Thinking Clearly about Psychology

Volume 1: Matters of Public Interest

Edited by Dante Cicchetti
and William M. Grove

Essays in honor of Paul E. Meehl

University of Minnesota Press
Minneapolis Oxford

۱۹۹۱

Human Mind versus Regression Equation: Five Contrasts

Lewis R. Goldberg

> *Why should people have been so surprised by the empirical results in my summary chapter? Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well it looks to me as if it's about $17.00 worth; what do you think?" The clerk adds it up. There are no strong arguments, from the armchair or from empirical studies of cognitive psychology, for believing that human beings can assign optimal weights in equations subjectively or that they apply their own weights consistently.*
>
> (Meehl, 1986, p. 372)

It has been said that the single most difficult thing to do in the world is to write with clarity and style. (It has also been said that fine writing is easy: One simply stares at the blank paper until the sweat on one's brow turns to blood.) Paul Meehl writes with the elan of no one else in our field, and to read him is to experience

The preparation of this chapter was supported by Grant MH-39077 from the National Institute of Mental Health, U. S. Public Health Service. The author is indebted to W. Scott Armstrong, Thomas Bezembinder, Berndt Brehmer, Matthias Burisch, William F. Chaplin, Lee J. Cronbach, Robyn M. Dawes, Adriaan D. de Groot, David Faust, Baruch Fischhoff, David C. Funder, Janice C. Goldberg, Kenneth R. Hammond, Sarah E. Hampson, Paul Herzberg, Willem K. B. Hofstee, Robert Hogan, John Horn, Irving L. Janis, Oliver P. John, Daniel Kahneman, Henry F. Kaiser, Kevin Lanning, Sarah Lichtenstein, John C. Loehlin, Robert R. McCrae, Warren T. Norman, Daniel Ozer, Dean Peabody, Leonard G. Rorer, Tina K. Rosolack, Myron Rothbart, Paul Slovic, Ross Stagner, Auke Tellegen, Amos Tversky, and Jerry S. Wiggins for their thoughtful reactions to previous drafts. All of the remaining errors in this chapter are the fault of those who were invited to correct them but have neglected to do so.

an aesthetic rush along with the inevitable intellectual jolt. Indeed, while waiting for my own sweat to congeal, I sometimes engage in a peculiar form of bibliotherapy: I read a piece or two (or seven if the blood feels thin) by Meehl. The particular piece does not seem to matter all that much, although I have my favorites. Any section of *Clinical versus statistical prediction* (Meehl, 1954) will do the job, so that monograph was always near at hand throughout the 1960s. Later, Meehl concocted an even stronger tonic, "Why I do not attend case conferences," and therefore the book in which that chapter appeared (Meehl, 1973) soon took its turn as my security blanket. Reprinted in that 1973 collection is the classic, "When shall we use our heads instead of the formula?" (Meehl, 1957), in which Meehl tried to provide a logical analysis of the conditions favoring human over actuarial predictions.

Meehl has frequently expressed disappointment in his readers' near-exclusive focus on the clinical versus statistical box-score of the moment, rather than on the merits of his logical analyses. In the latter, he posited some characteristics of prediction problems in applied settings that might lead the human mind to triumph over typical actuarial procedures. Such possibilities included open-ended prediction tasks, unanalyzed stimulus equivalences, empty cells, theory mediation, insufficient time, and highly configural prediction functions (Meehl, 1959). The last of these, the only one of the set that was later to be studied empirically, concerns the nature of the relations between predictor and criterion variables. Meehl assumed that for many important prediction problems these relations would be nonlinear and/or interactive in character. If so, he reasoned, the linear additive assumptions incorporated into the multiple regression model would severely attenuate the predictive accuracy of any regression equation. On the other hand, if professional experts have learned these complex relations and can use them in diagnostic decision-making, then the human predictions should be superior to those generated by a regression equation.

