CLINICAL versus STATISTICAL PREDICTION

A Theoretical Analysis and a Review of the Evidence

Paul E.Meehl

First softcover edition 1996 by Jason Aronson Inc.

Preface to the 1996 Printing

I AM pleased to see this reprinting of my book, first published in 1954 by the University of Minnesota Press as a special consideration for the then Chair of the Psychology Department after the manuscript had been rejected by several publishers who thought it would not sell. When the book went out of print some years ago, it had gone through seven printings and sold 13,200 copies.

This little book made me famous—in some quarters, infamous—overnight; but while almost all of the numerous prizes and awards that my profession has seen fit to bestow upon me mention this among my contributions, the practicing profession and a large segment—perhaps the majority—of academic clinicians either ignore it entirely or attempt to ward off its arguments, analyses, or empirical facts. Thus I am in the unusual position of being socially reinforced for writing something that hardly anybody believes! I have elsewhere tried to explain this resistance to the facts (Meehl, 1986) and have discussed the many objections that have been raised (Grove and Meehl, 1996).

One remarkable phenomenon following the publication of this book was the attribution to me of views and positions which I did not hold and that were in no way implied by what I had said. For example, many psychologists have claimed, formally or informally, that I think "objective psychological tests predict better than clinical interviews." How could anybody misread the text to that extent? I took great pains to make clear (pp. 15–18) the distinction between kind of data and mode of combination, pointing out that all combinations of these two factors are

found, and I illustrated them. I emphasized the confusion that results if one conflates psychometric tests with actuarial modes of combining data, of whatever nature the data may be. How anybody could read the book, even carelessly, and formulate my thesis as pitting psychometric scores against interview or history data is hard to imagine. Not only did I *not* say psychometric tests always predict better, in fact—while we have inadequate data to adjudicate this matter—I do not believe they do. Despite my long-term identification with the Minnesota Multiphasic Personality Inventory (MMPI), if I were asked to diagnose a mental patient and told that I could either have an MMPI profile or conduct a mental status examination, I would prefer the latter.

As a result of this book and articles I published shortly thereafter, numerous studies compared the efficacy of subjective clinical judgment with prediction via mechanical or actuarial methods. Accumulating over the years, they have tended overwhelmingly to come out in the same way as the small number of studies available in 1954. There is now a meta-analysis of studies of the comparative efficacy of clinical judgment and actuarial prediction methods (Grove et al., 2000; a summary is given in Grove and Meehl, 1996). Of 136 research studies, from a wide variety of predictive domains, not more than 5 percent show the clinician's informal predictive procedure to be more accurate than a statistical one. Despite this, clinicians—both practitioners and academics—continue to treat the subject either as nonexistent, or to misformulate it, or, most commonly, to equivocate, saying "there is a good deal to be said on both sides." The arguments offered against my thesis were, with three or four interesting exceptions, either uninformed or irrational. Objections I had anticipated in the book continued to be offered, and in such a way that an unwary reader would think that they had never occurred to me.

I have also heard (not, I should say, seen in print) various ad hominem complaints. One of my intellectual heroes, Sir Karl Popper, in his intellectual autobiography (Schilpp, 1974, pp. 963–974) talks about the "Popper legend," which he then contrasts with the "real Popper." I shall borrow

his technique and contrast some legends, as they have come to me by word of mouth or correspondence, with the fact in each case.

Legend: "Meehl is not a real clinician; he doesn't know what it's like to be on the firing line, having professional responsibility to diagnose and treat mentally disturbed people." *Fact*: I treated my first patient in 1942 (Meehl, 1989), and, except for short periods (e.g., an interruption when I was APA president), I was engaged continuously in the practice of psychotherapy until several years after my retirement from the University of Minnesota. For half a century I earned a portion of my income by private practice. I served as acting chief clinical psychologist in the VA hospital in Minneapolis and for several years was consultant to the VA Mental Hygiene Clinic, supervising trainees in diagnosis and psychotherapy. Although it had no financial benefit for me, I took the trouble to be boarded by the American Board of Professional Psychology (and was the first clinician to serve on ABPP who had taken the exam myself, rather than being grandfathered). During the 1950s, I was seeing patients from eight to fifteen hours a week in order to accumulate clinical experience hours to meet the ABPP requirements. It is simply false to say that I am "a pure lab and library" psychologist who doesn't know what it is like to work with "real-life patients."

Legend: "Meehl doesn't have a psychodynamic orientation." *Fact*: I came to psychology because of my fascination with Freud (Meehl, 1989). His picture hangs in my office. I benefited greatly from my analysis (Meehl, 1989), and I did control cases under the supervision of Rado-trained analyst Bernard C. Glueck, M.D. During my first decade of therapeutic practice my technique was fairly "classical" (insofar as that term is definable, cf. Meehl, 1983, 1993, 1995). Aside from the fact that the legend is false, the connection between being "psychodynamic" and being "nonstatistical" is in no way a logical relation implied by the character of the subject matter; rather, the correlation between the two in most psychologists is a consequence of certain historical accidents.

Legend: "If Meehl knew more about projective methods, he wouldn't take such a strong actuarial position." *Fact*: When I was a graduate student and for several years after obtaining my Ph.D., I gave the Thematic Apperception Test (TAT) to private psychotherapy patients for a psychiatric colleague. I early gave up the standard scoring of the Murray needs (as I think many practitioners do) and used the TAT stories essentially as one uses the material in an analytic hour. Despite my mentor Starke R. Hathaway's bias against projectives, in the late 1940s I took first Beck's and then Klopfer's Rorschach courses. (My first stage performance as a "blind Rorschacher" upon my return from Klopfer's course—Dr. Hathaway had put me on the spot to report the Rorschach in Saturday morning's grand rounds—was a roaring success. Relying on two "eye" responses, a bad O^{-}, and the "witches' cauldron" on card IX in an otherwise healthy record, I correctly diagnosed a patient [who had presented as hypochondriacal] as paranoid. It had taken the staff two weeks to discern this, especially as the patient had a normal MMPI with acceptable L and K.) I continued to give Rorschachs and TATs for some years, but finally concluded that either I was not very good at it or, for some unknown reason, the incremental information I was getting was not worth the additional time and cost to the patients. Although I have not given any projectives for many years, the legend that I am ignorant of them is false.

Legend: "The Minnesota department when Meehl was a student was strongly behavioristic, so he would naturally espouse a statistical emphasis." The social fact about the department is essentially correct, but the inference is unsound. Our archbehaviorist, B. F. Skinner, disliked psychologists' excessive reliance upon statistics. In his great book, *The Behavior of Organisms* (1938), there is not one statistical significance test; he maintained that proper control of the subject matter would result in smooth curves for individual organisms, and there would be no need to compute statistics. Our eminent applied psychologists Donald G. Paterson and Starke R. Hathaway emphasized quantification, and the latter was critical of psychoanalysis as a helping technique and as a theory of the mind. Although the imputed connection between

Minnesota behaviorism and emphasis on actuarial methods is a sociological mistake, there is in the legend an important element of psychological truth as regards my motivation in writing the book at all. The Minnesota animal learning theorists' behaviorism and the applied (counseling and clinical) psychologists' emphasis on psychometrics and statistical data collection shared what may be loosely described as a tough-minded or hard-headed emphasis on *objectivity*. This combination of social forces tended to engender considerable cognitive dissonance in a student who, like myself, came to the field via a passionate interest in psychoanalysis (Meehl, 1989). I think one reason for the book's success is that I was highly motivated to "get my ideas clear" about these methodologically different intellectual traditions. It seemed clear that no rational, fair-minded, intellectually alive person could simply brush off either of them, and that to espouse one of them to the complete exclusion of the other would be to settle for a deficient understanding of human motivetions and behavior. Nobody on the faculty was very helpful in parsing and then reintegrating these traditions. To a student less autonomous than I was, such a clash between the inner forces and intellectual history that brought me to the field on the one side and the strongly, articulately defended position of my esteemed academic mentors on the other would have produced misery, but for me it was an intellectual challenge. There is a deep sense in which that little book was written not for the profession, but for myself and a few of my Minnesota peer group who were trying to get clear about a deep and important question. I think I wrote a pretty good book because I genuinely understood and emotionally empathized with "both sides." (Writing years after the first edition of his *Epistle to the Romans*, Karl Barth said he felt like somebody who, steadying himself walking up a dark staircase, grasps a bell rope for support and finds that he has unintentionally stirred up the countryside! I remember thinking it certainly applied to me when I first read that remark.)

Finally, a legend which I have found more irksome and also harder to understand than any of the others. *Legend*: "Meehl made some required

bows to the clinician, and he displayed a superficial fair-mindedness, but his book is actually animated by anti-clinical prejudices." *Fact*: The easiest way to see whether I am grinding one axe or another is to do a content analysis. The 1954 Preface points out that in my list of honorific and pejorative terms (p. 4), culled from years of listening to conversations, lectures, and reading, the pro- and anti-clinician tallies are exactly equal, with no conscious intent on my part to make them so. Of course, if it is held that arguments I offered or evidence I adduced are unsoundly reasoned, that cannot be rebutted by an actuarial tally but only by examining arguments on the merits. The counterarguments, purported refutations of my arguments, or the dismissal of the empirical data are examined in detail in Grove and Meehl (1996). I have, however, done an informal content analysis by examining the content of the first complete paragraph on every seventh page (excluding Chapter 8, which summarizes the then available empirical comparisons), to get a total of 20 paragraphs. They can be classified as pro-clinical, pro-statistical, neutral, or essentially neutral (clarifying) but having implications that perhaps lean one way or the other. The tally shows that there are six neutral (clarifying, distinguishing) passages, such as pointing out the difference between a configural relation and a nonlinear but nonconfigural or atomistic relation—a clarification that does not harm or benefit either party; two clarifying with a pro-clinical implication; two clarifying passages with a pro-actuarial implication; two pro-actuarial as an implication from a clarification; seven plain pro-clinical; and one plain pro-actuarial. The only way I can explain such a distribution being perceived as mainly grinding a pro-actuarial, anti-clinical axe is that such critics never read the book. If one considers whole chapters, Chapters 1, 2, 7, and 9 are neutral, clarifying about concepts; Chapters 3 and 10 are mainly pro-statistical; Chapters 4, 5, and 6 are mainly pro-clinical; and Chapter 8 presents the empirical comparisons. Some readers seem to think that the main point of the book was the empirical comparisons, but I have left that chapter out of my tallies because the facts fell as they would, and I don't

consider myself responsible for how the studies (none of them conducted by me or by my students) came out. Since Grove's meta-analysis of all 136 interpretable studies now available shows the same clear trend, I see no reason for including that chapter in my content analysis. Of course, somebody might say that the greater emphasis on the clinician's unique contribution, mental processes that are not easily subsumable merely as a second-rate, less accurate linear regression equation, is the result of an ideological reaction formation on my part, but that is a kind of psychodynamic game that one simply cannot win, no matter how the content might be distributed.

Strangely, in writers purporting to be functioning in a science or the applications thereof, there is a tendency to write as if this were some kind of a whimsical personal preference, that one could have warm feelings for statistics or warm feelings to the contrary, and behave accordingly. Such an attitude is not only irrational and unscientific, it is unethical. When one is dealing with human lives and life opportunities, it is immoral to adopt a mode of decision-making which has been demonstrated repeatedly to be either inferior in success rate or, when equal, costlier to the client or the taxpayer. I do not expect clinicians to be convinced by 136 studies if they were not persuaded, at least until further notice, by the twenty described by me or by fifty-five or sixty-five or ninety-five described by others at various times. "A man convinced against his will is of the same opinion still." I have, over the years, developed a certain Buddhistic detachment about this. But however long it takes—I am sure it will be after my death—for psychologists to accept the finding, it will be an interesting episode for study by future historians of science.

References

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

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical vs. mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19-30.

Meehl, P. E. (1983). Subjectivity in psychoanalytic inference: The nagging persistence of Wilhelm Fliess's Achensee question. In J. Earman (Ed), *Minnesota studies in the philosophy of science*: Vol. 10. *Testing scientific theories*. Reprinted in *Psychoanalysis and Contemporary Thought*, 1994, *17*, 3-82.

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

Meehl, P. E. (1989). Autobiography. In G. Lindzey (Ed.), *History of psychology in autobiography*, Vol. VIII (pp. 337–389). Stanford, CA: Stanford University Press.

Meehl, P. E. (1993). If Freud could define psychoanalysis, why can't ABPP? *Psychoanalysis and Contemporary Thought*, *16*, 299-326.

