Reprinted from AMERICAN PSYCHOLOGIST, Vol. 23, No. 7, July 1968
Printed in U. S. A.

SIMPLE MODELS OR SIMPLE PROCESSES?

SOME RESEARCH ON CLINICAL JUDGMENTS [1]

LEWIS R. GOLDBERG

University of Oregon and Oregon Research Institute

IMAGINE the following situation: You are sitting unobserved in a physician's office watching a week of his professional activities. During the course of the week, some 100 patients come to his office, each telling him of his symptoms, which you record; after each patient leaves the office, and any requested laboratory findings have arrived, the physician records his diagnosis for that patient. At the end of the week you have collected a set of 100 symptom configurations, one for each patient, and a set of 100 corresponding diagnoses.

Alternatively, you are sitting unobserved in the office of a personnel officer of a large manufacturing concern. He has 100 folders on his desk, each containing information about a different applicant for 50 sales positions with his company. He spends his week carefully looking through the application materials for each applicant—examining the applicant's test scores, the ratings made by each of the company's three initial interviewers, and the reference forms from each of the applicant's past employers. When he has completed examining the materials for each applicant in turn, he records his selection decision. At the end of the week each of the 100 folders of application data has a corresponding personnel recommendation associated with it.

Again alternatively, you are watching a clinical psychologist function over the course of a month at a busy outpatient psychiatric clinic. Most of his day he spends administering tests and interviewing patients. But, for a few hours at the end of every day he gathers together all of the information he has collected on one patient, examines it all carefully, and proceeds to write a report of his findings. In this report he includes some descriptive statements about the patient and his problems, the patient's diagnosis, and some predictions of the likelihood of certain important consequences for the clinic (e.g., the probability of the patient committing suicide, his probable response to treatment, etc.). The data collected from the patient (test scores, interview notes, etc.) are stored in one folder; the resulting reports are sent elsewhere in the clinic. At the end of the month, you can gather together 100 patient folders, plus the 100 corresponding psychological reports.

Each of these three professional activities has as its central core a reliance upon what the practitioner might call "clinical wisdom," but which in psychology is more modestly called "clinical judgment." Each is an important human cognitive activity, typically carried out by a professional person, aimed at the prediction of significant outcomes in the life of another individual. When the same type of prediction is made repeatedly by the same judge, using the same type of information as a basis for his judgments, then the process becomes amenable to scientific study. And, not surprisingly, over the past 20 years the clinical judgment process has begun to be studied intensively by investigators all over the world.

THE FOCUS ON ACCURACY

Historically, the earliest research efforts centered on the accuracy of such clinical judgments. And,

[1] This is a revised version of an invited address presented at the meeting of The Netherlands Psychological Association, April 14, 1967, in Nijmegen, The Netherlands. An earlier version has been published in the Dutch journal, *Gawein*. The address was prepared while the author was serving as Fulbright Professor of Psychology at the University of Nijmegen (Psychologisch Laboratorium der Katholieke Universiteit, Nijmegen), during the 1966–67 academic year. The author is deeply indebted to the staff of the Oregon Research Institute, especially Paul J. Hoffman, Leonard G. Rorer, and Paul Slovic, for their help in formulating some of these ideas, for their stimulation and encouragement of the author's research, and for their own research efforts—many of which are discussed in this paper. Most of the Oregon Research Institute judgmental studies have been funded by Research Grants MH-04439, MH-10822, or MH-08160 from the National Institute of Mental Health, United States Public Health Service. While this paper is not intended as a comprehensive review of all studies of the clinical judgment process, a more complete bibliography is available from the author.

since World War II had sparked the emergence of clinical psychology as an applied speciality area (in which, at least at first, clinicians spent a good deal of their professional time making diagnostic judgments), it was natural that the first major focus of accuracy research was upon the diagnostic acumen of clinical psychologists themselves. Over the past 20 years, a flurry of such studies has appeared, the most dramatic and influential being the early ones reported by Kelly and Fiske (1951) and Holtzman and Sells (1954).

Studies of the accuracy of these sorts of judgments have yielded rather discouraging conclusions. For example, one surprising finding—that the amount of professional training and experience of the judge does not relate to his judgmental accuracy —has appeared in a number of studies (e.g., Goldberg, 1959; Hiler & Nesvig, 1965; Johnston & McNeal, 1967; Levy & Ulman, 1967; Luft, 1950; Oskamp, 1962, 1967; Schaeffer, 1964; Silverman, 1959; Stricker, 1967). Equally disheartening, there is now a host of studies demonstrating that the amount of information available to the judge is not related to the accuracy of his resulting inferences (e.g., Borke & Fiske, 1957; Giedt, 1955; Golden, 1964; Grant, Ives, & Ranzoni, 1952; Grigg, 1958; Hunt & Walker, 1966; Jones, 1959; Kostlan, 1954; Luft, 1951; Marks, 1961; Schwartz, 1967; Sines, 1959; Soskin, 1959; Winch & More, 1956). Let us look at Oskamp's (1965) study as one example of some of these findings.

Oskamp had 32 judges, including 8 experienced clinical psychologists, read background information about a published case, divided into four sections. After reading each section of the case in turn, and thus before seeing any other information, each judge answered a set of 25 questions about the personality of the target (questions for which the correct answers were known to the investigator). For each question, the judge also indicated his confidence in the accuracy of his prediction by indicating the percentage of questions answered with that much confidence that he would expect to answer correctly. Oskamp found that as the amount of information about the target increased, accuracy remained at about the same level, while confidence increased dramatically. In general, the average judge was slightly overconfident when he had only one-fourth of the total amount of data available to him (he estimated that he would be correct on 33% of the questions, while he was actually correct on 26%); by the time he had seen all of the information, however, he was extremely overconfident (53% estimated correct versus 28% actually correct). Oskamp (1965) concluded:

> the judges' confidence ratings show that *they become convinced of their own increasing understanding of the case.* As they received more information their confidence soared. Furthermore, their certainty about their decisions became entirely out of proportion to the actual correctness of those decisions [p. 264].

For another demonstration of this same phenomenon, see Ryback (1967).

Such findings relative to the validity of clinical judgments obviously raise questions as to their reliability. Within the judgment domain, we can distinguish at least three different types of inferential reliability (Goldberg & Werts, 1966): (*a*) *stability,* or reliability across time (for the same judge using the same data); (*b*) *consensus,* or reliability across judges (for the same data and the same occasion); and (*c*) *convergence,* or reliability across data sources (administered on the same occasion and interpreted by the same judge). While the relatively few investigations of judgmental stability have concluded that judges may show substantial consistency in their judgments over time, the vast majority of reliability studies have focused upon judgmental consensus and have come to widely disparate conclusions. Findings have ranged from extremely high consensus on some judgmental tasks (e.g., Bryan, Hunt, & Walker, 1966; Goldberg, 1966; Hunt & Jones, 1958a, 1958b; Hunt, Jones, & Hunt, 1957; Hunt, Walker, & Jones, 1960; Weitman, 1962; Winslow & Rapersand, 1964) to virtually no consensus on other tasks (e.g., Brodie, 1964; Grosz & Grossman, 1964; Gunderson, 1965a, 1965b; Howard, 1963; Marks, 1961; Ringuette & Kennedy, 1966; Watley, 1967; Watson, 1967).