Were it only so. As it has turned out, research findings suggest that highly configural predictor-criterion relations may be quite rare (e.g., Goldberg, 1965, 1969) and the human's ability to learn such complex relations, when they do occur, is far from optimal (e.g., Brehmer, 1979, 1980; Goldberg, 1968). As a consequence, it now appears that Meehl's initial concern about the limitations of the regression model no longer need be considered so seriously. Indeed, this highly popular and versatile technique could well be considered a prototypical example of standard actuarial methodology. In the present chapter, I will compare the features of multiple regression analysis with those of human numerical predictions. Specifically, I will ask: What does the regression equation do, and what would the human mind have to be able to do, in order for both mind and equation to forecast with equal accuracy? I will assume that readers share my understanding of the by-now-overwhelming empirical evidence favoring the use of actuarial methods (e.g., the regression equation) over unaided human intuition (the mind, however labeled) for predicting a diverse array of important human outcomes, and consequently I will not review that literature here. I will also assume that readers share my appreciation for the quite remarkable powers of the human mind over a wide range of contexts in which multiple-regression analyses are not appropriate, the prototypical example being that of pattern recognition.

In this chapter, I will describe five key features of a multiple-regression analysis, five problems in prediction that are automatically solved when one employs a regression algorithm. Specifically, a regression equation takes into account, and optimally[1] adjusts for, (a) differential *validities* within a set of predictors, (b) differences in *metrics* between the criterion and each of the predictors, (c) *consistency* of the forecasts from identical predictor patterns, (d) differential degrees of *redundancy* within sets of predictors, and (e) all *regression* effects associated with imperfect predictor-criterion associations. Historically, proponents of statistical predictions focused our attention on the first of these features, and perhaps as a consequence investigators of human decision-making explored them in roughly this order, as will I.

Adjusting for Differential Validities among the Predictors

The most obvious feature of a multiple-regression analysis is its optimal weighting of the predictors as a function of their differential validities: Other things being equal, the stronger the relation between a predictor and the criterion, the more weight that predictor is given in the regression equation. Clearly, humans cannot do this task perfectly, and perhaps they do it quite poorly. To discover the weights that experts implicitly assign to predictor information, Hammond (1955) and Hoffman (1960) suggested that we try to capture those weights by constructing a regression equation to predict each individual's judgments. Hoffman (1960), who labeled these equations "paramorphic representations" of the judgment process, assumed that "the regression weights signify, with certain limitations, the emphasis or importance attached to each of the predictor variables by the judge. Large coefficients mean, empirically, that the corresponding predictors can account for large proportions of the variance of judgment; and a predictor with a small beta coefficient contributes little beyond the contribution of other predictors" (p. 120). Later, Hammond, Hursch, and Todd (1964) and Tucker (1964) mathematically formulated some of the components of clinical inference within this multiple-regression framework. For more complete accounts of this analytical strategy and the research it spawned, see Hammond, McClelland, and Mumpower (1980).

All of this early work was based on the assumption that a major—if not the major—difference between human and equation was non-optimal versus optimal weighting of the predictor validities. Not until a decade later were we told that this difference may not be as critical as we had assumed. Dawes and Corrigan

(1974; see also Dawes, 1979) showed that equations with unit weights (sometimes even with random weights) produced predictions similar to those produced by models based on regression weights under three conditions: (a) a *correlational* index is used as the standard of comparison among the prediction methods, (b) the *signs* of the predictor-criterion relations are correctly identified, and (c) the *metrics* of the predictors have been transformed so as to eliminate any differences in predictor variances (e.g., they are all standardized in z-score form). Because the widely cited findings of Dawes and Corrigan (1974) have frequently been overgeneralized, it is important to try to clarify some of the basic issues they raise.

First of all, as has been brilliantly articulated by Cronbach and Gleser (1953) in another context, there are different statistical standards that can be used to index the accuracy of numerical predictions. Within the present context two standards have been invoked, those based on *relational* indices such as the correlation coefficient and those based on an average *discrepancy* between predicted and criterion values such as the standard error of prediction (i.e., the square root of the mean squared error). Dawes and Corrigan (1974) showed that under certain conditions the predictions generated by unit-weighted and regression-weighted models may be highly correlated. Indeed, as demonstrated some time ago by Gulliksen (1950, pp. 312–327) and more recently by Einhorn and Hogarth (1975) and von Winterfeldt and Edwards (1987), the size of the expected correlation between the predictions from any two models is a negative function of the number of predictors and a positive function of their average intercorrelation. Moreover, whereas the predictions from two models may be highly related, one could still relate positively and the other negatively with any other variable, including a criterion variable. For example, even when the predictions from two models correlate .80, one of them could correlate as high as .32 and the other as low as −.32 with the same criterion. In addition, correlations do not reflect any constant or systematic errors that affect the means and/or variances of the predictions. In the remainder of this chapter, I will assume the use of some accuracy measure that is so affected, such as the standard error of prediction.