Meehl, P. E. (1995). [Psychoanalysis is not yet a science: Comment on Shevrin.] *Journal of the American Psychoanalytic Association*, *43*, 1015-1023.

Schilpp, P. A. (Ed.) (1974). *The philosophy of Karl Popper* (Vol. 2). LaSalle, IL: Open Court.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton–Century.

Preface

THIS monograph is an expansion of lectures given in the years 1947-1950 to graduate colloquia at the universities of Chicago, Iowa, and Wisconsin, and of a lecture series delivered to staff and trainees at the Veterans Administration Mental Hygiene Clinic at Ft. Snelling, Minnesota. I am indebted to the staff and graduate students who attended these lectures for criticisms and suggestions which have contributed materially to the present form of the argument. Conversations and correspondence with Drs. E. S. Bordin, Robert C. Challman, Lee J. Cronbach, Herbert Feigl, James J. Jenkins, E. J. Shoben, Donald E. Super, and Joseph Zubin have also been very illuminating. Although I am compelled to disagree with some of his *theoretical* formulations of clinical method, the basic approach and clinical philosophy of my teacher and colleague Dr. Starke R. Hathaway are inextricably involved in most of what follows. I wish to thank Dr. Morris S. Viteles and Dr. Robert Y. Walker for their kindness in making available to me their personal copies of the out-of-print Dunlap and Wantman study (reference **38**). Dr. Richard Melton read the manuscript while working on his own thesis (**73**) and called my attention to two additional studies by Borden (**15**) and Hamlin (**48**). Dr. Albert Rosen located the Blenkner (**12**) paper. I am indebted to Dr. Charles Bird, Dr. William Schofield, and Dr. Charles Halbower for valuable editorial criticisms.

The manuscript of this book has been in substantially its present form since 1950, and I have not modified it as a result of nonempirical writings

published since then. Because of the special role of T. R. Sarbin's contributions on the topic under consideration, I especially urge the reader to consult Sarbin and Taft's *An Essay on Inference in the Psychological Sciences* (88), which treats, in much greater detail and with citation of empirical studies, some of the matters I raise speculatively in Chapter 7. But since, as I understand it, Sarbin's view on the main question is still fundamentally the same as that which he expressed in earlier publications, I have not attempted to incorporate the Sarbin-Tart monograph into my discussion.

Perhaps a general remark in clarification of my own position is in order. Students in my class in clinical psychology have often reacted to the lectures on this topic as to a projective technique, complaining that I was biased either for or against statistics (or the clinician), depending mainly on where the student himself stood! This I have, of course, found very reassuring. One clinical student suggested that I tally the pro-con ratio for the list of honorific and derogatory adjectives in Chapter 1 (page 4), and the reader will discover that this unedited sample of my verbal behavior puts my bias squarely at the midline. The style and sequence of the paper reflect my own ambivalence and real puzzlement, and I have deliberately left the document in this discursive form to retain the flavor of the mental conflict that besets most of us who do clinical work but try to be scientists. I have read and heard too many rapid-fire, once-over-lightly "resolutions" of this controversy to aim at contributing another such. The thing is just not that simple. I was therefore not surprised to discover that the same sections which one reader finds obvious and overelaborated, another singles out as especially useful for his particular difficulties. My thesis in a nutshell:

"There is no convincing reason to assume that explicitly formalized mathematical rules and the clinician's creativity are equally suited for any given kind of task, or that their comparative effectiveness is the same for different tasks. Current clinical practice should be much more critically examined with this in mind than it has been."

Preface to the 1996 Printing

It is my personal hunch, not proved by the presented data or strongly argued in the text, that a very considerable fraction of clinical time is being irrationally expended in the attempt to do, by dynamic formulations and staff conferences, selective and prognostic jobs that could be done more efficiently, in a small fraction of the clinical time, and by less skilled and lower paid personnel through the systematic and persistent cultivation of complex (but still clerical) statistical methods. This would free the skilled clinician for therapy and research, for both of which skilled time is so sorely needed.

Since I am myself a hybrid working clinician and rat psychologist, I feel that I am in a favorable position to see somewhat objectively, and I do not honestly think I am on either side of this debate. But I hope the reader will agree with me that fairmindedness cannot mean a mushy, middle-of-the-road position ("everyone is right!") on each of the issues when *separately* considered. When the major components of this long-standing controversy are teased apart by methodological analysis, I believe one can say some fairly definite things about them individually. When such definite positions are taken in defiance of clichés, toes are stepped on. Perhaps the most I can hope for is that I have stepped on clinical and statistical toes without favoritism. I hope that this scattering or my shots will incidentally help to disabuse non-Minnesota clinicians of the F–perception that there is a clear, monolithic "Minnesota line," predictable on the basis of conventional categories (nomothetic, dynamic, behavioristic, dust-bowl empiricist, global, objectivist, analytically oriented, and the like). Those who think about clinical issues in such terms cannot hope to understand the complexities of the reality.

Thanks are especially due to my wife Alyce, who knows how to protect a man at his work; and to Russell H. Linton, for many hours of informal psychotherapy.

PAUL E. MEEHL

University of Minnesota
June 11, 1954

Table of Contents

1

The Problem

ONE of the major methodological problems of clinical psychology concerns the relation between the "clinical" and "statistical" (or "actuarial") methods of prediction. Without prejudging the question as to whether these methods are fundamentally different, we can at least set forth the main difference between them as it appears superficially. The problem is to predict how a person is going to behave. In what manner should we go about this prediction?

We may order the individual to a class or set of classes on the basis of objective facts concerning his life history, his scores on psychometric tests, behavior ratings or check lists, or subjective judgments gained from interviews. The combination of all these data enables us to *classify* the subject; and once having made such a classification, we enter a statistical or actuarial table which gives the statistical frequencies of behaviors of various sorts for persons belonging to the class. The mechanical combining of information for classification purposes, and the resultant probability figure which is an empirically determined relative frequency, are the characteristics that define the actuarial or statistical type of prediction.

Alternatively, we may proceed on what seems, at least, to be a very different path. On the basis of interview impressions, other data from the history, and possibly also psychometric information of the same type as in the first sort of prediction, we formulate, as in a psychiatric staff con-

ference, some psychological hypothesis regarding the structure and the dynamics of this particular individual. On the basis of this hypothesis and certain reasonable expectations as to the course of outer events, we arrive at a prediction of what is going to happen. This type of procedure has been loosely called the clinical or case-study method of prediction.

Although all clinical psychologists make use of both sorts of predictions in varying degrees, and everyone admits some special merits and demerits of each type, it is nevertheless possible to characterize many clinicians as favoring one or the other. On this attitudinal continuum we would put such writers as Sarbin (**85**, **86**, **87**) at the one extreme together with Lundberg (**70**) and many users of "traditional" personality inventories. One usually thinks of Allport (**4**, **5**), Murray (**75**), the psychoanalytic group (e.g., **2**), psychiatrists generally, and most of the workers with a strong interest in projective techniques as being at the other end.

It is customary to apply honorific adjectives to the method preferred, and to refer pejoratively to the other method. For instance, the statistical method is often called operational, communicable, verifiable, public, objective, reliable, behavioral, testable, rigorous, scientific, precise, careful, trustworthy, experimental, quantitative, down-to-earth, hardheaded, empirical, mathematical, and sound. Those who dislike the method consider it mechanical, atomistic, additive, cut and dried, artificial, unreal, arbitrary, incomplete, dead, pedantic, fractionated, trivial, forced, static, superficial, rigid, sterile, academic, oversimplified, pseudoscientific, and blind. The clinical method, on the other hand, is labeled by its proponents as dynamic, global, meaningful, holistic, subtle, sympathetic, configural, patterned, organized, rich, deep, genuine, sensitive, sophisticated, real, living, concrete, natural, true to life, and understanding. The critics of the clinical method are likely to view it as mystical, transcendent, metaphysical, super-mundane, vague, hazy, subjective, unscientific, unreliable, crude, private, unverifiable, qualitative, primitive, prescientific, sloppy, uncontrolled, careless, verbalistic, intuitive, and muddleheaded. There are

also some words (e.g., positivistic, behavioristic) which are used sometimes favorably, sometimes unfavorably, depending upon the views of the speaker. Because of the extensive use of polemical words in discussions of the problem, I have listed them at the beginning for cathartic purposes so that we may proceed to our analysis unencumbered by the need to say them.

As a reminder of the flavor of this controversy, let us consider a few quotations without any attempt at a critical analysis of the kind of argument which is offered:

…the global approach at least respects the complexity of personality problems and seeks some elementary understanding before bursting into figures. (**50**, p. 50.)

Such standardization by its very nature ignores the individual.…All our theories of personality are at variance with the notion that the summation of a series of items determined by discrete frequency tables could ever be expected to give an accurate dynamic picture of an individual. (**74**, p. 233.)

Moreover, it has been claimed that psychostatistical manipulations and rigidly objective procedures are less applicable when carried over from the investigation of cognitive functions…to the more affective aspects of total personality. (**89**, p. 278.)

It would naturally be absurd ever to expect standardized tables based on statistical research which would enable one to determine whether a subject is schizophrenic, neurotic, or any other definite personality type—normal or abnormal. …There is no possibility of a rigid schematization, such as the establishment of standardized tables in which the scoring and interpretive value of every single Rorschach response would be listed.…Such a schematization would be incompatible with the basic principles of…any true personality diagnosis. (**65**, p. 21.)

In the latter [nonprojective] tests, the results of every individual examination can be interpreted only in terms of direct, descriptive, statistical data and, therefore, never can attain accuracy when applied to individuals. Statistics is a descriptive study of groups, and not of individuals. (**79**, p. 633.)

The statistical point of view must be supplemented by the clinical point of view. (**101**, p. 134.)

> …present statistical methods deal with averages and probabilities and not with specific dynamic combinations of factors. (**20**, p. 88.)

> A mathematical formula is possible and Zubin has attempted one; but it is in that rarefied mathematical atmosphere that has meaning only to mathematicians and statisticians. The present writer admired Zubin's effort, but found himself returning to inspection. (**9**, p. 85.)

> Indeed, psychological causation is always personal and never actuarial.…This is not to deny that actuarial prediction has its place (in dealing with masses of cases); it is good so far as it goes, but idiographic prediction goes further. (**5**, p. 156.)

> If predictions based on frequency were all that were possible, then a Hollerith machine worked on the basis of known frequencies by a robot could predict future behavior as well as a sensitive judge. (**5**, p. 159.)

Many other quotations of this sort could be given, although they are more frequent and uninhibited in informal discussions of clinical work than in journal articles.

I became fascinated by this problem at the 1947 meeting of the American Psychological Association, where Dr. E. Lowell Kelly presided at a symposium on clinical and statistical methods, a joint meeting of the clinical section of the A.P.A. and the Psychometric Society. Two comments could be made about this session. First, it was not very long before the usual arguments developed between the "clinicians" on the one side of the room and the "statisticians" on the other side. Dr. David Rapaport, for instance, said that certain statisticians apparently wanted him to substitute a Hollerith machine for his eyes and his brain. A second comment might be that the meeting was relatively poorly attended; which, considering the fundamental importance of the problem, seems to me a bad sign.

For this issue is not a trivial or academic one. In the first place, a psychologist's orientation on the matter has a considerable impact on his clinical practice. The degree and kind of validation he requires for a clinical instrument before using it to decide matters of commitments, shock, lobotomy, and psychotherapy, depend upon his conception of validation and

his notion of what the phrase "clinical validation" can reasonably mean. It is quite clear that a large number, perhaps the majority, of heated arguments about projective methods turn very shortly into a clinical-statistical controversy. And quite apart from a choice of testing instruments in the light of their validities, the distribution of clinical *time* is involved. How many hours of time of skilled psychological personnel can be profitably spent in staff conferences or team meetings in the attempt to make clinical judgments about the therapeutic potential of cases? This problem arises because therapists are in shortage. Every hour spent in thinking and talking about *whom* to treat, and *how,* and *how long* is being subtracted from the available pool of therapeutic time itself. The clerk or the statistician cannot do therapy; hence it is of the greatest importance to ascertain whether the clinician can do a better job of prediction than they can. If he cannot, we are wasting his precious time.

Furthermore, there are in every clinical setting occasions on which the predictions which would be made from a straight actuarial approach do not agree with the predictions made by a clinician. If some class to which the patient objectively belongs suggests a certain type of outcome on the basis of previous statistical experience, whereas the staff member who has been working with the patient feels that he understands the problem in terms of the individual dynamics of the case, it is necessary to decide whether practical decisions should be based on the actuarial findings or on the insight of the individual clinician.