The classic study of the convergence among clinical inferences was carried out by Little and Schneidman (1959), who concluded that the reliability of clinicians' judgments leaves "much to be desired" (a most dramatic understatement if one examines their important findings). In a more recent study, Goldberg and Werts (1966) concluded that "an experienced clinician's judgments from one data source do *not* correlate with another clinician's judgments from another data source, even though both clinicians are diagnosing

the very same patient on—ostensibly—the very same trait [p. 205]." Most of the other studies of judgmental convergence (e.g., Howard, 1962, 1963; Phelan, 1964, 1965; Vandenberg, Rosenzweig, Moore, & Dukay, 1964; Wallach & Schooff, 1965) have tended to confirm this somewhat dismal general picture.

If one considers the rather typical findings that clinical judgments tend to be (*a*) rather unreliable (in at least two of the three senses of that term), (*b*) only minimally related to the confidence and to the amount of experience of the judge, (*c*) relatively unaffected by the amount of information available to the judge, and (*d*) rather low in validity on an absolute basis, it should come as no great surprise that such judgments are increasingly under attack by those who wish to substitute actuarial prediction systems for the human judge in many applied settings. Since I assume that virtually all psychologists are acquainted with what has come to be known as the "clinical versus statistical prediction controversy" (e.g., Gough, 1962; Meehl, 1954, 1956, 1957, 1959, 1960; Sawyer, 1966), I can summarize this ever-growing body of literature by pointing out that over a rather large array of clinical judgment tasks (including by now some which were specifically selected to show the clinician at his best and the actuary at his worst), rather simple actuarial formulae typically can be constructed to perform at a level of validity no lower than that of the clinical expert.

THE FOCUS ON THE JUDGMENT PROCESS

As a consequence of these sorts of findings, the research focus among judgmental investigators has begun to turn from validity studies to investigations of the process of clinical inference, the aim of which is to "represent" (or "simulate" or "model") the hidden cognitive processes of the clinician as he makes his judgmental decisions (Hoffman, 1960). Hopefully, by understanding this process more completely than we do today, clinical training programs can be made more effective and judgmental accuracy can thereby be increased.

An investigator of the clinical judgment process might express his aims through the following questions: By what psychological model can one best depict the cognitive activities of a judge? More specifically, what model allows one to use the same data available to the judge and combine these data so as to simulate most accurately the judgments he actually makes? To return to the three illustrative examples described at the beginning of this paper, these questions could be reformulated, respectively:

1. By what model can the 100 symptom configurations from each of the 100 patients be combined so as to generate most accurately the physician's resulting diagnoses?

2. By what model can the information extracted from the 100 applicant folders be combined so as to produce the most accurate prediction of the personnel officer's selection decisions?

3. By what model can the information from the psychological folders of the 100 psychiatric patients be combined so as to most accurately reproduce the material found in the 100 psychological reports?

All of these questions have some common elements, namely, (*a*) a search for some formal (i.e., specifiable) model, which (*b*) uses as its "input" the information (data, cues, symptoms, etc.) initially presented to the judge, and (*c*) combines the data in some optimal manner, so as to (*d*) produce as accurate as possible a copy of the responses of the judge—(*e*) regardless of the actual validity of those judgments themselves. Note that such a model is always an intraindividual one; that is, it is intended as a representation of the cognitive activities of a single judge. Moreover, the test of the model is not how well it works as a representation of the state of the world (e.g., how well it predicts who will or will not be a successful employee), but rather how well it predicts the inferential products of the judge himself.

In mathematical terms, one begins with a cue matrix of size $M \times N$, where $M =$ the number of variables presented to the judge and $N =$ the number of targets for which the judge is asked to predict. One wishes to discover some combinatorial model which will reproduce as accurately as possible the vector of N responses produced by the judge to the same cue matrix. For this process to be amenable to mathematical analyses, the original cue matrix and the terminal judgmental response vector should be in a quantified format (or in a format easily transformable into a set of numbers). While it is fashionable to lament about the difficulty of transforming "behavioral" data into quantitative form, this difficulty may be more apparent than real. For if the cues (and resulting judgments) can be represented in even so simple a form as a

binary digit (e.g., the patient has characteristic X versus the patient does not have this characteristic), then quantification is straightforward (e.g., $X = 0$; non-$X = 1$). Since a good deal of the data available to many clinicians is already in quantitative form (e.g., test scores, laboratory findings) or can be easily transformed quantitatively with no apparent loss of information (e.g., trait ratings), it is typical for most judgmental investigators to simply present the data to the judge and ask for the judgmental responses in a previously quantified format.

What sort of judgmental model should one try? Since introspective accounts describe the clinical judgment process as curvilinear, configural, and sequential (e.g., McArthur, 1954; Meehl, 1954, 1960; Parker, 1958), one possible strategy is to begin with fairly complex representations, perhaps with an eye to seeing how they may eventually be simplified. For example, Kleinmuntz (1963a, 1963b, 1963c) had a clinician "think aloud" into a tape recorder as he made judgments about the adjustment of college students on the basis of their MMPI profiles. Kleinmuntz then used these introspections to construct a computer program simulating the clinician's thought processes. The resulting program was a complex sequential (e.g., hierarchical or "tree") representation of the clinician's verbal reports.

The research of investigators at two major centers for research on clinical inference—Oregon Research Institute and the Behavior Research Laboratory of the University of Colorado—has proceeded from a diametrically opposite strategy (see Hammond, Hursch, & Todd, 1964; Hoffman, 1960), namely, to start with an extremely simple model and then to proceed to introduce complications only so far as is necessary to reproduce the inferential responses of a particular judge. Rather than beginning with a model which is already complex (e.g., curvilinear, configural, sequential) as Kleinmuntz did, we have opted to start with what is perhaps the simplest of all models: a linear, additive, regression model (of the sort now used rather universally for a host of applied prediction problems). That is, we begin with the hopefully naive assumption that the responses of a person in a judgment task can be reproduced by a mathematical model of the form:

$$Z = b_1X_1 + b_2X_2 \cdots + b_kX_k$$

where Z is the vector of judgmental responses, $X_1 \ldots X_k$ are the values of the matrix of K cues by N targets presented to the judge, and $b_1 \ldots b_k$ are constants representing the "weight" of each cue in the judgmental model. In practice, the X values (the $N \times K$ matrix of cues) are known to the investigators (they are the stimulus or input variables presented to the judge) and the Z values are produced by the judge during the course of the experiment. The b values (regression weights) are found from one subset of the judge's responses by a standard linear regression analysis, and the "accuracy" of this linear model can then be ascertained by cross-validating these regression weights on the other subset of the judge's responses. The resulting correlation coefficient (R_a) represents the extent of agreement between the linear model and the inferential products of the judge.