Using such a standard, what do we know about the accuracy with which people can learn to assess the relative validity of different predictors? The classic work on this problem was carried out by Chapman and Chapman (1967) and subsequently replicated and extended by Golding and Rorer (1972). In a nutshell, they showed that we can rather easily be led to see what we expect to see, even when what we expect is not there. Specifically, when people expect to find predictor-criterion relations in a stimulus set in which those relations are absent, they "find" them anyway. What this suggests is that our ability to learn the differential validities of predictors is far from perfect (see Brehmer, 1980; Gaeth & Shanteau, 1984). Unlike the regression equation, which performs this chore optimally, the human mind does not. Perhaps the most pessimistic assessment of the powers of the human mind in this regard has been provided by Nisbett and Ross (1980), who concluded that the evidence shows that people are poor at detecting many sorts of covariation: "The layperson apparently does not understand intuitively the basic logic of covariation assessment underlying the 2 × 2 contingency table. Perception of covariation in the social domain is largely a function of preexisting theories and only very secondarily a function of true covariation. In the absence of theories, people's covariation detection capacities are extremely limited" (p. 111).

Aligning the Metrics of the Predictors with That of the Criterion

In the studies by Dawes and his colleagues, both the "improper" and the "proper" weights under study were always applied to standardized (z-scored) predictors. Indeed, the very notion of "unit" or equal weights makes no sense unless applied to predictors scaled in the same metric. Other things being equal, the contribution of each part to a composite is a direct function of the relative variances of the parts; specifically, the larger the variance of a part, the more it contributes to the variance of the composite. One function of a regression equation is to align the metrics of each of the predictors with that of the criterion, and it does this in an optimal manner. Clearly, humans will not align diverse metrics so keenly, and this cognitive limitation could result in their intuitive weights differing from those they intended.

Some studies of the effects of differing predictor metrics were carried out by Paul Hoffman in the early 1960s. In the one published report of this work (Knox & Hoffman, 1962), judgments based on predictor scores expressed in T-score metric were compared with judgments based on the same scores expressed in percentiles. College students rated the intelligence or sociability of (fictitious) target persons on two occasions, using eight potential predictors of sociability and nine of intelligence. Each profile was displayed in either T-score or percentile format. Within each of the eight cells generated by three experimental conditions (T score vs. percentile, intelligence vs. sociability, test vs. retest), a regression model was fitted to the judgments of each subject. Of the two metrics under study, the percentile format was associated with greater variance of the judgments (and therefore higher test-retest reliability and higher predictability by the paramorphic representations), leading the investigators to conclude that "judgments from profiles are influenced not only by the underlying meaning of the plotted scores but by their graphical location as well" (Knox & Hoffman, 1962, p. 14). On the other hand, differences in metric were not associated with different patterns of relative weights of the predictors. Moreover, because the profiles were fictitious, no assessment of differential accuracy was possible. Finally, within each of the eight conditions all predictors were scaled in the same metric, thus rendering impossible any analyses of potential difficulties in aligning diverse metrics within the same task.

Such difficulties were hypothesized, and investigated indirectly, in one of the five studies reported by Slovic and MacPhillamy (1974). In these studies, sets of stimuli were compared on features, some of which are common to all stimuli and some of which are unique to each one. The major hypothesis was that people would tend to weight the common features more highly than the unique ones. Subjects compared pairs of hypothetical high school students on their potential college success. Scores were available for both students in each pair on one common attribute (e.g., English skills) and for each student on one unique attribute. For half of the subjects, scores on the three (one common and two unique) attributes were all in the same metric, whereas for the other half of the subjects each of the three attributes was scored in a different metric. Unfortunately, all analyses were focused on the effects of common vs. unique attributes, rather than on the comparison between same and different metrics (labeled "equal vs. unequal" units). The investigators concluded that the "common-dimension effect was as strong in the equal as in the unequal-units conditions, contrary to the expectation that it would be easier to use the unique information in the equal-units condition since there was less data transformation for the judge to consider" (Slovic & MacPhillamy, 1974, p. 180).