The professional relationships of the psychologist are also profoundly influenced by his position on this issue. The use of psychometric devices, a statistical orientation, and the possession of statistical skills constitute *unique* tools of the psychologist. In the matter of history-taking and, if properly trained, counseling and psychotherapy, the psychologist, psychiatrist, and social worker are all capable; yet each of the three disciplines has its own unique kind of contribution. The professional prestige of the clinical psychologist and the kind of professional satisfaction he gets from

his work will be influenced profoundly by his orientation with respect to clinical and statistical methods.

In addition, it is desirable to have *some* rational formulation of what we do in practice. Two such apparently different methods of prediction should be somehow understood in their logical relationship to one another. Which differences between them are basic, and which are merely apparent? Why does one method of prediction "work" better in one case, the other in another? In the interests of intellectual consistency some rational reconstruction of the relationship of the two techniques needs to be given.

Finally, a clinician's view on this matter has a considerable impact on the character of his research. What sorts of things the psychologist decides to study, what methods he employs in studying them, and (unfortunately) the kind of results he finds depend partly upon his position on this clinical-actuarial continuum.

Some of the questions which are often involved in the clinical statistical discussion may be stated: Which of the two methods works better? How much mathematics and statistics should be required in the training of clinical psychologists? What should be done in individual cases when the actuarial and clinical predictions are not in agreement? Can it be argued that the statistical approach is suited for research but the clinical or case-study approach is the only one suited for clinical practice? Since clinically we are concerned with individuals and not group trends, should we therefore be paying less attention to the results of statistical methods when we work in the clinic? Do statistical methods imply an ignoring of "dynamic" factors? Can statistical methods be applied to all phases of projective techniques? If not, what limitations are there? What is to be substituted for them? Are there kinds of questions which it is simply absurd to try to formulate statistically? Is there a kind of clinical validation which brings its own credentials and is freed of the traditional problems of validity? Doesn't a global approach make statistical procedures outmoded? What relation exists between the statistical-clinical and nomothetic-idiographic dichotomies? Are these dichotomies or actually continua? What about the statistics of the single case? (See **7**, **28**, **29**.) How about

taking the person as your population from which samples are taken? Aren't statistical methods appropriate only to inventories of the old type? Could not all clinical inferences be, in theory, made in a. formal, statistical fashion? As science advances, can't we expect to see the gradual replacement of the clinician's judgment and synthesis by automatic, cut-and-dried manipulation of data?

Some of these questions are either pseudo-problems or involve large components of pseudo-issues, and others have either mathematical or empirical answers. All of them have need of semantic clarification.

2

Some Preliminary Distinctions

DISCUSSIONS of the problem tend to lump together issues which are logically independent, simply because of certain sociological clusterings in the opinions of psychological practitioners. Thus, if your remarks show that you are favorable to a fairly orthodox brand of Freudian theory, others are likely to assume that you are global, intuitive, antibehaviorist, projectivist, and nonstatistical. There is no doubt that certain clusters actually exist in the behavior of psychologists, such as described in Murray's list of differences between centralists and peripheralists (**75**, pp. 6-10).

If I know that a psychologist is a Hullian in learning theory, that he has done experiments on albino rats, and that he owns a copy of Skinner, I can predict somewhat better than chance that he will be mildly suspicious of the Rorschach, that he would put his bets on actuarial methods of prediction, and that he thinks that candidates for the doctorate in clinical psychology ought to learn a little undergraduate mathematics. Nonetheless, there is no *logical* implication from one of these opinions to the others. If you bet this way you stand to win, but in attempting a rational analysis of the issues involved we must not take these sociological groupings for granted as a basis for argument.

It should be emphasized that I am concerned in this monograph wholly with the problem of prediction and am not talking about psychotherapy. It is evident that one cannot manipulate the behavior of a person by filling numbers into a multiple regression equation. Of course, certain

aspect of prediction—e.g., prognosis with insulin shock treatment in schizophrenia, or choice of interviewing technique—have a direct therapeutic import. The application of concrete predictions to therapeutic problems is a practical issue and will not be treated except tangentially. I am concerned here solely with the empirical problem of making correct predictions about the course of events, and with a logical analysis of this enterprise.

The first clarifying possibility that occurs to me is that there may be two different kinds of statistics or, perhaps I should better say, two different ways of applying statistics. I do not have any great confidence in this distinction but find it helpful in thinking about this issue. There are no standard words for these two methods, and I should propose the distinction between what may be called the *discriminative* (or *validating*) use of statistics on the one hand, and the *structural* (or *analytic*) use of statistics on the other. As a first approximation, we may say that the discriminative or validating use of statistics is the use which makes few or no psychological assumptions about the nature or structure of the behavior being investigated. The use of such methods is almost wholly neutral as regards theory. The only assumptions made are certain very basic or broad assumptions, usually directly confirmable within the data, involving such things as the shape of the population frequency function and the randomness of certain series. In the pure case of this use of statistics, the only assumptions required are those of the theory of probability. Even here, the empirical conditions for applicability—e.g., the existence or nonexistence of randomness—can usually be subjected to a direct empirical test within the material collected.

Typical questions of the discriminative or validating type would be as follows: "Is the trait or attribute x associated in any way (not merely in the sense of Pearson r) with the attribute y in a group of persons defined by so-and-so?" "When Mr. A mentions his brother in an interview, is he more likely to talk about his thwarted ambitions than he is in those interviews in which he does not make any mention of his brother?" "Can a group of

educated judges match these personality sketches better than chance with the names of people they know?" "If I, clinician Z, make any use or combination I choose of the MMPI profiles of patients called schizophrenic at this hospital, in the attempt to predict the rated outcomes of insulin shock therapy, can I do so significantly better than I could by flipping pennies or entering a table of random numbers?" The prototype of this kind of statistics, it seems to me, would be the method of correct matchings. We do not, except in designing the experiment, make any implicit assumptions concerning the judges, the kind or data they are using, the mode of combining information, etc. The use of statistics consists in a direct application or pure combinatorial analysis in which the reference base is the "chance" hypothesis; and the probabilities by, for example, Chapman's tables (**30**) are precise upon this basis.

As distinguished from the discriminative or validating use of statistics, I have proposed the term *structural* (or *analytic*) use. This use of statistics presupposes certain empirical assumptions about the behavior—or constructs used to "explain" the behavior (**71**)—which are not themselves directly confirmed in the analysis. If these assumptions are false, or to the extent that they are poor approximations, the inferences are untrustworthy. Often the complete statement of the required hypothesis concerning the behavior or constructs may be of a high order of complexity. As examples of such a use of statistical method, I would consider such inferences as these: "I have solved the multiple factor equations backward from this individual's test scores, and I conclude that he has an amount a ± e of Factor II as a primary ability." "The orthogonal solution of the intercorrelation matrix of these symptoms indicates the presence of a psychological dimension hysteria-dysthymia which is similar to the extravert-introvert continuum and which is uncorrelated with a trait of general neuroticism." "The analysis of covariance indicates that the observed differences in trait A among different social strata are attributable solely to differences in verbal intelligence." The prototype of this use of statistics is factor analysis.

I do not know whether it is possible to assign most statistical tools or techniques to these two classes without regard to the particular use to which one puts them. Even in the case of factor analysis, if one is willing to look upon the factor matrix as "nothing but" an arbitrary simplification of an intercorrelation matrix, no psychological issues are involved. It is difficult to see what is the value of such an approach, for either theoretical or practical purposes. If we are interested in a straight prediction problem, as Burt has pointed out (**21**) factor analysis cannot enable us to improve upon straight regression procedures, where the sampling problem has been better worked out. If it is the intention to use results of the analysis for the improvement of testing instruments so that they will have greater *inherent* validity and "purity," the likelihood of achieving this depends upon the adequacy of our psychological inferences made as a result of the factor analysis.

If, for example, a particular solution of the rotation problem gives us three factors which do not correspond at all to the underlying dynamics (causal agents) which have in fact given rise to the observed correlations, we shall not find any particular improvement in prediction when we make up new test items on the basis of the pseudo-insight gained from an inspection of the old factor matrix. Even such statistically simple procedures as partial correlations or the discriminant function are discriminative or structural depending upon what we do with them. Sometimes we use the discriminant function simply to give the optimal weight to the members of a predictive battery, and the assumption of linearity and absence of pattern interactions of the predictive variables are assumptions which are testable within the data. On the other hand, we may be interested in making a psychological interpretation of the weight in the discriminant function and in speaking about the contribution (in the causal-determinative sense) of the dimensions measured. Such statements are sometimes made in such a manner that they must be considered structural-analytic application of this neutral statistical tool. If the purely *statistical* assumptions are fulfilled, a partial correlation simply tells us what the correlation surface of two variables is like on a slice of the box

determined by looking only at the triads of numbers in which the third number has a constant value. But in actual research it is very rare that we are willing to confine ourselves to such a cautious claim. We want to know, for example, whether the relationship between achievement test score and socioeconomic status is attributable to the factor of intelligence. The well-known problems involving whether one partials out too much, what direction the relationship runs, and so on arise because the statistical analysis of the data does not make these structural-analytic distinctions.

We may tentatively conclude, then, that this distinction refers both to the *aim* of a statistical procedure and (as a consequence of the aim) the *assumptions* of a nonstatistical character which must be made in order for this aim to be reached on the basis of the statistical findings. The method of correct matchings, simple significance tests, and straight prediction systems will usually be found to be discriminative-validating; whereas factor analysis, the analysis of covariance, and most applications of partial correlations will typically be used in a structural-analytic way.

As I have said, I am not sure of the value of this distinction and I am not arguing that it reflects a fundamental logical difference between the two kinds. But it seems to me that discussions regarding the use of statistical methods in clinical work are sometimes confused because arguments for or against one of these uses of statistics are erroneously treated or reacted to as arguments for or against the other. For example, in response to a demand for validation data, clinicians will sometimes state that they "do not work in a mechanical, additive way" and that the usual statistical procedures are therefore not applicable to their clinical behavior. More often than not, this is hokum. Again, some clinicians object to factor analysis because it uses basic equations with no cross-product terms, and because the assumption of constant factor loadings over the population is implausible. Here the clinician is (I think perhaps validly) calling into question a psychological presupposition needed for a particular structural-analytic use. But this does not in the least free him from the

obligation of showing *statistically* that his own predictions, on different assumptions, tend to be correct. That is, it does not enable him to avoid the discriminative-validating use of statistics. Unless these functions are separated, confusion results continually.

A second distinction is that between the source or type of information employed in making predictions, and the manner in which this information is combined for predictive purposes. It appears to me that Allport has contributed somewhat to this confusion. I should distinguish first, as regards *data,* between psychometric and nonpsychometric kinds of information. As a completely different dichotomy (or continuum) I should distinguish as regards *method* between mechanical (or formal) methods of combining data and nonmechanical (or informal) methods (so-called judgmental, clinical, impressionistic, or subjective).

With reference to the kind of data used, by *psychometric* I mean *tests* in the fairly strict use of that term. If the data arise from a systematic behavior sample having the following four cardinal properties of a psychological test, I shall consider them psychometric: (1) standardized conditions of administration, (2) immediate recording of the behavior or behavior products, (3) objective classification of the responses ("scoring"), (4) norms. It seems that this division between psychometric and nonpsychometric samples of behavior is also actually a continuum rather than a dichotomy. Any kind of information which is not based upon tests in the above-defined sense I shall call nonpsychometric or case-study data. Examples of this would be remarks made during an interview, the social history, a police record, a rating by the examining physician, facts about present marital or employment status, subjective impressions from the patient's voice, expressive movements, etc. Note that case-study data need not be "subjective" or "impressionistic," although they may be.

As for the combining method, by *mechanical* (or *staistical*) I mean that the prediction is arrived at by some straightforward application of an equation or table to the data. I do not mean the word in its usual pejorative sense. This table, let me emphasize, does not have lo be a table of

individuals. The elements of such a table may be episodes or occasions in the life history of one person. The defining property is that no judging or inferring or weighing is done by a skilled clinician. Once the data have been gathered from whatever source and of whatever type, the prediction itself could be turned over to a clerical worker. By *nonmechanical* or *informal* methods of combining I mean those of any other sort. It must be stressed that "nonmechanical" is *not* to be identified with "intuitive" or with any mode of combining data that has the connotation of subjectivism or irrationality. It may be intuitive in special cases; on the other hand the clinician making this sort of prediction may give explicit reasons for his predictions from the data but they are not a mechanical consequence of a table or equation plus rules for applying it. That Sherlock Holmes does not employ an actuarial table is not tantamount to saying that his procedures are nonrational! (A minor point here is that the clinician may talk about a score which in itself is actuarial in the sense that it is, say, a sigma score. But unless there is some direct and strict relation between this score and a prediction that is tabulable, he is not predicting mechanically in the way I am using the term.)