Since we routinely ask each judge to make his judgments on two occasions (typically these "retest" protocols are sandwiched among the original protocols so that the judge is unaware of the fact that he is ever judging the exact same protocol twice), it is possible to compute a reliability coefficient (r_{tt}) to represent the stability of the judge's responses (or, alternatively, the extent to which one can predict his judgments from his own previous judgments of the same stimuli). This reliability coefficient can be viewed as the upper limit to the predictability of any model which we might construct. To the extent that the value of R_a approaches the value of r_{tt}, the model can be seen as representing the cognitive processes of the judge. When R_a and r_{tt} are identical for a particular judge, we have a perfect "paramorphic representation" of his judgment processes. Hoffman (1960) introduced this term to indicate that we do not pretend to be mapping any mind in an "isomorphic" fashion, but are merely seeking to discover some model which accurately generates the judgmental responses themselves.

Since clinicians generally describe their cognitive processes as complex ones involving the curvilinear, configural, and sequential utilization of cues, one might expect that the linear additive model would provide a rather poor representation of their judgments. Consequently, we might anticipate the need to introduce into the model mathematical expressions to represent these more complex processes. For example, if the judge is using a particular cue (X) in a curvilinear fashion (e.g., a personnel

officer may feel that applicants who score in the middle range of a standardized intelligence test will be more successful salesmen than those who score at either extreme), then we may be able to approximate this judgmental process by adding to the model terms like X^2, or X^3, X^4, etc. That is, we can represent curvilinear cue utilization generally by introducing into the more basic equation terms of the form bX^a, where X represents the cue value, a is a power constant reflecting the particular curvilinear use of that cue by the judge, and b is once again the weight of the entire term in the overall judgmental model.

While clinicians frequently attest that they use cues in a curvilinear fashion, even more commonly do they call attention to their use of cues in a configural (or interactive) manner. What they mean is that their judgments are not simply dependent on the value of a particular cue, but rather that the relationship between cue X_1 and their response is dependent upon (i.e., interacts with) the value of a second cue, X_2. For example, a physician might feel that body temperature is positively related to the likelihood of some disease if a patient has symptom Y, while temperature has no relevance for this diagnosis if the patient does not have symptom Y. Therefore, once again we must find mathematical expressions which approximate such configural cue usage. One way to express the interactive use of two cues, X_1 and X_2, is by the product term, $X_1 \cdot X_2$. Higher order interactions could be introduced into the basic equation by using even more complex cross-products (e.g., $X_1 \cdot X_2 \cdot X_3$, a term analogous to the three-way interaction line in the classical analysis of variance).

What should be clear from these examples is that we can systematically begin to introduce more complex terms into the basic multiple regression model and see whether the new models are more adequate representations of the judge's mental processes than was the original linear one. In general, we can introduce curvilinearity in cue utilization, for example,

$$\sum_{i=1}^{k} b_i X_i^{a_i}$$

configurality, for example,

$$\sum_{i=1}^{k-1} \sum_{j=2}^{k} b_{ij} X_i \cdot X_j \qquad (i<j)$$

and, of course, much more complex sets of terms, for example,

$$\sum_{i=1}^{k-1} \sum_{j=2}^{k} b_{ij} X_i^{a_i} \cdot X_j^{a_j} \qquad (i<j)$$

While the introduction of additional terms into the model can never serve to decrease its accuracy in the sample of judgments used to derive the b weights, these extra terms may simply serve to explain chance characteristics of the particular judgments from the derivation sample and thus can severely attenuate the accuracy of the resulting model upon its cross-validation in another sample of judgments. However, when the judge is actually using the cues in a curvilinear or in a configural manner, then the introduction of the mathematical approximations of these processes should serve to improve the model.

While the preceding discussion has focused primarily on the use of multiple regression techniques, it could just as easily have been formulated in terms of the fixed-model analysis of variance (ANOVA), both systems simply being alternative formulations of a general linear model. Since the structural elements underlying both the multiple regression and the ANOVA model are formally equivalent, it is often possible to use the latter in judgment studies—thereby capitalizing on the well-known descriptive and inferential properties of ANOVA (Hoffman, Slovic, & Rorer, 1968). However, the ANOVA model imposes two important restrictions on the cue values to be used in judgment research: (a) the cues must be treated as categorical rather than continuous variables; and (b) the cues must be orthogonal (uncorrelated). While these restrictions make the ANOVA model less suitable for some real life judgment situations (for example, differential diagnosis from the profile of highly correlated MMPI scale scores), there are many real situations—plus a host of contrived situations—where the restrictions are not too severe. In some of these cases, it is possible to use a completely crossed experimental design (all possible combinations of each of the cue levels), provided that neither the number of cues nor the number of levels per cue is too large.

When judgments are analyzed in terms of the ANOVA model, a significant main effect for cue X_1 implies that the judge's responses varied systematically with X_1 as the levels of the other cues

were held constant. Provided sufficient levels of the factor were included in the design, the main effect may be divided into effects due to linear, quadratic, and cubic (i.e., curvilinear) trends. Similarly, a significant interaction between cues X_1 and X_2 implies that the judge was responding to particular patterns of those cues (i.e., the configural effect of variation of cue X_1 upon judgment differed as a function of the corresponding level taken by cue X_2). Moreover, it is possible to calculate an index of the importance of individual or configural use of a cue, relative to the importance of other cues. The index ω^2, described by Hays (1963), provides a rough estimate of the proportion of the total variation in a person's judgments which can be predicted from a knowledge of the particular levels of a given cue or of a configural pattern of cues. An alternative technique for expressing the extent of configural cue usage in the judgment process has been proposed by Hammond et al. (1964).

THE SEARCH FOR CONFIGURAL JUDGES

With this technical digression now out of the way, let us return to some empirical studies of the clinical judgment process. You will recall that while our research strategy forces us to begin with a simple linear additive model, this model should soon give way to more complex ones, as configural and curvilinear terms are added to fit the judgmental processes of each particular judge. However, in study after study our initial hopes went unrealized; the accuracy of the linear model was almost always at approximately the same level as the reliability of the judgments themselves, and—no doubt because of this—the introduction of more complex terms into the basic equation rarely served to significantly increase the cross-validity of the new model. Hammond and Summers (1965) have reviewed a series of studies in which the same general finding has emerged: for a number of different judgment tasks and across a considerable range of judges, the simple linear model appeared to characterize quite adequately the judgmental processes involved—in spite of the reports of the judges that they were using cues in a highly configural manner.

Three possible hypotheses spring to mind to account for these findings: (*a*) human judges behave in fact remarkably like linear data processors, but somehow they believe that they are more complex than they really are; (*b*) human judges behave in fact in a rather configural fashion, but the power of the linear regression model is so great that it serves to obscure the real configural processes in judgment; [2] (*c*) human judges behave in fact in a decidedly linear fashion on most judgmental tasks (their reports notwithstanding), but for some kinds of tasks they use more complex judgmental processes.