Additional indirect evidence about human limitations in aligning different metrics comes from some recent studies by Tversky, Sattath, and Slovic (1988) that focused on individuals' judgments about their preferences. Based on the initial findings of Slovic (1975), these investigators showed that attributes that are scaled in a metric that is compatible with that of the judgmental response format will be weighted more highly than are attributes scaled in any other metric. Said another way, in intuitive decision-making the weight of an attribute is enhanced by its compatibility with the output format or metric. The probable reason for this effect is that the incompatibility between input and output requires additional mental transformations, which in turn demand increased effort, which can lead to strain and to errors. For example, in one study subjects were asked to predict the decisions of a college admissions committee from applicants' *ranks* on one attribute (e.g., an ability test) and from *categorical* values on another (e.g., a strong versus a weak pattern of extracurricular activities). Subjects were told that the committee ranked all applicants and accepted the top quarter, and they were randomly assigned to predict either the rank or the acceptance decision for each applicant. As hypothesized, subjects asked to predict the rankings weighted the attribute scaled in ranks more highly than the attribute presented categorically, whereas the reverse pattern characterized subjects asked to predict the categorical decisions.

In summary, then, although none of the studies to date has focused directly on numerical predictions in applied settings, all of the indirect evidence suggests that the task of aligning attributes scaled in different metrics is not likely to be a forte of the human mind. Regression equations, on the other hand, handle this aspect of prediction tasks as facilely as they handle other aspects.

Forecasting with Consistency

In the quotation with which this chapter begins, Meehl noted that "There are no strong arguments . . . for believing that human beings can . . . apply their own weights consistently." Hubris demands that I now finish that sentence: "the query from which Lew Goldberg derived such fascinating and fundamental results" (Meehl, 1986, p. 372). The results to which Meehl alludes (Goldberg, 1970) have been cited frequently enough, and have now been replicated in enough new contexts, that I can describe them here quite briefly. When a paramorphic representation of a person's judgments is substituted for that person in the same predictive context, the person's model typically turns out to be more accurate than the person's own judgments. How can that be? The paramorphic model is a regression equation, and one crucial feature of all such equations is that their test-retest reliabilities are perfect. Given the same pattern of predictor values, the equation must generate identical predictions. Humans, on the other hand, are not like equations: Their repeated judgments of the same predictor configurations are often different. Indeed, one can be almost completely unreliable in a predictive situation in which one expects to be perfectly reliable, as was revealed in an intensive study of an experienced graphologist whose test-retest reliability turned out to be near zero (Goldberg, 1986). Because unreliability inherently limits potential validity, it is undesirable in any predictive context.

Over the past decade, there have been frequent demonstrations that seemingly innocuous and normatively inconsequential changes in the formulation of prediction problems can produce substantial inconsistencies when persons respond to two or more forms of the same problem. These inconsistencies are referred to as "framing" effects, and they have now been obtained across a quite diverse set of judgmental tasks. Because superb reviews of framing effects are readily available (e.g., Kahneman & Tversky, 1984; Slovic, Fischhoff, & Lichtenstein, 1982), I will say no more about them here other than to point to their relevance for any comparisons between the human mind and the regression equation.

Accounting for Predictor Redundancies

Not all of the attributes we use as predictors come packaged independently of each other; indeed, most attributes covary in nature to some extent. Such covariation implies information redundancy, and redundant predictor information adds no zest to the predictive punch. A multiple regression equation automatically takes the intercorrelations among the predictors into account in deriving the

Table 1. Types of Predictor-Criterion Correlational Patterns

Correlation	Redundancy Strong	Redundancy Weak	Independence	Enhancement Weak	Enhancement Strong
Validity of the first predictor (r_{yx_1})	.50	.50	.50	.50	.50
Validity of the second predictor (r_{yx_2})	.40	.40	.40	.40	.40
Correlation between the two predictors ($r_{x_1x_2}$)	.60	.30	.00	−.30	−.60
Multiple Correlation ($R_{y.x_1x_2}$)	.52	.56	.64	.76	1.00

predictor weights. Humans are unlikely to do this optimally. Even worse, they tend to value redundancy for its own sake, because redundancy leads them to have increased confidence in their predictions (Slovic, 1966).

Other things being equal, predictive accuracy is increased when all predictors are mutually orthogonal, and accuracy is enhanced even more when the predictors are correlated with each other in ways other than the ways they are correlated with the criterion. Table 1 provides examples in the most simple case—two predictors (x_1 and x_2) and one criterion (y)—of three types of information configurations: redundancy, independence, and enhancement. In all cases, the validity of one predictor is .50 and that of the other is .40. When the predictors are redundant, use of them both does not improve predictive accuracy much beyond what can be attained by the best of them alone. When the predictors are unrelated, predictability is increased considerably. When the sign of the correlation between the predictors is the oppposite of the signs of their validity coefficients, predictability is enhanced substantially. Indeed, when the two predictors are correlated approximately −.60, multiple prediction is perfect.