It is obvious that before we ask what mode of combination of data is used in reaching a prediction, the *data* have already to be somehow given. Thus, a statistical clerk may combine, by purely mechanical, explicitly stated rules, sociometric judgments made by fraternity brothers. Given such judgments, the clerk is proceeding statistically in my sense—the clerk needs only to be able to read, write, and figure to get out a prediction. In the extreme case the clerk might not even know the source of the ratings, or the empirical meanings coordinated to the numbers he is given and the predicted score at which he arrives. But if we inquire into the fraternity brothers' judgments themselves (which might even be couched in predictive terms, e.g., "Who would be the best arranger of a picnic?"), these judgments are not arrived at mechanically or statistically, *in my sense.* They are human judgments, the rules for which are buried in the judges'

heads; we cannot train a clerk to observe the subjects' behavior and then, by straightforward mechanical means, duplicate (except for ordinary clerical errors) the judges' judging behavior. I am not here concerned with which would be *better* in predicting the final criterion of picnic management; I am simply pointing out that there is an obvious, noncontroversial operational difference between the clerk's activity and the judge's, a point which is established immediately we realize that a second clerk (or a machine) can be easily made to duplicate the predictions of the first *starting with the same data,* a possibility no one seriously claims with respect to the judgments of the judge. Whether what the judge "adds" is error or not is here quite beside the point.

We see from this example, however, that the question "Is this prediction clinical or statistical?" is likely to be an ellipsis. The expanded form would be "Is this prediction, given such-and-such data expressed in so-and-so form, clinical or statistical?" Thus, we have immediately a question of *levels,* in the sense that the transition from a certain class of statements, scores, or behavioral adjectives to the prediction proper may be purely mechanical, following explicit rules; whereas this evidential class itself may consist of members all, some, or none of which were arrived at by human judgment, at least partly inexplicit. There is no need for persistent ambiguity here, since in any real case we can specify the level of data with respect to which the query "Clinical or statistical?" is being raised. That the answer varies as we treat different levels or stages of the same total predictive process is only to be expected. The use of a Stanford-Binet score in a regression equation is statistical, and *starting with* the score as the datum, this regression method can be significantly compared with a competitor prediction by a clinician looking at the same set of numbers. Yet at a much lower level, the scoring of the individual item responses, there occurs a process of human judgment which, no matter how reliable it can be made by short training, is still not quite clerical or mechanical in character. In several of the empirical studies we shall review in Chapter 8 (e.g., Wittman's, **103**) the reader should bear this matter of levels in mind,

since judgmental components enter into the total predictive chain at a level below that for which the crucial comparison of clinical and statistical is being made.

Let us pause for a moment to consider the fact that all four combinations of data with methods are constantly occurring in clinical practice. For this reason any discussion of the problem that does not distinguish between method and data is likely to lead to confusion:

1. *Psychometric data combined mechanically.* An intelligence test and a test of reading speed are combined in a multiple regression equation for the prediction of college grades.

2. *Psychometric data combined nonmechanically.* A clinician skilled in the interpretation of the Strong Vocational Interest blank, the Rorschach, or the Minnesota Multiphasic gives a personality description and guesses a prognosis from inspecting a profile of one of these devices.

3. *Nonpsychometric data combined mechanically.* Parole prediction tables in criminology use data such as age at first sentence, size of community, and marital status, but these data are combined by a statistical table in a mechanical fashion to arrive at a prediction.

4. *Nonpsychometric data combined nonmechanically.* On the basis of the history, an interview, and observation of the patient's behavior on the ward, a psychiatrist decides to give the patient electroshock.

More complex combinations also occur. A very common one is the combination of high school rank with ACE score in a regression equation to predict academic grades. This is of course an instance of psychometric plus nonpsychometric data, combined mechanically. The most common case of all in clinical practice is that of psychometric plus nonpsychometric data combined nonmechanically, where we have the history, an interview, ward behavior, and the results of standardized and semistandardized psychological examinations combined in a staff conference in the attempt to yield a diagnosis (in the broad sense of that word) which in turn entails some sort of prediction.

3

The Rationality of Inference from Class Membership

ONE point which I feel is really crucial is Allport's seeming implication that inference from class membership is somehow inherently fallacious. He does not explicitly state this, but some of the arguments leave one wondering if he does not believe it. For instance, in his monograph on personal documents we find the following paragraph:

> Where this reasoning seriously trips is in prediction applied to the single case instead of to a population of cases. A fatal nonsequitur occurs in the reasoning that if 80% of the delinquents who come from broken homes are recidivists, then *this* delinquent from a broken home has an 80% chance of becoming a recidivist. The truth of the matter is that *this* delinquent has either 100% certainty of becoming a repeater or 100% certainty of going straight. If all the causes in his case were known, we could predict for him perfectly (barring environmental accidents). His chances are determined by the pattern of his life and not by the frequencies found in the population at large. Indeed, *psychological causation is always personal and never actuarial.* (**5**, p. 156.)

In general, I agree with the content of this paragraph and admit the importance of Allport's point. However, the phrase "a fatal nonsequitur" *could* be a source of confusion, because one gets the impression that Allport believes it is a nonsequitur *because* it is based upon an inference from the fact of class membership. I should like to stress that *if nothing is*

rationally inferable from membership in a class, no empirical prediction is ever possible. There is, in Allport's paragraph, a subtle implication that by non-actuarial methods you can predict "for sure." It is interesting to note that in spite of his dislike for actuarial concepts he begins the crucial sentence with "His chances are determined." The whole notion of someone's "chances" is, as Sarbin has emphasized, an implicitly actuarial notion.

The superiority in some cases of making such predictions from a study of the occurrences in the individual life over trying to make them on the basis of his membership in a class of *persons* can be established without departing from actuarial reasoning if we construct a table such as that shown on page 21. Here, situations are represented along the horizontal—e.g., days of the week—and persons along the vertical. The marginal totals in the table give us the over-all frequencies for situations such as the probability that a person will go to the movies on Saturday night if we know nothing about the person; the corresponding marginal totals going the other direction give us the probability that Professor A will go to the movies when we don't know which night it is. It is apparent that in general the maximation of "hits" will be achieved when the probability figure used to arrive at our prediction is that of the smallest possible subset, i.e., P_{A7} rather than P_A or P_7 (and, a fortiori, P). Special cases exist in which it makes no difference.

It should not be implied, as Allport seems to, that we can *always* do better knowing the frequency for Jones than we can knowing the frequency for the class to which Jones has been ordered. In the event that the modal frequency is attached to the identical prediction whether the analysis is by situations or by persons, we would be predicting the same thing by both methods and the success frequency for the table as a whole would be the same. However, if there is at least one row or column in which we would reverse the prediction on the basis of the subclass frequency, we will stand to improve our guesses. It is obvious that the best prediction would be that based upon the P value for a given entry, i.e., for what I

		SITUATIONS							
		1	2	3	4	5	6	7	Total
PERSONS	A	P_{A1}	P_{A2}	·	·	·	·	P_{A7}	P_A
	B	P_{B1}	P_{B2}	·	·	·	·	P_{B7}	P_B
	C	·	·	·	·	·	·	·	·
	·	·	·	·	·	·	·	·	·
	·	·	·	·	·	·	·	·	·
	·	·	·	·	·	·	·	·	·
	·	·	·	·	·	·	·	·	·
	·	·	·	·	·	·	·	·	·
	Z	P_{Z1}	P_{Z2}	·	·	·	·	P_{Z7}	P_Z
	Total	P_1	P_2	·	·	·	·	P_7	P

shall call an occasion, meaning a person-situation interaction. To carry the argument further, it might be that we could improve even over this kind of guess if the situations for Professor A were themselves ordered as to time, so that whereas the over-all frequency is .75, an analysis for such a time series would lead us to conclude that the relative frequencies were not random with respect to successive occasions. These are fairly obvious points but I stress them in order to make clear that Allport can defend his interest in Jones as an individual without departing at any point from an

analysis which is still essentially statistical; for his conclusions can be based simply upon an analysis of certain class inclusion relationships among frequencies. I do not, however, wish to defend the Lundberg-Sarbin position that *all* prediction is of this sort, as will be clear from the discussion of the probability of hypotheses in Chapter 6.

As to Allport's emphasis upon the distinction between prediction from categories and predictions for the individual, it should be clear that in principle all laws even of the so-called causal-dynamic type refer to classes of events. "Adding more information about the person" is taken by Allport and Alexander (**2**) as a relatively unanalyzed idea. But a case can be made that this always consists in assigning him to a still narrower subclass, that is, to a class having more restricting properties. The question of the optimal subclass has been considered by Reichenbach (**82**, p. 316) and from his point of view there is no such thing as *the* probability of an event. There are as many probabilities as there are specifiable classes. No one of them is any truer than the other, but nevertheless from the standpoint of prediction, there is a best class, and this best class is always to be defined in the same way. It is the smallest class, i.e., extensionally smallest and intensionally most complex, for which the N is large enough to generate stable relative frequencies.

Paradoxically, the uniqueness of individual events which Allport is at such pains to emphasize in all his writing forces us to assume that it *is* rational to entertain expectancies about the future on the basis of class membership. The alternative view, if made explicit, would have to be something like this: "Nothing can be rationally said about an individual instance on the basis of its class membership, because the members of the class differ with respect to other predicates than the defining one, or differ on some quantitative dimension as regards the defining predicate itself, or there is a qualitative difference in arriving at the same dimensional point." Even the ordinary practical decisions of everyday life become strictly impossible to rationalize if one really argues consistently that it is not

rational to decide in any particular instance on the basis of a known or estimated frequency in some class to which the unique instance belongs.

This can be made very clear by considering the case of a regression system leading to a multiple R of .999. Surely Allport would not deny the rationality of predicting the individual subject's behavior on this basis. But if this is reasonable, is not .990 reasonable? And then, why not .90, and thus .75, and, to be consistent, .25? Surely there is no miracle that renders such prediction suddenly irrational, no discontinuity in the situation such that, say, to predict for an *individual* when R = .9 is legitimate, but not when R < .9. The only conceivable discontinuity in the logic as related to the statistics would be at R = 1.00; but if Allport were to maintain that it is irrational to predict for individuals when the prediction system involves an R in the open interval (–1, 1) he would have to abandon *all* prediction, and not only in the social sciences at that!

4

The Special Powers of the Clinician

STOUFFER (**95**) has treated the question "What can the clinician do with his facts beyond that which can be done by the mechanical application of an actuarial table or a regression equation?" In his discussion Stouffer chiefly emphasizes the fact that the clinician can in special cases give more weight to a factor than it is given in the actuarial table. On what basis can he validly do this? As has been pointed out (e.g., by Lundberg, **70**, p. 382), if he does so, he must be using some law or other based upon his previous experience, and this law, argues Lundberg, is actuarial. The sense in which Lundberg's use of the term "actuarial" in this context is legitimate we shall consider later. At least it is admitted by all that there are special instances in which the clinician can apply some knowledge which is not included in the table or which, if it is, is not given the weight that he feels it should be given in the case at hand.

Whether the clinician tends to improve over the table under these conditions is an empirical problem. For instance, suppose that in the table given in Chapter 3 we are trying to predict whether a given professor will attend the movies on a given night. On the basis of the values in this table and a failure to show any time-series change in the relative frequency when the occasions are ordered as to time, we arrive at a probability of .90 that he will attend the neighborhood theater, the present night being Friday. The clinician, however, knows in addition to these facts that

Professor A has recently broken his leg. This single fact is sufficient to change the probability of .90 to a probability of approximately zero. Sarbin or Lundberg might reply to this that either such a fact is important in the prediction or it isn't. If it is important, that is, if it *should* be taken into account (whether the clinician thinks it should or not), it can in principle be discovered by the use of actuarial tables. If the word "actuarial" is used this broadly, so as to be synonymous with "inductive," I doubt that any clinician would care to argue the issue. Whether this is a useful way to use the word I shall consider below. I should like merely to point out that a statistical study of a large number of professor-situation occasions of the present type, in which factors were decided upon on the basis of the establishment of statistically significant differences between the movie-goers and the non-movie-goers, would presumably not result in the isolation of broken legs as an important variable. The simple reason is, of course, that this is a factor of extreme rarity in *both* of the criterion groups. In other words, such a factor does not appear as statistically important in the mass event, but if the clinician knows this fact in the case of Professor A he (correctly) allows it to override all other data in the table.