During the past few years, my colleagues at Oregon Research Institute and I have been systematically experimenting to see which of these three hypotheses is the most plausible. Our general goals have been (*a*) to discover and use some alternative judgmental models which allow more rigorous checks on the process of cue utilization (e.g., Hoffman, 1967), and (*b*) to discover and study some new judgmental tasks—tasks where configural cue utilization is most likely to be necessary for making accurate inferences and therefore where configural judgmental processes are most likely to be found. The remainder of this paper will focus primarily on our efforts to achieve this latter goal.

The search for inherently configural tasks has led to three major fields: physical medicine, psychiatry, and clinical psychology. Subject matter experts in each of these fields were consulted in search of diagnostic decisions of a clearly configural nature, and three judgmental tasks—one from each field—were finally selected for intensive study. Experienced medical gastroenterologists chose the first purportedly highly configural task: the differential diagnosis of a benign versus malignant gastric ulcer from the signs which are visible on a stomach X ray. The staff of a large psychiatric hospital provided a second important clinical judgment task: the decision to permit temporary liberty for a chronic patient committed to a psychiatric hospital. And finally, Paul Meehl (1959) chose the third purportedly highly configural judgment task: the differential diagnosis of psychosis versus neurosis from a patient's MMPI profile.

Let us begin with the problem from medicine,

[2] In the same way, a straight line can provide an excellent approximation of many curved lines, exemplified by the fact that we often use a straight line to navigate between two cities even though the real route is along a curved arc. For an excellent discussion of this point, see Ghiselli (1964, pp. 3–7). For a more complete treatment of this topic in judgment research, see Hoffman (1968).

the diagnosis of benign versus malignant gastric ulcers (Hoffman et al., 1968). Physicians have assured us that there are seven major signs which can be seen in X rays of gastric ulcer patients and that this diagnostic problem can be assessed only by the configural (interactive) use of these seven cues. The seven cues are either present or absent in a given X ray, and one of the cues can only occur when another one is present; consequently, two of the seven cues can be combined into one variable having three levels, while each of the other five cues has two levels (absent versus present). There are thus 3×2^5, or 96, possible combinations of all seven cues. Nine judges, six experienced radiologists and three radiology residents, were asked to make differential diagnoses for 192 presumably real, but actually hypothetical, patients (two administrations of each of the 96 possible cue combinations). The judges made their diagnoses on a seven-point scale, from "definitely benign," through "uncertain," to "definitely malignant." The inferences of each judge were analyzed by the ANOVA model to ascertain the proportion of the variance in his diagnoses associated with each of the 6 possible main effects (i.e., linear use of the cues), each of the 15 possible two-way interactions, each of the 20 possible three-way interactions, each of the 15 possible four-way interactions, each of the 6 possible five-way interactions, and the 1 six-way interaction.

The major finding was that the largest of the 57 possible interactions, for the most configural judge, accounted for but 3% of the variance of his responses. In the investigators' own words (Hoffman et al., 1968):

> the largest main effect usually accounted for 10 to 40 times as much of the total variance in the judgments as the largest interaction. On the average, roughly 90% of a judge's reliable variation of response could be predicted by a simple formula combining only individual symptoms in an additive fashion and completely ignoring interactions [pp. 343–344].

> it should be noted that the performance of the judges in this study was rather adequately accounted for in terms of linear effects, in spite of the fact that a deliberate attempt had been made to select a task in which persons would combine cues configurally [p. 347].

While these findings may be disheartening to judgment researchers, another finding could be more generally terrifying. When one examines the degree of agreement between physicians for this diagnostic problem, these interjudge correlations were distressingly low. Of the 36 coefficients of consensus, 3 were negative—the median correlation being only .38. The intrajudge test-retest correlations were reasonably high (ranging .60–.92, *Mdn* = .80), and the task itself was certainly not seen as an impossibly difficult one. Yet, these findings suggest that diagnostic agreement in clinical medicine may not be much greater than that found in clinical psychology—some food for thought during your next visit to the family doctor.[3]

Let us turn now to some ANOVA analyses of another judgmental task, the decision whether or not to grant temporary liberty to a psychiatric patient (Rorer, Hoffman, Dickman, & Slovic, 1967). The six presumably most relevant variables for making this decision were used in this study. With two levels of each variable (e.g., "Does the patient have a problem with drinking?" "Yes" versus "No"), there were thus 2^6, or 64, possible cue combinations. Twenty-four members of the professional staff of a psychiatric hospital—6 physicians, 12 nurses, 3 clinical psychologists, and 3 psychiatric social workers—served as judges. Each of them decided whether 128 presumably real, but actually hypothetical, patients (two administrations of each of the 64 possible cue configurations) should be granted the privilege of leaving the hospital for 8 hours on a weekend. Again, as in the previous study, the judgments from each judge were analyzed individually to ascertain the proportion of his response variance which was associated with each of the six main effects and each of the possible two-way, three-way, four-way, five-way, and six-way interactions.

The results were, unfortunately, remarkably similar to those from the previous study. On the average, less than 2% of the variance of these judgments was associated with the largest interaction term, these percentages ranging across the 24 judges from virtually zero to less than 6%. And again, one of the most striking findings was the great diversity—the startling lack of interjudge agreement—among clinicians for this judgment task.

Let us now turn to the third purportedly configural judgment task, the differential diagnosis of neurotic from psychotic patients by means of their MMPI profiles. Paul Meehl (1959) initially

[3] For some intriguing corroborative evidence concerning this seemingly subversive statement see Garland (1959, 1960).

focused research on this task on the grounds that: "the differences between psychotic and neurotic profiles are considered in MMPI lore to be highly configural in character, so that an atomistic treatment by combining single scales linearly should theoretically be a very poor substitute for a configural approach [p. 104]." Meehl collected 861 MMPI profiles from seven hospitals and clinics throughout the United States; each of these profiles was produced from the MMPI responses of a psychiatric patient who had been diagnosed by the psychiatric staff as being rather clearly either psychotic or neurotic—the total sample containing approximately equal numbers of both diagnostic groups. Twenty-nine clinicians (13 PhD clinical psychologists, plus 16 advanced graduate students in clinical psychology) attempted to diagnose each of the 861 patients from their MMPI profiles; the 29 judges rated each profile on an 11-step forced-normal distribution from least to most psychotic. After making some preliminary comparisons of the validity of the clinicians' judgments with the validities achieved by various actuarial techniques (Meehl, 1959), Meehl generously turned over these valuable data to Oregon Research Institute for further analyses.