Given sets of stimulus materials corresponding to each of these configurations in a learning task, humans much prefer redundancy to either of the other two conditions, even though they should be able to predict most accurately in their least preferred configuration. To demonstrate this effect, Kahneman and Tversky (1973) asked subjects to predict grade-point average on the basis of two pairs of aptitude tests. Subjects were told that one pair of tests (e.g., creative thinking and symbolic ability) was highly related, whereas the other pair of tests (e.g., mental flexibility and systematic reasoning) was unrelated. The scores they encountered conformed to these expectations. Subjects were told that "all tests were found equally successful in predicting college performance." Although in this situation higher predictive accuracy can be achieved with the unrelated than with the related pair of tests, subjects were significantly more confident in predicting from the related pair. Specifically, "they were more confident in a context of inferior predictive validity" (Kahneman & Tversky, 1973, p. 249).

Other things being equal, the more certain we are of being correct, the more extreme are the numerical values we give; when we are in doubt, we tend to respond more cautiously, and thus more neutrally. As a consequence, any increment in redundancy within a predictor set tends to increase the variance of intuitive predictions. And, to the extent to which our confidence is misplaced—that is, we are "more confident in a context of inferior predictive validity"—we are doing the opposite of what we should do under the circumstances. What we should do is covered in the next section.

Adjusting for Regression Effects

Regression to the mean is like the weather: Everybody talks about it, but few of us do anything about it. However, it is unlike the weather, because most of us fail to recognize it, even when it hits us on the nose. Consider the following regression axiom: In the context of numerical prediction problems, the variance of our predictions should never be larger than that of the criterion we seek to predict. (Never, not just hardly ever.) Indeed, virtually always the variance of our predictions should be much smaller than that of the criterion. As their name implies, regression equations are specifically constructed to handle regression effects optimally. Like the other four features of this analytic method, the fifth just comes with the territory.

Although it has long been known, at least by statistics instructors, that the concept of regression was not part of our human intuitions, it remained for Kahneman and Tversky (1973) to document the extent of this cognitive limitation, to analyze some conditions under which this failing is particularly severe, and to provide a theoretical rationale to explain both how we make intuitive predictions and why they tend to be nonregressive in character. Briefly, these investigators have shown that intuitive predictions are often based on a heuristic called "representativeness"—specifically, that we tend to assess outcomes by the extent to which the outcomes resemble or "represent" the main features of the evidence. Predictions based on resemblance are not regressive, whereas all predictions should be.

The best summary of the evidence on the use and misuse of the representativeness heuristic in making intuitive predictions has been provided by Nisbett and Ross (1980; see especially chapter 7), who argued that the tendency to be insufficiently regressive has two causes. On the one hand, people may often overestimate the degree of covariation among events in the social domain, at least when such events can be linked by plausible causal theories, scripts, or stereotypes. More important, however, is the fact that people fail to make regressive predictions even when they recognize that the available predictor variables are only weakly related to the criterion. That is, incorrect theories about the strength of relations lead to exaggerated beliefs about the utility of potential predictors, and intuitive prediction strategies result in nonregressive predictions even when the weakness of the relations is recognized. Nisbett and Ross (1980) concluded that "when it comes to predictions, a little knowledge (that is, knowledge of the target's score on a weakly related predictor variable) is a dangerous thing" (p. 153).

Summary

Of the five types of problems solved automatically by the regression equation, which is the most troublesome for the human mind? In a first draft of this chapter, I questioned whether the relative importance of the five problems in differentiating human from equation was not the very opposite of their historical order of investigation (and the opposite of their order of presentation here). I speculated that the most obvious limitation of the human mind as compared to the regression equation—failure to give proper weight to differences in the validities of the predictors—may not turn out to be the key villain in this mystery story. Rather, the very feature of the regression equation that gives it its name may be the most important of the five contrasts. Now older and wiser (see the note on the first page of this chapter), I lean toward reversing my initial importance ranking. In so doing, I thereby demonstrate the very deficiencies that I have been chronicling—an inability to figure out the relative validities of each of the five features, confusion stemming from the incommensurability of their metrics, obvious inconsistency in my forecasting, difficulty combining nonredundant features, and finally a failure to regress to the mean, which in this case would amount to a prediction that all the features are quite similar in their importance. Indeed, all that is clear to me at this point is that as humans we do none of them optimally.