The actuary may counter by saying, "If the factor in question is so rare, why bother with it?" It is the tremendous interest in the individual case that defines the clinician. Furthermore, speaking of the mass of cases, there may be *many (different) rare kinds of factors.* The cases which they largely determine add up to a very sizable minority of all the cases for which prediction is made. The situation is somewhat like the old paradox that "an improbable event is one that hardly ever happens, but nevertheless something improbable happens almost every day." An improbable factor of a given type may occur with extreme rarity, but improbable factors as a class, each of which considered singly will not appear in a statistical analysis as significant, may contribute heavily to the "misses."

In passing, it may be pointed out that these rare cases furnish one of the respects in which the human brain can be a very sensitive indicator. To

take a simple example, I recently attended a staff conference in which one of the psychologists made a correct diagnosis of an absent patient after hearing an inadequate, if not downright misleading history, because he recognized the profile pattern on the MMPI as very similar to an unusual pattern he had seen over four years previously in a case of alcoholic hallucinosis. It is of course true that with a sample of only one case, there is a very sizable chance that he is wrong. (It just happened in the present instance that he was right, so he now has N = .2!) This raises an entirely different although legitimate question, namely, that of validating the clinician. I agree with Sarbin and Lundberg that this is a thoroughly actuarial problem, involving the discriminative use of statistics. We need here Reichenbach's distinction between the "context of discovery" and the "context of justification" (**82**, p. 7). The clinician may be led, as in the present instance, to a guess which turns out to be correct because his brain is capable of that special "noticing the unusual" and "isolating the pattern" which is at present not characteristic of the traditional statistical techniques. Once he has been so led to a formulable sort of guess, we can check up on him actuarially.

The whole problem of the miraculous brain is intimately involved in the problem of clinical and statistical methods of prediction. Clinicians often hold the view that no equation or table could possibly duplicate the rich experience of the sensitive worker. Here psychology has its precedent from medicine, in the old country doctor who is a "brilliant clinician," as evidenced by the fact that he seems to be able to "smell" diphtheria merely by walking into the sick room. Murray has stated that no instrument could have the analyzing and integrating power of the human brain. Having extracted from a man the best possible explicit verbal description of his wife's face, we can without much difficulty find hundreds of women in a given community who meet the requirements of his description; and yet the man himself would be able in a split-second glance to distinguish his wife's face from the others. Opposed to this kind of datum, one may ask whether modern fire-control methods could have been con-

structed by the use of clinical intuition, or without the aid of explicit mathematical analysis? It is not difficult for the protagonists of the clinical or actuarial view to cite both kinds of evidence, to show either that the brain is a good instrument or that it is a relatively poor one.

I do not feel that there is enough empirical evidence at hand to decide this question or, better, to determine in which situations the brain is a powerful device and in which it is relatively weak. An obvious hypothesis, suggested by such researches as those of S. G. Estes (**40**) and the mass of material gathered by the Gestalt psychologists, is that the brain's "superiority" shows up heavily at the level of perception itself. At the level of subtle cues of a primarily social type, any normal person has had a very long history of rewards and punishments with respect to responses to such cues. Responses to certain configurations of sense data as being indicators of the inner states of other organisms are presumably acquired very early. It is even possible that some of these configurations do not require to be learned but are given as part of our biological heredity. (Cf. Goodenough, **46**.) If the term "facts" is used with sufficient broadness to include *perceptual facts* of this type, we return to the argument about actuarial methods of combination when the facts are given, as, for example, immediate or impressionistic clinical judgments. In other words, if we are willing to call such immediate impressionistic responses to social cues "facts" in the sense that the clinician is here operating as a testing instrument of a sort, it is still an open question whether the fact that the patient acts hostile or dominant *ought to be given the weight that the clinician gives it in arriving at his predictions.* An empirical example of this sort of thing will be found in the study of Wittman discussed below.

In any case, psychologists should be sophisticated about the errors of observing, recording, retaining, and recalling to which the human brain is subject. We, of all people, ought to be highly suspicious of ourselves. For us, the problem of the adequacy of the analyzing and integrating human brain is to be approached through an empirical investigation of its success.

I once worked with a psychologist who has been very much interested in the clinical use of a certain test. An extremely sensitive and able clinician, he had administered this test to somewhat over 600 patients of varied diagnoses and intelligence levels. To many of these patients he had also administered a Wechsler-Bellevue Intelligence Test. He stated to me on one occasion that he felt he could do a pretty good job of estimating IQ's from the new test, although he had never checked himself against the Wechsler systematically. He had "noticed" that on the whole the correlation seemed "pretty good." The correlation between the IQ of a group of cases and this clinician's guesses from his favorite test was .04. The point of this anecdote is that this clinician was quite sure that he could do it, and the fact of the matter was that he could not. It should not be necessary to admonish psychologists on this subject, but recent conversations have convinced me that there are some clinical psychologists who are so busy being clinicians that they tend to forget they are psychologists. The kind of skepticism about human observation and inference which was engendered in large part by the classical studies in the psychology of testimony and by the early work on judging people—e.g., that of Hollingworth (**55**)—can be carried to extremes and undoubtedly has been so carried by some superskeptics. Nevertheless, we have no right to assume that entering the clinic has resulted in some miraculous mutations and made us singularly free from the ordinary human errors which characterized our psychological ancestors. There are some published investigations of the hypothesizing and predicting behavior of clinicians which ought to make us rather cautious and humble in our claims (**6**, **39**, **64**, **68**, **69**, **97**, **98**).

5

The Theoretical Argument of T. R. Sarbin

THE most radical of the recent actuarial debaters is T. R. Sarbin, whose systematic treatment and review of the few experimental studies appeared several years ago (**87**). I find it surprising that only one clinician was impelled to respond (**32**). Since Sarbin has been a practicing clinician, the case he makes is all the more interesting and we owe it to ourselves to take his arguments with great seriousness. I have learned a great deal from study of Sarbin's paper and for some time was persuaded that his position was wholly correct. But I feel now that the argument as he states it requires qualification if not some basic revisions.

The course of Sarbin's argument runs something like this: No predictions made about a single case in clinical work are ever certain, but are always probable. The notion of probability is inherently a frequency notion, hence statements about the probability of a given event are statements about frequencies, although they may not seem to be so. Frequencies refer to the occurrence of events in a class; therefore all predictions, even those that from their appearance seem to be predictions about individual concrete events or persons, have actually an implicit reference to a class.

The basic premise by which Sarbin attempts to show that the clinician is always predicting actuarially and from classes whether he knows it or not is an appeal to the criterion of verifiability (or, as the newer

terminology of positivism has it, the criterion of confirmability). All empirical statements must be capable in principle of confirmation or disconfirmation. If I say before throwing a die, "The probability of this die coming up an ace is 1/6," how is such a statement to be confirmed? I throw the die and it comes up an ace. Obviously the fact of an ace is compatible with other statements of the probability, in fact somewhat more so, at least in the sense that the thing which occurred was "improbable" by having a probability of less than 1/2. On the other hand, if it comes up other than an ace, this statement is still not confirmed because there is no principle of probability theory which states that the improbable may not occur. It is evident that a decision between a statement that the probability of an ace is 1/6 and the statement that it is 1/7 cannot be made on the basis of the outcome of my throw. Sarbin argues that unless such probability predictions are to be completely meaningless they must be confirmable in principle and hence they must refer implicitly to a class. For it is only if we have a reference class to which the event in question can be ordered that the possibility of determining or estimating a relative frequency exists.

Sarbin applies the same reasoning to the case of prediction of single events in the clinical situation. The clinician is interested in predicting whether Jones will commit suicide within a year. The clinician, unless he is actually utilizing actuarial tables, does not assign numerical values to these predictions; but as Sarbin and Lundberg point out, the appearance of words like "probable" and "likely" involves reference to an actuarial notion. The failure to realize this sometimes results in amusing paradoxes, as in the reference to "his chances" in the paragraph from Allport cited above. A similar slip occurs in Alexander (**2**, p. 441). So that whether P is actually stated metrically or in a phrase such as "rather likely" or "quite probable" or "can be reasonably expected," Sarbin says is irrelevant. "The probability is P that Jones will kill himself" is of the same type as the prediction about the die. If this prediction actually refers to a single event, i.e., the suicide of Jones, Sarbin argues that it is unverifiable in principle

and consequently excluded by the verifiability criterion of meaning. The only way in which it can have a meaning attached to it is by ordering it, as in the case of the die, to a class for which a success frequency of such predictions is defined. Therefore the clinician, if he is doing anything that is empirically meaningful, is doing a second-rate job of actuarial prediction. There is fundamentally no logical difference (and here Sarbin is arguing the same position as Lundberg) between the clinical or case-study method and the actuarial method. The only difference is on two quantitative continua, namely that the actuarial method is more *explicit* and more *precise.*

The argument seems to proceed quite inexorably to its end, and yet it is very difficult for the clinician to feel as though it is an adequate description of what he is doing, even implicitly. We do not feel that we are carrying on this kind of enterprise when we discuss the hypothetical dynamics of a case in a staff conference. Lundberg asserts that this merely means that the clinician does not "know" all the information which is contained in his own previous experiences, and that whether the clinician recognizes the actuarial character of his predictions is irrelevant. It is clear that the clinician's feeling about the matter cannot be used as a rational argument, but it perhaps justifies us in scrutinizing Sarbin's development very critically.

The first thing that one might say about Sarbin's exposition is that he does not distinguish carefully between *how you get there* and *how you check the trustworthiness of* your judgment. It is clear (and in fact trivial) that the case study leads only to probable judgment, and of course all knowledge about the empirical world is confined to probable judgment. In order to assess the confidence that we ought reasonably to place in the predictions of the clinician, it seems straightforward to keep a record of his guesses and to determine his success frequency. This procedure, as has been pointed out above, is quite independent of any analysis of the processes whereby the clinician *arrives* at his judgment. Such an investigation can be carried on with a predicting organism whose mode of operation is com-

pletely enigmatic. (Cf. Reichenbach's clairvoyant, **82**, p. 358.) We must study the judgments of the clinician and arrive at some reasonable statement as to his successes by ordering his predictions to a class; but does it follow that anything of an actuarial character is being carried on by the clinician himself? That he is operating implicitly in an actuarial fashion may be *true*, but this involves a different question from that which is involved in the matter of confirming his guesses.

This brings us to a second consideration, touched upon indirectly by Chein (**32**), which may deprive Sarbin's argument of much of its force. It seems to me that Sarbin is not distinguishing between a sentence about Jones and a sentence *about the sentence about Jones.* The grammatical form of the prediction as Sarbin gives it—"Student X has one chance in 6 of meeting the standards of competition in the university"—undoubtedly contributes to this confusion. But it seems to me that we have here to deal with two sentences.

The first sentence is the prediction proper, "The student X will not succeed at the university" or, in our example, "Jones will kill himself." It is obvious that these sentences are, in their content, references to single events, occurring in the life history of particular persons; nevertheless they are not excluded by the application of the confirmability criterion of meaning. It is not difficult to verify the suicide or nonsuicide of Jones. The suicide of Jones is a specific event which will or will not occur and which does not present any new problems for confirmation or disconfirmation beyond those involved in any particularistic hypothesis. If a year hence Jones is dead by suicide, the prediction is confirmed; if not, the prediction is disconfirmed. I do not think it legitimate to invoke the confirmability criterion of meaning in trying to prove that the clinician is here proceeding actuarially. When the clinician says that he intends to speak about Jones and *not* about a group of individuals in making such a prediction to the patient's relatives, the court, or a social agency, it is understandable that he should resist Sarbin's insistence that he is not talking about Jones. Here, I believe the clinician is right and Sarbin wrong.

It is only when we wish to assign a confidence, weight, or probability to such a statement that the meaning criterion comes into operation. The sentence assigning this confidence is *about* the sentence which speaks of Jones. In *this* respect (if we accept the generalized frequency interpretation of the probability concept) Sarbin is presumably right; but I do not think that even here the clinician is "wrong" since I am not aware of any statement by a clinician which denies the frequency interpretation of this second kind of statement. Allport, for example, is commonly taken as standing in the clearest opposition to Sarbin's view; yet he has several times reiterated the need for studying the accuracy of such judgments, the individual differences in this accuracy, correlates and determiners of these variations, the relation between the confidence of a clinician and his tendency to be right, and so on. I think all clinicians would agree that if they assign a numerical probability to a prediction about Jones, although the prediction about Jones has a specific content tied to Jones and is itself directly confirmable by the individual event of the future, the justification for the probability number must lie in the establishment of some sort of empirical frequency.