An extensive investigation of the validity of the clinicians' judgments, as compared to that of numerous MMPI signs and indexes, has already been published (Goldberg, 1965). As in many previous judgment studies, accuracy on this task was not associated with the amount of professional experience of the judge; the average PhD psychologist achieved a validity coefficient identical to that of the average graduate student. Moreover, an unweighted composite of five MMPI scale scores ($L + Pa + Sc - Hy - Pt$) achieved a validity coefficient ($r = .44$) greater than that of the average judge ($r = .28$), greater than that of the pooled ratings of all 29 clinicians ($r = .35$), and even greater than that of the single most accurate judge ($r = .39$). Moreover, I recently discovered a moderator for the above index, namely another unweighted linear composite ($D + Pd + Sc - F - Hs - Pa$); when some 1,248 patients were divided into three subsamples on the basis of their scores on the moderator variable (i.e., high versus medium versus low moderator scores), the validity coefficients for the three groups were .27, .42, and .58, respectively.

When one turns from analyses of validity to those focused on the judgment process, conclusions become more difficult. For unlike the two previously described judgmental tasks, this one has some serious limitations. Two of these problems are inherent to the task, while a third stems from Meehl's (1959) experimental procedures: (*a*) the 11 MMPI scale scores presented to the judges are not orthogonal (the 55 intercorrelations range up to almost .80—for example, *Hs* versus *Hy*—8 of them being higher than .50); (*b*) each scale score is a relatively continuous variable covering a considerable range of scale values; and (*c*) the 29 clinicians in Meehl's original study judged each of the 861 profiles only once (i.e., no repeated profiles were presented). For reasons *a* and *b* the ANOVA model is inappropriate for these data, and for reason *c* it is impossible to ascertain to what extent various judgment models approach the reliability of the judges' responses, since these reliability values are not known. Nonetheless, it has been possible to make some estimates about the nature of these clinical judgments (Wiggins & Hoffman, 1968).

Wiggins and Hoffman (1968) compared—as representations of the cognitive processes of each of the 29 clinicians—the following three models: (*a*) the standard linear regression model; (*b*) a quadratic model, which added to the first model all squared terms (e.g., X_1^2) and cross-product terms (e.g., $X_1 \cdot X_2$); and (*c*) a "sign" model, which included a set of 70 MMPI diagnostic signs from the psychometric literature. While Wiggins and Hoffman (1968) interpreted their findings as indicating that, for some judges, one of the nonlinear models provided a slightly better representation of their judgments than the linear model, nonetheless, the most overwhelming finding from this study was how much of the variance in clinicians' judgments could be represented by the linear model. For example, if one compares the judgment correlations produced by the linear model with those from each of the two configural models (see Wiggins & Hoffman, 1968, Table 3), one finds that (*a*) the linear model was equal to, or superior to, the quadratic model for 23 of the 29 judges (and at best, for the most configural judge, the quadratic model produced a correlation with his judgments which was only .03 greater than that of the linear model); and (*b*) the linear model was equal to, or superior to, the sign model for 17 judges (the superiority of the sign model being but .04 for

the single most configural judge). In the authors' own words,

> A note of caution should be added to the discussion of differences between linear and configural judges. Though the differences appear reliable, their magnitude is not large; the judgments of even the most seemingly configural clinicians can often be estimated with good precision by a linear model [pp. 76–77].

Once again, the linear model provided an excellent representation of the judgments of most of these clinicians, even for a task which they believed to be a highly configural one.

The point of this discussion is *not* to assert that clinicians, including the many clinicians studied in the experiments already described, cannot and do not use cue relationships more complex than simple linear ones. In the first two of these studies, for example, there were one or more statistically significant interactions in the judgment models of at least some of the clinicians, and in the third study there were clinicians whose judgments were at least slightly better represented by a model other than the linear one. Moreover, Paul Slovic (1968) has recently demonstrated that the judgments of each of two professional stockbrokers, asked to predict future stock prices from 11 dichotomized indexes, showed significant interactions which are explainable in terms of the theoretical orientations of the brokers themselves. And, in a number of other judgmental studies (e.g., Slovic, 1966), evidence of configural cue utilization has been uncovered. Clearly, clinical judgments *can* involve the configural utilization of cues. What are, then, the implications from these judgmental investigations?

First of all, it is important to realize that the very power of the linear regression model to predict observations generated by a large class of nonlinear processes can serve to obscure our understanding of all but the more gross types of configural judgments. Yntema and Torgerson (1961) and Rorer (1967) have both demonstrated rather dramatically how observations generated by nonlinear processes can become interpreted as linear ones when analyzed by standard regression and ANOVA methodology; Hoffman et al. (1968) and Hoffman (1968) provide an excellent discussion of this problem as it applies to the judgment process. If we return once again to the three competing hypotheses which provided the framework for launching these judgmental investigations, I would now assert that our original hypothesis (*b*)—that judges can process information in a configural fashion, but that the general linear model is powerful enough to reproduce most of these judgments with very small error—is, at this point, certainly the most compelling one.

Consequently, if one's sole purpose is to reproduce the responses of most clinical judges, then a simple linear model will normally permit the reproduction of 90%–100% of their reliable judgmental variance, probably in most—if not all—clinical judgment tasks. While Meehl (1959) has suggested that one potential superiority of the clinician over the actuary lies in the human's ability to process cues in a configural fashion, it is important to realize that this is neither an inherent advantage of the human judge (i.e., the actuary can include nonlinear terms in his equations), nor is this attribute—in any case—likely to be the clinician's "ace in the hole." If the clinician does have a long suit—and the numerous clinical versus statistical studies have not yet demonstrated that he has—it is extremely unlikely that it will stem from his alleged ability to process information in a complex configural manner.

LEARNING CLINICAL INFERENCE

If "clinical wisdom" results in linearly reproducible judgments of rather low validity, it becomes sensible to ask whether these judgments could not be improved through training. Leonard G. Rorer and I reasoned that the major cause of the low validity coefficients reported for the judgments of practicing clinicians is the fact that in most, if not all, clinical settings there is no realistic opportunity for the clinician to improve his predictive accuracy. For learning to occur, some systematic feedback regarding the accuracy of the judgmental response must be linked to the particular cue configuration which led the clinician to make that judgment. But, in clinical practice feedback is virtually nonexistent, and in the relatively rare cases when feedback does occur the long interval of time which elapses between the prediction and the feedback serves to ensure that the initial cue configuration leading to the prediction has disappeared from the clinician's memory. As an example, say a clinician infers the prognosis "high suicide potential" from the MMPI profile of Patient A and writes in his report a statement like "Patient A has a high risk of committing suicide and therefore should be carefully watched." In

most cases Patient A eventually returns to the community or moves to another hospital, and the clinician does not know whether the patient ever attempted suicide (accurate inference) or not (inaccurate inference). And if in 3 years the clinician happens to read in the newspaper that Patient A committed suicide, he is unlikely to be able to recall the particular MMPI profile configuration which initially led to this (successful) prediction, with the result that the "cue configuration → suicide inference" link is in no way strengthened.[4]

What is necessary for clinical inference to be learned, Rorer and I reasoned, is that the clinician obtain immediate feedback concerning the accuracy of his judgments—ideally feedback which occurs after the judgmental response has been formulated but before the removal of the cue configuration which led the clinician to that response. Moreover, if the cues are related to the criterion in some curvilinear and/or configural manner, then the clinician should be able to learn these more complex relationships, modify his own judgmental processes to incorporate such configural elements, and thereby begin to make judgments for which the best representation is a more complex model than the linear one.