In the 1950s, it was Paul Meehl who brought to the attention of both the clinical and the scientific communities in psychology the necessity of comparing the accuracy of intuitive and actuarial procedures so as to better understand "what can the clinician do well?" In so doing, he forced us to become aware of our cognitive limitations, and he led us to think more carefully about the logic of the predictive enterprise. During the 1960s and 1970s, his name was primarily associated with his summaries of the empirical comparisons between intuitive and actuarial predictions. Now in the 1990s, we should be ready to appreciate his contributions for what *he* valued most in them, his extraordinarily prescient insights into the assets and the liabilities associated with human intuitions. To the extent to which we do so, we will begin to attentuate his anguish about his readers' reactions, or lack thereof, to his 1954 monograph:

> People sometimes ask me whether I am disappointed by the relatively feeble impact of that book and of the many studies that were stimulated by it. . . . I have learned to develop a certain Buddhistic detachment about the matter. Suppose a social worker confidently tells me that of course we can predict how this delinquent will do on probation by reflecting on psychodynamic inferences and subjective impressions, recorded in a 10-page presentence investigation, despite the malignant rap sheet record and acting-out psychometrics, and the officer's comment that "he's a real mean, tough street kid." Well, I remind myself that Omniscient Jones has not put me in charge of reforming the world (Meehl, 1986, p. 375).

Note

1. The coefficients in a multiple-regression equation are "optimal" only in the sample in which they were derived, and consequently the validity of a regression equation will tend to shrink when it is applied in new samples. In general, the robustness of the equation is negatively related to the number of predictor variables used to derive it, and positively related to the size of the derivation sample and its representativeness of the population to which the equation is to be applied.

References

Brehmer, B. (1979). Effects of practice on utilization of nonlinear rules in inference tasks. *Scandinavian Journal of Psychology, 20,* 141–149.

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica, 45,* 223–241.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72,* 193–204.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50,* 456–473.

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

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

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

Gaeth, G. J., & Shanteau, J. (1984). Reducing the influence of irrelevant information on experienced decision makers. *Organizational Behavior and Human Performance, 33,* 263–282.

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

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

Goldberg, L. R. (1969). The search for configural relationships in personality assessment: The diagnosis of psychosis vs. neurosis from the MMPI. *Multivariate Behavioral Research, 4,* 523–536.

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

Goldberg, L. R. (1986). Some informal explorations and ruminations about graphology. In B. Nevo (Ed.), *Scientific aspects of graphology* (pp. 281–293). Springfield, IL: Charles C. Thomas.

Golding, S. L., & Rorer, L. G. (1972). Illusory correlation and subjective judgment. *Journal of Abnormal Psychology, 80,* 249–260.

Gulliksen, H. (1950). *Theory of mental tests.* NY: Wiley.

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

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71,* 438–456.

Hammond, K. R., McClelland, G. H., & Mumpower, J. (1980). *Human judgment and decision making: Theories, methods, and procedures.* NY: Praeger.

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39,* 341–350.

Knox, R. E., & Hoffman, P. J. (1962). Effects of variation of profile format on intelligence and sociability judgments. *Journal of Applied Psychology, 46,* 14–20.

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

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4,* 268–273.

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

Meehl, P. E. (1967). What can the clinician do well? In D. N. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 594–599). NY: McGraw-Hill.

Meehl, P. E. (1973). *Psychodiagnosis: Selected papers.* Minneapolis: University of Minnesota Press.

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

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Slovic, P. (1966). Cue-consistency and cue-utilization in judgment. *American Journal of Psychology, 79,* 427–434.

Slovic, P. (1975). Choice between equally valued alternatives. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 280–287.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1982). Response mode, framing, and information-processing effects in risk assessment. In R. Hogarth (Ed.), *New directions for methodology of social and behavioral science: Question framing and response consistency: No. 11* (pp. 21–35). San Francisco: Jossey-Bass.

Slovic, P., & MacPhillamy, D. (1974). Dimensional commensurability and cue utilization in comparative judgment. *Organizational Behavior and Human Performance, 11,* 172–194.

Tucker, L. R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review, 71,* 528–530.

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

von Winterfeldt, D., & Edwards, W. (1987). *Decision analysis and behavioral research.* NY: Cambridge University Press.