I should like to point out in elaboration of Sarbin's discussion that there are alternative ways of looking at this probability statement which are, in terms of present formulations, equally legitimate. Even if we agree that the assignment of a probability number (such as 3/4) to an individual prediction can only have meaning in terms of a relative frequency for a class, it does not follow that this class is a class of *individuals.* Nor need it be repeated occurrences in the life history of one person, which is the only alternative mentioned by Lundberg (**70**). Just as there are an indefinitely large number of classes to which Jones can be ordered, each of which will have its own "correct" relative frequency, so also the *prediction* about Jones may be ordered to classes not even defined by the properties of Jones or Jones' situation but rather by some "non-Jones" characteristics.

The crudest example is to order the prediction (treated as a sentence occurring in the clinician's verbal behavior) to the entire class of sentences

the clinician emits qua clinician. This is the largest class and although its relative frequency is very stable, it is too broad to be very informative. To be sure, the establishment of such a number for each clinician would be of theoretical and practical interest, and a crude guess as to this relative frequency is made by most of us about our colleagues in clinical work. We may define narrower classes, e.g., what is the relative success frequency of clinician A when he is concerned with the prediction of suicide? Or, what is the relative success frequency of clinician A when he is making predictions about patients of a given sort? Or, what is the relative success frequency of clinician A when he attaches to his individual prediction the statement "I am very certain about this one"?

Sarbin leaves the reader with the impression that there is a *true* probability which the crude mental operations of the clinician poorly approximate; in point of fact there are hierarchies of probability, and only an empirical study of frequencies will tell us on which system of classifying the predictions we ought to lay our bets. The best bets will be based upon the relative frequency of success of predictions for joint (multiple predicate) classes, including the clinician, the situation, the nature of the predicted events, and all the information about the individual. None of these procedures for assigning confidence to the concrete prediction of the clinician restrict him in the psychological operations he goes through in coming to the prediction. It is for this reason that we can admit with Sarbin the necessity for attaching an empirical meaning to a numerical probability, without immediately concluding with him that the clinician is a second-rate substitute for a Hollerith machine. This latter statement, which both Sarbin and Lundberg appear to believe, may or may not be true; the important point is that whether true or not, it cannot be established by Sarbin's appeal to the positivist meaning-criterion.

An even more fundamental difficulty with Sarbin's argument might lie in the application of the meaning criterion even to the numerical probability. Sarbin's discussion takes it for granted that there is only one legitimate usage of the probability notion, that is, he holds to the "identity

conception" associated with the views of Reichenbach and other frequentists. According to this view, all probability statements, whether they refer to the a priori "likelihood" in idealized games of chance, the empirical frequencies of insurance statistics, the inferred frequency distributions of values of unobserved variables (such as components of the momentum of a hydrogen molecule), or even the probability of theories and hypotheses—*all* these sorts of probability are reducible in principle to relative frequencies; and the justification for a statement of probability always lies, in the last analysis, in the establishment of a relative frequency.

As opposed to this identity conception, we have the distinction made by Carnap (**22**, **23**, **24**, **25**, **26**, **27**) between probability$_1$ and probability$_2$, in which an effort is made to maintain an empiricist definition of factual meaning without reducing every statement of the probability of a hypothesis to a success frequency. I am not competent to discuss the technicalities of this argument, and will only briefly indicate what I understand to be Carnap's position. Consider a hypothesis *h* which we hold with some confidence on the basis of evidence *e.* We say that *h* is probable upon *e* to a degree *p.* If the statement about the probability of *h* upon *e* is interpreted as itself an empirical statement, then it is difficult to give it meaning within the confirmability criterion without interpreting it directly as some sort of a relative frequency, e.g., by ordering *h* to a class of hypotheses of a certain sort whose relative success frequency in the past is fairly well known. But it is hard for people, including many scientists and logicians, to think of the probability of a specific hypothesis as a frequency statement, even an implicit one.

Carnap takes the bull by the horns and attempts to solve the problem by denying that the probability statement relating *h* to *e* is an empirical statement at all. He argues that the relationship of *h* to *e* is a special kind of linguistic relation, different from, but analogous to, the relationship that exists between the conclusion of a syllogism in deductive logic and its premises. That is, to say that *h* is probable to a degree *p* upon the evidence

e is to say that certain kinds of formal relationships, discernible by a study of the sentences in the light of a knowledge of the semantical system of a language, obtain. This kind of probability, which Carnap calls "degree of confirmation," seems to some to be closer to what we think of as "support of a hypothesis" than the relative frequencies of Reichenbach. It is true that the rules for establishing the degree of confirmation of a hypothesis upon its evidence have not been worked out in any detail and in fact are only described by Carnap in general terms for an extremely simple case. For the actual world in which we live, in which the various possible "state descriptions" and their weights which enter into the determination of Carnap's "degree of confirmation" are not even known to us, an actual computation of the probabilities cannot be carried through.

In this sense Carnap's treatment merely gives us a hint as to the direction in which a nonfrequency interpretation of probabilities might proceed. However, so far as I know, the frequentists are in pretty much the same position when it comes to the calculation of actual pragmatic probabilities in scientific hypothesizing and ordinary life. To decide upon this issue is beyond the scope of the present discussion, and in this respect the psychologist interested in Sarbin's point of view will simply have to wait upon the further developments in technical inductive logic. I do not mean to invoke the name of Carnap in *ad verecundiam* against Sarbin, but it is only fair to point out to clinical readers, who may perhaps be unfamiliar with the logic of science literature, that we can quote nonfrequentist scripture when Sarbin quotes frequentist scripture at us. The logical status of probability concepts is one of the most technical and obscure problems of modern philosophy and logic of science, and it would be very dangerous for us to draw any such far-reaching conclusions about clinical methods as Sarbin draws until the logicians have agreed upon the sketch of a solution at least.

6

The Problem of the Logical Reconstruction of Clinical Activity

THE problem with which we are presented is, on the one hand, that of giving a behavioral description of what the clinician does, which is a task of the empirical sociology and psychology of science; and, secondly, carrying out a rational reconstruction of this activity, i.e., showing from the logical standpoint in what way his predictions are related to their grounds. Most of the resistance which I as a clinician feel against the Sarbin-Lundberg interpretation of clinical work springs from the belief that although at bottom, in a most general epistemological sense, their analysis is substantially correct, yet it is stated in such a manner as to give an oversimplified picture of my clinical activities. Lundberg (**70**) has endeavored to reduce this sort of resistance on the part of clinicians by arguing that the whole clinical-actuarial issue is based upon a misunderstanding, and that if the clinician had a really adequate comprehension of the actuarial position he would no longer find the interpretation objectionable. I believe that Lundberg is in part correct in this view, but I shall attempt to show that some of his reduction of the clinical process to procedures which are fundamentally actuarial involves oversimplifications which, if not technically incorrect, are at least so far removed quantitatively from the usual usage of the word "actuarial" that the employment of this word is downright misleading.

In what follows I am not concerned with the empirical question of the relative efficiency of clinical and actuarial predictions (when these terms are used in the usual sense). This is an experimental problem, on which the evidence is as yet inadequate; and what evidence we have will be reviewed later in the present work. Let me state very explicitly that in what follows in the present section I shall be concerned with a purely a priori discussion of the clinical method, and *am not intending to show by any argument whether it is or is not advantageous to make use of procedures over and above an actuarial table or a regression equation.* I shall attempt to show that in *principle* there could be situations in which the Sarbin-Lundberg analysis does not hold up as a description of what takes place, leaving open the question as to whether what does take place "pays off" in terms of an increase in objective success-frequency. I am concerned here with that part of Sarbin's argument which is devoted to showing that it is irrational to *expect* the clinician to improve upon strict actuarial methods, and the allied aspect of his and Lundberg's position that the clinician is always doing what actually amounts to actuarial prediction anyway.

It might seem at first blush that these two opinions are contradictory, but this is not the case. What Sarbin and Lundberg are maintaining is that fundamentally clinical prediction is always actuarial when that word is understood in its broadest sense; and since this is always actually the case whether the clinician knows it or not, it is to his advantage to make these actuarial predictions *explicitly* actuarial. In other words, when Rapaport objects to having a Hollerith machine substituted for his eyes and brains, Sarbin and Lundberg would say that in no case do Rapaport's eyes and brains do anything fundamentally different from what is done by the Hollerith machine, and that the latter mechanism is "obviously" capable of doing the job in question better. If the clinician predicts the future on any basis other than clairvoyance, he is presumably making use of some psychological law. In many actual cases, these laws are of the character of a regression system in that many variables enter into the optimal predic-

tion of the criterion. If a given variable does not, in fact, make any difference, the clinician should not be utilizing it; and taking it into account will not in the long run have any effect except to reduce his accuracy. If the variable *does* have an effect, this effect is measured by the weight which the variable receives in the predictive equation. There is some weight, or more generally, some manner of combining the variables in the predictive function, which is optimal. No combination which the clinician can make can, by definition, do better than this optimal function. It is practically certain that the clinician's brain will not be able to determine the weight as well as the Hollerith machine. Ergo, the clinician cannot possibly do better; and, in general, is practically certain to do worse. It is this version of the actuarial argument which I wish now to consider in greater detail.

For purposes of discussion let us consider the two extreme cases of the clinical-actuarial continuum without prejudging whether there are any qualitative differences. Let us suppose that the factual (observational) material from which prediction is made consists of the protocols of a diagnostic interview, a history obtained from a social agency, and results from a couple of psychological tests, say the MMPI and the Rorschach. We are interested in predicting whether the patient will respond favorably, i.e., remain out of trouble and subjectively relatively free of anxiety and conflict, if he goes unpunished for a delinquency he has committed and is persuaded to change his occupation from F to G and alter his place of residence. Let us take for granted that some reasonably objective criterion of "favorable response" has been set up. I shall assume that the clinician is a skilled psychologist with a wide experience of cases, and that his use of statistics of an explicit mechanical sort does not extend in the present instance beyond the use of the norm data to express scores on the two psychological tests. Any statistical experience for use with the history and interview material and the psychometrics is buried in the reaction tendencies of this clinician's nervous system. He may or may not give verbal reasons for predicting as he does, but at least he does not appear to

proceed in a straightforward mechanical fashion. At the other extreme, let us conceive of a large and complex actuarial table or, alternatively, a multiple-variable prediction equation in which the variables employed are the psychometrics and some quantification based on a classification of events in the history and interview material. A clerical worker is to take this material, enter the actuarial table or substitute in the prediction equation, and grind out mechanically, by straightforward arithmetical procedures, without the use of any judgment or interpretive inference (**61**) a number which represents the optimal prediction of the criterion here involved. That we are usually concerned to predict several aspects of adjustment, and hence would rarely want a single number, is not relevant here. And for any tender-minded clinician who objects to the whole idea, let him substitute a collection of adjectives such as he naturally uses every day. I am interested in a careful scrutiny of these kinds of predictive process both from the standpoint of the behavior of the predicting organism (clinician or clerical worker), and from that of the objective (formal) relation of the prediction to its evidence.

Certain general questions about "lawfulness" and "uniqueness" must be considered before we proceed. I shall assume that there are general laws such as the laws of drive reduction, learning, perceptual organization, and that these laws are known by the clinician. This is not to say that I am giving the case to Sarbin by denying or even qualifying Allport's uniqueness thesis. As Allport has pointed out, his (1937) position does not deny determinism or lawfulness, since the idiographic approach is entirely consistent with the view that such general laws do not preclude uniqueness but are simply the laws describing "how uniqueness comes about" (**4**, p. 558).

This uniqueness is not confined to clinical material, or even to the human case, but holds in the study of all sorts or behavior. In the laboratory investigation of the behavior of the white rat this Allportian uniqueness holds strictly, and for at least two reasons. First, the fundamental laws or the learning process, such as the statement that habit

strength is related to the number of reinforcements by a simple positive growth function, obviously involve the possibility of different values of the parameters. (In what follows, the framework of S-R-reinforcement theory is used; of course the present argument concerning uniqueness applies, *mutatis mutandis,* to any view.) In the second place, the history of no two rats is identical even in a well-controlled experimental study involving the same number of reinforcements. For purposes of the development of nomothetic learning theory, it is convenient to neglect, as in all scientific abstractions, the many individual aspects of the organism's response and to order all responses sharing certain rather rough defining properties to a response class. Until single reaction occasions are thus grouped and equated, it is impossible even to begin counting responses and hence to obtain any measure of response strength, oscillation, etc. The concrete explication, confirmation, or application of a law such as that of habit growth already presupposes certain qualitative decisions. These are necessary (in any individual case) before we can even assign a value of a habit to such a continuum as habit strength. We speak of the rat "pressing the lever," and in general we do not pay much attention to the minor variations in topography and in the intensive and durational properties of the various instances of what is loosely called a response.