To test this hypothesis, Rorer and I designed a study in which judges were given immediate feedback on the same task previously described, namely, the differential diagnosis of psychosis versus neurosis from MMPI profiles. Three groups of judges —termed expert, middle, and naive—were studied. The expert group was composed of three clinical psychologists who had had extensive MMPI experience. The naive group was composed of 10 nonpsychologists who were unfamiliar with the MMPI and who were told only that their task was to learn to differentiate "N" from "P" profiles. The middle group was composed of 10 psychology graduate students who had at least a passing familiarity with the MMPI and some idea of the difference between a neurotic and a psychotic patient.

The judges received alternate weeks of training and testing. Five sets of 60 training profiles, each of which contained the criterion diagnosis on the back of the profile sheet, were assembled from 300 profiles drawn at random from one hospital sample. Thirty of these profiles in each set were repeated so that there was a total of 90 profiles in each training set. Ten testing sets were constructed, each set including profiles from a different clinical sample (one of which was the same as that used in the training set). Whenever possible, the testing set was composed of 100 profiles, 50 of which were then repeated, so that there was typically a total of 150 profiles in each testing set. Judges were instructed to diagnose the profiles from one set per day for 5 days per week. The judges were asked to classify each profile in turn and also to indicate their confidence in each of their judgments.

While all of the analyses of these data have not been completed, some preliminary results are available (Goldberg & Rorer, 1965; Rorer & Slovic, 1966). Let us first look at the levels of accuracy achieved after 9 weeks of daily training and 8 alternate weeks of daily testing. By this point, the judges had already received 90 training profiles per day (450 per week) for a total of over 4,000 training profiles (each followed by immediate feedback), plus another 6,000 testing profiles—over 10,000 profiles in all. But, while all three groups of judges manifested some learning on the training profiles, only the naive group showed *any* generalization of this training in improving their accuracy on the testing profiles. The average naive judge was correct about 52% of the time at the beginning, and after 17 weeks he had increased his accuracy to about 58%. The middle and expert judges were virtually indistinguishable, both groups achieving an average accuracy percentage around 65% at the beginning of training and the same figure after 17 weeks. Thus, even after 4,000 training profiles, the average accuracy percentage for the naive judges was still substantially below that manifested by the expert and middle judges. For the expert and middle judges, training on this task turned out to almost completely sample specific; there was virtually no cross-sample generalization of learning as a result of intensive training on over 4,000 MMPI profiles!

Faced with these startling findings, a number of experimental variations in the training procedures were introduced in an effort to increase judgmental accuracy. Two naive and two middle subjects

[4] B. F. Skinner (1968) has made much the same point in rebutting the belief in the accumulated wisdom of the classroom teacher: "It is actually very difficult for teachers to profit from experience. They almost never learn about their long-term successes or failures, and their short-term effects are not easily traced to the practices from which they presumably arose [pp. 112–113]."

were assigned to each of the following five subgroups:

Standard condition. These subjects continued what they had been doing all along. They were therefore a control group for the other four experimental variations.

Group training. Two subjects worked together and agreed on a response. There was one naive pair and one middle pair. They were tested both individually and as a pair.

Generalization training. Subjects in the generalization training group were given training on previously unused profiles from those installations on which the judges had achieved their poorest results.

Formula training. These subjects were given the formula $(L + Pa + Sc - Hy - Pt)$ and told that it would increase the accuracy of their judgments. They were encouraged to use the formula as a guide to indicate the scales to which they might profitably attend.

Value training. The judges in this group, including all three experts, were given the numerical value of the formula for each profile and the optimum cutting score. They were told that this formula would achieve approximately 70% accuracy and that it would be more accurate for extreme values than for values close to the cutting score. They were free simply to report the formula diagnosis for every profile (a procedure which in every case would have allowed them to increase their judgmental accuracy), though they were encouraged to try to find ways in which they might improve on the formula decision.

After 8 more weeks (4 of training and 4 of testing), we found that those groups given value training (including all of the experts) had, on the average, increased their accuracy to a bit below 70% correct. But, none of the other experimental groups showed any substantial learning. Giving judges the optimal formula (formula training) resulted in a rapid increase in diagnostic accuracy (especially for the naive group), but this effect gradually wore away over time. By the end of the study the formula training groups were again achieving approximately the same level of accuracy as the standard training control groups. On the other hand, giving judges the actual values of the optimal formula for each profile (value training) did result in a stable increase in diagnostic accuracy, though the accuracy of these judges' diagnoses was not as high as would have been achieved by simply using the formula itself.

The thousands of judgments collected during those months of intensive training should yield many more nuggets than these few which I have scraped off the top. But, I doubt whether the conclusions we can already draw will have to be drastically changed. It now appears that our initial formulation of the problem of learning clinical inference was far too simple—that a good deal more than outcome feedback is necessary for judges to learn a task as difficult as the present one. The research of Chapman and Chapman (1967) serves to reinforce this belief by providing an even more stunning example of the pitfalls of relying solely upon feedback to improve the accuracy of clinical inferences.

In what is perhaps the most ingenious series of studies of clinical judgment ever carried out, Chapman and Chapman (1967) have demonstrated how prior expectations of the relationships between cues and criteria can lead to faulty observation and inference, even under seemingly excellent conditions for learning. The Chapmans exposed subjects to human figure drawings, each of which was paired with two criterion statements concerning the characteristics of the patients who allegedly drew the figures. Though these training materials were constructed so that there was no relationship between the cues and the criterion statements, most subjects erroneously "learned" the cue-criterion links which they had expected to see. In fact, the "illusory correlation" phenomena demonstrated by the Chapmans was such a powerful one that many subjects trained on materials where the cue-criterion relationships were constructed to be the opposite of those expected still persisted in "learning" the erroneous relationships! For further documentation of this pervasive source of bias in the learning of clinical (and other) types of inference see Chapman (1967).

The intriguing research of the Chapmans illustrates the ease with which one can "learn" relationships which do *not* exist. Our own MMPI learning research, plus that of others (e.g., Crow, 1957; Sechrest, Gallimore, & Hersch, 1967; Soskin, 1954), demonstrates the problems which can be encountered in learning those relationships which *do* exist. What now seems clear is that at least three conditions—all of which are are missing from the typical clinical setting—must hold if more complex

clinical inferences are to be learned. First of all, some form of feedback (e.g., Skinner, 1968; Todd & Hammond, 1965) is a necessary, though not necessarily a sufficient, condition for learning to occur. Second, at least for problems of the complexity of many encountered in clinical practice, it may be necessary to be able to disturb the natural sequence of cue presentations—to rearrange the order of cases—so that one's hypotheses can be immediately verified or discounted. It does little good to formulate a rule for profile Type A, only to have to wait for another 100 profiles before an additional manifestation of Type A appears; what one must do is group together all Type A profiles in order to be able to verify one's initial inference. In the clinical setting this means studying those patients who manifest some particular cue configuration of interest, rather than taking patients as they come in the door. Finally, as the Chapmans' (1967) clever research so vividly demonstrates, it may often be necessary to *tally* the accuracy of one's hypotheses, thereby letting some variant of a paper-and-pencil boxscore substitute for the more ephemeral storage capacities of the unaided human brain.

