *False impressions: The hunt for big-time art fakes.* New York: Simon & Schuster.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

Huntford, R. (1999). *The last place on earth.* New York: Modern Library.

Kelly, E. L. (1954). Evaluation of the interview as a selection technique. In *Proceedings of the 1953 Invitational Conference on Testing Problems* (pp. 116–123). Princeton, NJ: Educational Testing Service.

Knaus, W. A., & Wagner, D. P. (1989). APACHE: A nonproprietary measure of severity of illness. *Annals of Internal Medicine, 110,* 327–328.

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

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

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

Pepper, S. C. (1942). *World hypotheses.* Berkeley: University of California Press.

Pinker, S. (1997). *How the mind works.* New York: Norton.

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

Shepard, R. N. (1964). Attention and the metric structure of the stimulus. *Journal of Mathematical Psychology, 1,* 54–87.

Sherden, W. A. (1998). *The fortune sellers: The big business of buying and selling predictions.* New York: Wiley.

Tversky, A., Sattah, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95,* 371–384.

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