As Skinner has pointed out, there is a difference between operationally specifying and identifying members of a response class, which can usually be done to any desired degree of accuracy, and specifying a response class which fractionates the behavior in the way it is fractionated by the organism as a result of its unique reactional biography. The criterion for the behavioral reality of a response class is dynamic lawfulness. In studying the extinction curve of a rat in a Skinner box, we might choose to count only those lever pressings which were made with a force of 4 to 4.5 grams and with the right paw. While this specifies a class operationally, a response class so defined would show a much lower degree of orderliness than one simply defined by the fact that the lever is pressed. Lewin's well-known distinction between phenotypic and genotypic classification is

essentially an insistence that behaviors ought to be classified together not on a basis of arbitrary topographical or other superficial resemblances, but on the basis of their dynamic lawfulness. The definition of *response* is one of the least adequately treated problems in modern rigorizations of behavior theory; for example, Hull's *Principles of Behavior* nowhere gives a general definition of this pivotal notion.

The fundamental correspondence between the human and animal case should not mislead us into neglecting those differences which, even if merely quantitative, are of tremendous importance. The chief among these differences is in *the kind of defining property which is necessary to specify lawful response classes* at the level of human social behavior. In the animal case, we ordinarily have access to an organism throughout its experimental history, and *we* have set up the conditions of reinforcement in such a manner that the *defining properties of the maximally lawful response class are relatively simple physicalistic ones.* The reason that "pressing the lever" is an adequate description of the response in the Skinner box is simply that it is this physicalistically defined property of the response class which we, as the experimenters, have made the condition of reinforcement. The mechanical inevitability of this as we use the recording apparatus makes it easy to overlook the behavioral principle reflected. It would presumably be possible, even in the rat, to define the properties of the reinforced response-class by an ingenious manipulation of the reinforcement history so that a naive experimenter would be hard pressed to specify these defining properties by a study of the behavior. Any defining properties would be characterizable by some disjunction or conjunction of properties of topographic, intensive, and temporal dimensions; but it is clear that they could be made more complicated than is the case when our experimentation is directed at the nomothetic aim of discovering the general laws of learning, and the terms of the disjunction might be very heterogeneous. Even at the level of the rat, there are complications which have, as yet, hardly been touched. The Skinnerian emphasis upon the generic nature of stimulus and response is an important one, but it already

obscures the fact that there is *not* a complete equivalence of all members of a response (or stimulus) class, and that the inductive and extinctive effect of the emission of topographically different class members is not known. The principle of cumulative causation is very important here because of the unknown but possibly marked influence of generalization effects. For example, the previous history of the rat in the acquisition of chain-pulling behavior may alter the characteristics of the modal response; so that whereas, from the standpoint of the experimenter, the reinforcement conditions are the same as for any other rat, the response is harder for the animal. This will result in an alteration of all the parameters of the learning process, and change the quantitative characteristics of the extinction curve. A rat clinician, ignorant of the previous history of chain-pulling experience, might infer, for example, a lower state of drive or a generally greater ease of extinction for the organism at hand, and thus fall into error.

In the human case, this generic nature of stimulus and response presents tremendous difficulties to a physicalistic analysis. To take an obvious example, how do we classify behavior as aggressive? If Mr. B says things which might imply that he has a compulsive, anxiety-driven need for economic status, and subsequently Mr. A, who is usually bored by talk of money, tells Mr. B many things about the tremendous wealth of Mr. C, we are likely to take this as indicating that Mr. A is aggressing against Mr. B. Furthermore, if we know Mr. B very well, we may realize that he is actually not motivated as Mr. A infers and that the "symptoms" of a high economic status drive were actually a function of other aspects of Mr. B's personality. In other words, Mr. A's response is classified as aggression even though it does not tend to inflict tissue injury on Mr. B, does not cause Mr. B any kind of anxiety, and would not be a remark classifiable as aggressive when made to any *arbitrary* member of Mr. A's culture. The behavior which is important to clinicians always involves, at least indirectly, interaction with other human organisms; and the problem of specifying response classes and of taking certain reactions as indicative

of certain habit strengths or states of need is, therefore, a fantastically complicated one. The relevance of these considerations to the problem of prediction by the clerical worker will appear in the paragraphs below.

I do not mean to cast any doubt here upon the epistemological thesis of physicalism. There is no question as to whether behavior protocols furnish the confirmation base of all psychological assertions about others, nor whether any behavior interval can be "described in the physical language." We are concerned here with the *classifying* of such dated behavior-intervals to yield measures of strength and, later, inferences as to the determinative inner conditions. "He took off his hat," "He stood rigidly with hands at sides," "He spoke quietly to the judge," are all behavior descriptions in or close to the physical thing language. But one defining property of this set of responses, by which we recognize a state of *respect,* cannot be stated physicalistically. *The culture reinforces in such a way that responses may covary in strength and yet have no common topography.*

The laws which are of a truly *general* (nomethetic) nature may exist at a much lower (molecular) level of analysis than we generally suppose. For example, the Hullian principle relating strength of response classes to number of reinforcements as an independent variable may itself be a consequence of Guthrie-type laws (which is what a Guthrian would presumably argue). If this should turn out to be the case, an ingenious manipulation of the animal's experimental history might yield a single organism *for which Hullian laws did not hold.* This is not to deny determineism nor to doubt that general laws exist. It is simply to say that such laws as usually obtain are themselves derivative. If there is a sufficient stereotypy due to the experimental traditionalism of ordinary life and of the laboratory, such a fact will not be discovered. Certain initial conditions, and a sufficient isolation of a particular physical system, will lead to exceptionless regularities which are, however, consequences or more fundamental principles. A consideration of a system obeying the same fundamental laws but with different initial conditions will enable us to

discover the derivative nature of the principle that we have been taking as completely general. Most psychologists would probably feel this to be the case with many laws of (capitalistic) economics, for example. All "natural" cats slay rats; but Kuo showed this to be modifiable. Another example would be the close approach of a very large comet upon the motions of the planets as specified by Kepler's laws. The prediction and understanding of the apparent irregularities which would immediately arise require a passage to a more basic level of causal analysis as represented by the formulations of Newton. For a discussion of the general problem of novelty as related to the generality and level of laws the reader may refer to the excellent paper by Bergmann (**10**).

Let us consider the predictive activity of the clinician in the light of these remarks. As clinicians we would usually say that to the extent that we do more than a second-rate job of actuarial prediction, we endeavor to form a conception of "this person"; and it is from this conception, combined with certain admittedly actuarial expectations as to the external events of the future, that our prediction is derived. Most of us would argue, for example, that the behavior we are trying to predict is a consequence of inner variables and is not a *causal consequence* of the facts utilized by the clerical worker. Everyone admits that behavior is determined by the state of the field and organism at the time it occurs. The facts of the psychometrics, the history, and the interview are not related by direct causal laws to the events we wish to predict. The immediate basis of the predicted behavior is the state of the person in conjunction with the assumed future state of the stimulating field. There is, because of our lack of specific information (and our lack of knowledge of laws) merely a crude and fragmentary relationship between the predictive data and our hypothesis concerning the inner state or structure of the person at hand. If this were not the case, of course the prediction would not be actuarial, i.e., probabilistic, but would be strictly deterministic. (I neglect here what I consider to be Sarbin's misapplication of the Heisenberg principle to the behavior case, for a detailed refutation of which see London (**66**).)

What the clinician does is to utilize the given facts, together with crudely formulated laws, to invent a hypothesis concerning the state of certain intervening variables or hypothetical constructs in his patient. On the basis of such diverse evidence as the Rorschach and Multiphasic profiles, a slip of the tongue during the interview, and a social worker's description of the patient's mother, the clinician arrives at such statements as "this patient has strong oral-dependent attitudes, against which he has set up dominant-aggressive reaction-formations." Statements of this sort, which are often mixtures of propositions about habit strength, generalization gradients, topographic properties, drive levels, and even the *parameters* in learning functions and satiation functions themselves, are all covered by the phase "forming the concept of this person." I do not believe that as clinicians we ought to be threatened or feel depreciated by such a general (and correspondingly empty!) formulation of our activities. To say that in forming a conception of a person I am assessing his needs and his modes of satisfying those needs (including the all-important need to reduce anxiety, and with it the immense collection of self-reinforced habits which we call his *defences)* in no wise detracts from a recognition of the tremendous possibilities for variations and complications that arise when a more specific description of these needs and habits is undertaken in seriousness. Let us now ask, what would Sarbin have to say about this process as contrasted to the activity of the clerical worker?

In the first place, he would point out that the "laws" which the clinician makes use of are actuarial. Certainly this is true, at least in the sense that all laws are based upon inductions, and all inductions are actuarial in the general sense of Reichenbach. There is no reason why the clinician should be hesitant to admit this, *so long as he detects no equivocation in the word "actuarial,"* i.e., so long as the philosophic or epistemological use of the term "actuarial" is not surreptitiously changed into the more customary use, in which we speak about statistical tables the elements of which are persons. In order to avoid any possibility of this confusion, I shall rephrase Sarbin's viewpoint more neutrally, and

simply say that the clinician ought to admit freely that the laws he employs are inductive. This is clearly trivial.

It is perhaps worth mentioning that some of the laws which the clinician uses are not laws in which the present *S* and *R* facts occur as independent and dependent variables, respectively. That is to say, some of the laws are correlational R-R laws (cf. Spence, **93**) and others are rather ill-established laws concerning hypothetical inner events (cf. Feigl, **41**, p. 42; Spence, **94**, p. 73). That these laws must have been *suggested* initially by observations of behavior, and that they must ultimately be supported by behavioral data, is not tantamount to saying that they are laws relating *directly* the data given the clerical worker to the behavior which she is asked to predict. I am not interested here in the question of how well such laws are supported at the present time, but simply wish to indicate that the statement "If the clinician uses laws he must be proceeding on the basis of some previous inductive experience" does not necessarily imply that such laws and his use of them are of the same sort as the multiple regression equation.

Can this performance, "forming a conception of Patient A," be duplicated by the clerical worker? I am sure that no one will seriously maintain it can in *fact* be duplicated by the clerical worker; the question is whether it could be duplicated in principle. Here we are on very dangerous ground and I do not have any dogmatic pronouncements to make. I should like simply to raise some questions which I think cast doubt on the view that, in principle, the clerical worker could here duplicate the predictive behavior of the clinician. In the first place, certain facts will be seen by the clinician to support hypotheses as to the internal economics and dynamics of the patient, although instances of these facts simply do not occur in the actuarial table. It may be asked, *how can* they be seen to support the hypothesis, unless there is a second-rate actuarial table in the clinician's head? And, if this is the case, all we need to do is to get that table out of the clinician's head and on paper, and we will shortly discover that the clerical worker can do a better job because the actuarial table will assign a

better weight, In spite of the plausibility of this argument and a personal disposition in its favor, I remain suspicious of it, What appears to me convincing when thus stated in abstract terms, seems very unreal when I consider concrete cases. Let me give a clinical example of the sort that I cannot readily fit into this mold.

A patient has been developing insight into her ambivalent attitude toward her husband. She begins to show some gross manifestations of hostility against him; for example, she tears up a series of short stories he wrote some years ago, telling him he knows perfectly well that they were no good anyway, Do we deal here with a relatively unmixed expression of hostility previously repressed by the patient, or are there other components in her need structure contributing to this behavior? She reports that one evening, feeling very nervous, she went out alone to a movie; and as she was walking home, wondered if he would be "peacefully sleeping" upon her arrival. Entering the bedroom, she was terrified to see, for a fraction of a second, a large black bird ("a raven, I guess") perched on her pillow next to her husband's head. Asked to give her thoughts in connection with a raven, she says that she shouldn't have called it a raven, it was probably just a crow; in fact she doubts that she said raven in the first place. Insistence that she did say raven elicits irritation. She recalls "vaguely, some poem we read in high school, I guess I don't know anything else about it."