But, what do we call that process which is characterized by a disruption of the naturally occuring order of observations, plus immediate feedback on cue-criterion links, followed by some concrete form of tallying the accuracy of one's hypotheses? We call it RESEARCH.

REFERENCES

BORKE, H., & FISKE, D. W. Factors influencing the prediction of behavior from a diagnostic interview. *Journal of Consulting Psychology,* 1957, 21, 78–80.

BRODIE, C. M. Clinical prediction of personality traits displayed in specific situations. *Journal of Clinical Psychology,* 1964, 20, 459–461.

BRYAN, J. H., HUNT, W. A., & WALKER, R. E. Reliability of estimating intellectual ability from transcribed interviews. *Journal of Clinical Psychology,* 1966, 22, 360.

CHAPMAN, L. J. Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior,* 1967, 6, 151–155.

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

CROW, W. J. The effect of training upon accuracy and variability in interpersonal perception. *Journal of Abnormal and Social Psychology,* 1957, 55, 355–359.

GARLAND, L. H. Studies of the accuracy of diagnostic procedures. *American Journal of Roentgenology, Radium Therapy, and Nuclear Medicine,* 1959, 82, 25–38.

GARLAND, L. H. The problem of observer error. *Bulletin of the New York Academy of Medicine,* 1960, 36, 569–584.

GHISELLI, E. E. *Theory of psychological measurement.* New York: McGraw-Hill, 1964.

GIEDT, F. H. Comparison of visual, content, and auditory cues in interviewing. *Journal of Consulting Psychology,* 1955, 19, 407–416.

GOLDBERG, L. R. The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender-Gestalt test. *Journal of Consulting Psychology,* 1959, 23, 25–33.

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

GOLDBERG, L. R. Reliability of Peace Corps selection boards: A study of interjudge agreement before and after board discussions. *Journal of Applied Psychology,* 1966, 50, 400–408.

GOLDBERG, L. R., & RORER, L. G. Learning clinical inference: The results of intensive training on clinicians' ability to diagnose psychosis versus neurosis from the MMPI. Paper presented at the meeting of the Western Psychological Association, Honolulu, June 1965.

GOLDBERG, L. R., & WERTS, C. E. The reliability of clinicians' judgments: A multitrait-multimethod approach. *Journal of Consulting Psychology,* 1966, 30, 199–206.

GOLDEN, M. Some effects of combining psychological tests on clinical inferences. *Journal of Consulting Psychology,* 1964, 28, 440–446.

GOUGH, H. G. Clinical versus statistical prediction in psychology. In L. Postman (Ed.), *Psychology in the making.* New York: Knopf, 1962.

GRANT, M., IVES, V., & RANZONI, J. Reliability and validity of judges' ratings of adjustment on the Rorschach. *Psychological Monographs,* 1952, 66(2, Whole No. 334).

GRIGG, A. E. Experience of clinicians, and speech characteristics and statements of clients as variables in clinical judgment. *Journal of Consulting Psychology,* 1958, 22, 315–319.

GROSZ, H. J., & GROSSMAN, K. G. The sources of observer variation and bias in clinical judgments: I. The item of psychiatric history. *Journal of Nervous and Mental Disease,* 1964, 138, 105–113.

GUNDERSON, E. K. E. Determinants of reliability in personality ratings. *Journal of Clinical Psychology,* 1965, 21, 164–169. (a)

GUNDERSON, E. K. E. The reliability of personality ratings under varied assessment conditions. *Journal of Clinical Psychology,* 1965, 21, 161–164. (b)

HAMMOND, K. R., HURSCH, C. J., & TODD, F. J. Analyzing the components of clinical inference. *Psychological Review,* 1964, 71, 438–456.

HAMMOND, K. R., & SUMMERS, D. A. Cognitive dependence on linear and nonlinear cues. *Psychological Review,* 1965, 72, 215–224.

HAYS, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

HILER, E. W., & NESVIG, D. An evaluation of criteria used by clinicians to infer pathology from figure drawings. *Journal of Consulting Psychology,* 1965, 29, 520–529.

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

HOFFMAN, P. J. Non-shrinkable, wrinkle-resistant configural prediction. Paper presented at the meeting of the American Psychological Association, Washington, D. C., September 1967.

HOFFMAN, P. J. Cue-consistency and configurality in human judgment. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

HOFFMAN, P. J., SLOVIC, P., & RORER, L. G. An analysis of variance model for the assessment of configural cue utilization in clinical judgment. *Psychological Bulletin,* 1968, 69, 338–349.

HOLTZMAN, W. H., & SELLS, S. B. Prediction of flying success by clinical analysis of test protocols. *Journal of Abnormal and Social Psychology,* 1954, 49, 485–490.

HOWARD, K. I. The convergent and discriminant validation of ipsative ratings from three projective instruments. *Journal of Clinical Psychology,* 1962, 18, 183–188.

HOWARD, K. I. Ratings of projective test protocols as a function of degree of inference. *Educational and Psychological Measurement,* 1963, 23, 267–275.

HUNT, W. A., & JONES, N. F. Clinical judgment of some aspects of schizophrenic thinking. *Journal of Clinical Psychology,* 1958, 14, 235–239. (a)

HUNT, W. A., & JONES, N. F. The reliability of clinical judgments of asocial tendency. *Journal of Clinical Psychology,* 1958, 14, 233–235. (b)

HUNT, W. A., JONES, N. F., & HUNT, E. B. Reliability of clinical judgment as a function of clinical experience. *Journal of Clinical Psychology,* 1957, 13, 377–378.

HUNT, W. A., & WALKER, R. E. Validity of diagnostic judgment as a function of amount of test information. *Journal of Clinical Psychology,* 1966, 22, 154–155.

HUNT, W. A., WALKER, R. E., & JONES, N. F. The validity of clinical ratings for estimating severity of schizophrenia. *Journal of Clinical Psychology,* 1960, 16, 391–393.

JOHNSTON, R., & MCNEAL, B. F. Statistical versus clinical prediction: Length of neuropsychiatric hospital stay. *Journal of Abnormal Psychology,* 1967, 72, 335–340.

JONES, N. F., JR. The validity of clinical judgments of schizophrenic pathology based on verbal responses to intelligence test items. *Journal of Clinical Psychology,* 1959, 15, 396–400.

KELLY, E. L., & FISKE, D. W. *The prediction of performance in clinical psychology.* Ann Arbor: University of Michigan Press, 1951.

KLEINMUNTZ, B. MMPI decision rules for the identification of college maladjustment: A digital computer approach. *Psychological Monographs,* 1963, 77(14, Whole No. 577). (a)

KLEINMUNTZ, B. Personality test interpretation by digital computer. *Science,* 1963, 139, 416–418. (b)

KLEINMUNTZ, B. Profile analysis revisited: A heuristic approach. *Journal of Counseling Psychology,* 1963, 10, 315–324. (c)

KOSTLAN, A. A method for the empirical study of psychodiagnosis. *Journal of Consulting Psychology,* 1954, 18, 83–88.

LEVY, B. I., & ULMAN, E. Judging psychopathology from painting. *Journal of Abnormal Psychology,* 1967, 72, 182–187.

LITTLE, K. B., & SHNEIDMAN, E. S. Congruencies among interpretations of psychological test and anamnestic data. *Psychological Monographs,* 1959, 73(6, Whole No. 476).

LUFT, J. Implicit hypotheses and clinical predictions. *Journal of Abnormal and Social Psychology,* 1950, 45, 756–760.

LUFT, J. Differences in prediction based on hearing versus reading verbatim clinical interviews. *Journal of Consulting Psychology,* 1951, 15, 115–119.

MARKS, P. A. An assessment of the diagnostic process in a child guidance setting. *Psychological Monographs,* 1961, 75(3, Whole No. 507).

MCARTHUR, C. Analyzing the clinical process. *Journal of Counseling Psychology,* 1954, 1, 203–208.

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

MEEHL, P. E. Wanted—A good cookbook. *American Psychologist,* 1956, 11, 263–272.

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

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

MEEHL, P. E. The cognitive activity of the clinician *American Psychologist,* 1960, 15, 19–27.

OSKAMP, S. The relationship of clinical experience and training methods to several criteria of clinical prediction. *Psychological Monographs,* 1962, 76(28, Whole No. 547).

OSKAMP, S. Overconfidence in case-study judgments. *Journal of Consulting Psychology,* 1965, 29, 261–265.

OSKAMP, S. Clinical judgment from the MMPI: Simple or complex? *Journal of Clinical Psychology,* 1967, 23, 411–415.

PARKER, C. A. As a clinician thinks . . . *Journal of Counseling Psychology,* 1958, 5, 253–262.

PHELAN, J. G. Rationale employed by clinical psychologists in diagnostic judgment. *Journal of Clinical Psychology,* 1964, 20, 454–458.

PHELAN, J. G. Use of matching method in measuring reliability of individual clinician's diagnostic judgment. *Psychological Reports,* 1965, 16, 491–497.

RINGUETTE, E. L., & KENNEDY, T. An experimental study of the double bind hypothesis. *Journal of Abnormal Psychology,* 1966, 71, 136–141.

RORER, L. G. Conditions facilitating discovery of moderators. Paper presented at the meeting of the American Psychological Association, Washington, D. C., September 1967.

RORER, L. G., HOFFMAN, P. J., DICKMAN, H. D., & SLOVIC, P. Configural judgments revealed. *Proceedings of the*

75th Annual Convention of the American Psychological Association, 1967, 2, 195–196.

RORER, L. G., & SLOVIC, P. The measurement of changes in judgmental strategy. *American Psychologist,* 1966, 21, 641–642. (Abstract)

RYBACK, D. Confidence and accuracy as a function of experience in judgment-making in the absence of systematic feedback. *Perceptual and Motor Skills,* 1967, 24, 331–334.

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

SCHAEFFER, R. W. Clinical psychologists' ability to use the Draw-A-Person test as an indicator of personality adjustment. *Journal of Consulting Psychology,* 1964, 28, 383.

SCHWARTZ, M. L. Validity and reliability in clinical judgments of C-V-S protocols as a function of amount of information and diagnostic category. *Psychological Reports,* 1967, 20, 767–774.

SECHREST, L., GALLIMORE, R., & HERSCH, P. D. Feedback and accuracy of clinical predictions. *Journal of Consulting Psychology,* 1967, 31, 1–11.

SILVERMAN, L. H. A *Q*-sort study of the validity of evaluations made from projective techniques. *Psychological Monographs,* 1959, 73(7, Whole No. 477).

SINES, L. K. The relative contribution of four kinds of data to accuracy in personality assessment. *Journal of Consulting Psychology,* 1959, 23, 483–492.

SKINNER, B. F. *The technology of teaching.* New York: Appleton-Century-Crofts, 1968.

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

SLOVIC, P. Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. Paper presented at the meeting of the Western Psychological Association, San Diego, March 1968.

SOSKIN, W. F. Bias in postdiction from projective tests. *Journal of Abnormal and Social Psychology,* 1954, 49, 69–74.

SOSKIN, W. F. Influence of four types of data on diagnostic conceptualization in psychological testing. *Journal of Abnormal and Social Psychology,* 1959, 58, 69–78.

STRICKER, G. Actuarial, naive clinical, and sophisticated clinical prediction of pathology from figure drawings. *Journal of Consulting Psychology,* 1967, 31, 492–494.

TODD, F. J., & HAMMOND, K. R. Differential feedback in two multiple-cue probability learning tasks. *Behavioral Science,* 1965, 10, 429–435.

VANDENBERG, S. G., ROSENZWEIG, N., MOORE, K. R., & DUKAY, A. F. Diagnostic agreements among psychiatrists and "blind" Rorschach raters or the education of an interdisciplinary research team. *Psychological Reports,* 1964, 15, 211–224.

WALLACH, M. S., & SCHOOFF, K. Reliability of degree of disturbance ratings. *Journal of Clinical Psychology,* 1965, 21, 273–275.

WATLEY, D. J. Counselor predictive skill and differential judgments of occupational suitability. *Journal of Counseling Psychology,* 1967, 14, 309–313.

WATSON, C. G. Relationship of distortion to DAP diagnostic accuracy among psychologists at three levels of sophistication. *Journal of Consulting Psychology,* 1967, 31, 142–146.

WEITMAN, M. Some variables related to bias in clinical judgment. *Journal of Clinical Psychology,* 1962, 18, 504–506.

WIGGINS, N., & HOFFMAN, P. J. Three models of clinical judgment. *Journal of Abnormal Psychology,* 1968, 73, 70–77.

WINCH, R. F., & MORE, D. M. Does TAT add information to interviews? Statistical analysis of the increment. *Journal of Clinical Psychology,* 1956, 12, 316–321.

WINSLOW, C. N., & RAPERSAND, I. Postdiction of the outcome of somatic therapy from the Rorschach records of schizophrenic patients. *Journal of Consulting Psychology,* 1964, 28, 243–247.

YNTEMA, D. B., & TORGERSON, W. S. Man-computer cooperation in decisions requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics,* 1961, Vol. HFE-2, No. 1, 20–26.