What prediction enters the listener's mind with this reference? The prediction is mediated by a miniature dynamic hypothesis. The reference is almost certainly to Poe's poem; one guesses that the thematically important content determining her hallucination is connected with the preceding thought about her husband peacefully sleeping. The hypothesis forms itself: Nervous and upset, she goes out alone to a movie while her husband, unmindful of her, is able to "sleep peacefully." The fantasy is that, like Poe's Lenore, she will die or at least go away and leave him alone, with the bird croaking "Nevermore." Then he'll be sorry, not able to sleep peacefully, etc. We formulate the further hypothesis, which

includes our hypothesis about the determination of the particular hallucination, that she is concerned about her husband's need for her, and would like to know how important she is to him. This leads to a prediction as to the leading themes we expect in the rest of the session. The prediction has a wide latitude, i.e., a *class* character is specified for the behavior, as always, But we anticipate that her (unguided) associations will touch upon the theme of punishing her husband, by going away somehow, that he would be sorry if she did, and the like. We also permit ourselves some leeway as to time, in that the development of the theme may not begin strongly until the next session, etc. But we do not make a vacuous prediction, since *some* manifestations of the Lenore fantasy are to be expected, and fairly soon. Her subsequent remarks in the same interview return repeatedly to the general topic of her husband's lack of concern for her condition, and his "sublime confidence" that she will "never do anything rash," which turns out in further talk to cover both suicide and unexpectedly leaving him. Fortified by these confirmations, we begin to attach considerable weight to the hypothesis that her hostile reactions are overdetermined, being in part attempts at testing the limits of his love and acceptance. Systematic attention to this hypothesis is well rewarded in the succeeding sessions.

The interesting question here is this: What are the general statistical uniformities which are allegedly able to generate the initial hypothesis? I presume the situation of a woman hallucinating a raven next to her husband's head is unique, and hence cannot define a reference class for any relative frequency, either known or unknown. To what larger class can the event be ordered? It would be a nonsensical classification, and would completely cut across the categories and dimensions which are really involved here, to consider the *obvious* larger classes, e.g., having hallucinations of birds. I do not suppose anyone would seriously maintain that hallucinating birds is statistically associated with the desire to test a husband's love, or the unconscious fantasy of leaving him. The general principles involved here are not difficult to state; but what

impresses me is their relatively vacuous character *insofar as generating the particular hypothesis is concerned.*

We are making use of such general statements as these: "When a person describes an experience and subsequently corrects his description and refuses with some emotion to admit to his original description, it is frequently the case that the original description was correct and that it involves material which is dynamically important and which must be defended against." "The only poem involving a raven which is read with any frequency in high school classes is Poe's poem." "One basis on which a literary production may be associated with a situation or state of need in a person familiar with it, is an unconscious identification with one of the characters or an identification of one's situation with that portrayed." These are the principal statistical generalizations which form the matrix for the construction of the present hypothesis. I take it as obvious that the hypothesis could not be mechanically ground out from these statements, even if the frequency words were replaced by numerical frequencies, through an application of probability calculus. The hypothesis "She saw a raven because she was thinking of herself as dead or departed, which would injure her husband and also make him realize how important she was" is psychologically *suggested* by these facts; and I think it is fair to say it would not be suggested to a clerical worker, even if she were fully cognizant of the meaning of the above general propositions. It seems to me that even if we acquaint the clerical worker with the statistical frequencies and make sure that she understands the meaning of all the concepts involved, we still have to create in her a readiness *to invent particular hypotheses that exemplify the general principle in a specific instance.* And when we have done this last, which I do not think can be done *wholly* by stating general rules, we have trained a clerical worker to the point that she is now actually a skilled clinician.

Reik (**83**) gives numerous examples of clinical hypothesis formation, which are instructive (and discouraging) to try to formulate actuarially. A fascinating case of *postdiction* based on only a fragment of behavior during analysis:

One session at this time took the following course. After a few sentences about the uneventful day, the patient fell into a long silence. She assured me that nothing was in her thoughts. Silence from me. After many minutes she complained about a toothache. She told me that she had been to the dentist yesterday. He had given her an injection and then had pulled a wisdom tooth. The spot was hurting again. New and longer silence. She pointed to my bookcase in the corner and said, "There's a book standing on its head." Without the slightest hesitation and in a reproachful voice I said, "But why did you not tell me that you had had an abortion?" (**83**, p. 263.)

Reik gives us his introspection on this bit of postdiction, to which I refer the interested reader. But let us ask, how can we arrive at such a postdiction actuarially, making the generalizations and frequencies explicit so that the clerical worker can duplicate Reik? The tooth extraction as a symbol of birth we can put into a crude actuarial "law." The silence, we can teach our clerical worker, is usually resistance, conscious or unconscious. Where does this leave us? "There is a probability P that the patient is resisting something about birth." So far, so good. But this interpretation does not have the dramatic, time-saving quality of Reik's, and it is much less specific. How work the "book on its head" into the actuarial mold? This fragment gives Reik his image of the fetus, and hence mediates the final touch of his postdiction. Speaking in very general terms (and it is impossible to speak otherwise, not merely because of the inadequate state of theory, but because the kind of behavior with which we are here dealing has an intrinsic vagueness, involving continuous gradation in topography and marked variability from one individual to the other in the defining properties of the response class), we might say that "any words or images which indicate properties belonging to a fetus may acquire induced strength from mentation concerning fetuses." It will be necessary to sensitize the clerical worker to this very broad defining property, so that when a specific member of the response class occurs, never before observed and hence not present in an actuarial table even of colossal N, he will respond to it as an instance of the class.

I do not mean to use the word *sensitize* in any mystical or undefinable sense. I refer simply to the fact that categorization of a particular patient's response by the clerical worker is in itself a *response,* which must become elicitable by a great diversity of patient responses seen as clinical stimuli. The defining properties of this latter class will, in general, not be simple. The majority of the individual forms (physicalistically defined) will not be listable in an actuarial table, partly because they will simply never have occurred in any recorded clinical experience to date; and partly because the number of them, thus botanized, would become too cumbrous for any practical use. The complicated kinds of mutual interaction between internal variables and external events which characterizes human clinical material result in a situation in which a response having a specified topography, emitted in a specified stimulus field, may indicate different states of internal variables depending upon all well-confirmed hypotheses about the individual. In ordinary life, we recognize this when we say that the same behavior will suggest, in the extreme case, even opposite interpretations when the behavior occurs in two individuals concerning whose personality structure we have already considerable knowledge.

What I am trying to indicate is that the general laws which relate the strength of responses to certain antecedent conditions even when they are adequately worked out, have to do with the *form* of behavior covariation. But we can only talk about these laws as applied to a particular case when we have already specified to some degree at least *what end terms* (stimulus class, response class, and so on) are involved in the particular case. This amounts in itself to the setting up of particularistic hypotheses as to the unique organization of stimulus and response classes and the intervening variables of the particular patient at hand. A partial formulation of some of these hypotheses will lead to a basis for classifying other responses by the same individuals. I think it is this kind of relationship which clinicians have in mind when they emphasize the necessity for knowing the "meaning" of a given segment of behavior to the "whole person."

Although the word "meaning" is sometimes used chiefly for its rhetorical effect, it seems to me that it indicates in this context a genuine problem in the classification of behavior.

We could presumably train the clerical worker, both by a feeble attempt at stating general properties of a given response class, and by the multiplication of many instances, to respond to a segment of a patient's behavior in this categorical fashion. Unfortunately, the verbal response "there's a book on its head" is only one of a thousand different sensitizations which must be achieved if a clerical worker is to be able to order behavior to such meaningful classes. Suppose we indoctrinate the clerical worker with the whole system of dynamic theory by means of which individual behavior segments are seen as supportive of this or that particular hypothesis; and then we make this abstract knowledge available for practical use by exposing the clerical worker to innumerable instances of each sort. It seems to me that this is the only way in which we can avoid the consequences of the uniqueness for the mechanical application of an actuarial table. With Sarbin and Lundberg, I argue that every skilled clinician must be making use of some laws, however vague, which may be of considerable generality, but which nevertheless make it possible for him to order his material with respect to a given patient in terms of some general nomothetic basic psychodynamics. The problem is, however, to make these highly general laws available to the clerical worker, and to build into her nervous system the appropriate reaction tendencies so that she can use them in the formulation of the individual case, many if not most of whose evidential behaviors will have occurred too rarely to be in any actuarial table. That is, a set of *kinds* of hypotheses such as "this man has set up a reaction formation against impulse *X*," as well as a readiness to perceive a physicalistically diverse collection of behavior segments as supportive of this or that particular hypothesis, must be taught to the clerical worker. In principle, given sufficient intelligence and motivation, there is no reason why this cannot be done; but as I have indicated above, such a trained clerical worker has been made into a skilled clinician.

To summarize the argument just completed, one might proceed somewhat as follows. The so-called general nomothetic laws of behavior are laws relating responses to stimuli via certain intervening variables whose states are in turn specified by antecedents, e.g., hours of deprivation of food. Presumably the form of these laws is genuinely nomothetic for a given species. The parameters vary from organism to organism within the species but are in principle inferable from values of other parameters and from certain combinations of dynamic changes in strength which are themselves observable. The end terms involved in these laws however, are variable from organism to organism; the same is true of the intervening variables. That is to say, the intervening variables relate the facts only via *tentative response classes.* It is necessary for the observer to have in mind habits, traits, derived needs, and the like before he can see how a given behavior datum supports propositions concerning the state of these variables.

If there were a very small number of habits, all manifesting themselves In the same way and having little or no variation in their topography, it would be a simple problem of inverse probability to construct a particularistic hypothesis concerning the system of inner states necessary to account for the observed behavior strengths of individuals. But in fact the "$_sH_R$" involves an *H* whose dimensions (or, properties) vary greatly from individual to individual and from time to time. Consequently the formulation of the state of a particular organism involves the hypothesization of the *forms* of a set of H's (and analogously, a set of D's for drives). It is not feasible literally to list all the tremendous collection of such habit and need forms in an actuarial table. First, because really new forms constantly occur; and, secondly, because particular combinations of such narrowly specified forms will have no entries in such a table even from an extensive clinical experience. On the other hand, it is very difficult to specify such classes by their general defining properties, because in the case of human social behavior the defining properties are, in general, not physicalistic. One has to think of the hypothetical habits, traits, or needs, including

specifications of individual properties in some detail, *before he can understand that a given behavior datum supports hypotheses concerning it.* If one listed enough concrete possibilities to give an idea of the response class, and indicated in terms of general causal laws how they could be grouped together, and had the clerical worker overlearn these so that they were at sufficiently high strength in *her* verbal behavior to come out in a particular clinical instance, the clerical worker would have been transformed into a skilled clinician.

No matter how convincing the previous considerations seem, one still has the uneasy feeling that something must be wrong. I find in myself the tendency to say something to this effect: After all, the behavior *is* lawful. If it is lawful, everything about it, including the topography and the dimensions of particular responses, must be a function of some variables which are determinable. Therefore, it is not possible that the clinician could do anything that the clerical worker could not do in principle. The apparent inconsistency of this train of thought with what precedes can be, I think, readily resolved. It is a tautology for a determinist to say that if we knew all the parameters in all the equations of the behavior acquisition functions for an organism at birth, and *if* we knew all the situations to which he was exposed, the specification of the response classes would follow directly from this knowledge and everything would proceed in a "mechanical" fashion. But these initial parameters, and these previous experiences, are not known to us. I think it is no exaggeration to say that they will *never* be known to the practicing clinician. The experiences which determine the topography or a given response class and the mutual interrelations between needs and habit strength are for the most part permanently inaccessible to us when we come to consider the adult organism. The most fantastically detailed social history and the deepest psychoanalysis could only, from the purely physical standpoint or the nature or verbal descriptions, give us a fraction of all the events in the reactional biography which has determined what the individual is at the present time. Furthermore, these events are not available in any record or

anyone's nervous system regardless of the time and effort that we would be willing to put forward in obtaining them. Nobody *knows,* at the present time, what the patient's older brother said to him at the dinner table when the patient was four and one-half years old. What we see before us is the cumulative result of literally thousands of single learnings and unlearnings, not the least of which are those elusive kinds of learning which are involved in the internal responses which we call fantasy, not observed by anyone and since forgotten by the patient. For this reason, we are perpetually in the situation of trying to reconstruct initial conditions from a study of the results. It is here that the necessity of being able to think up the best hypotheses concerning the organization of the individual's personality arises, in spite of the assumption of complete determinism.

I think that there may be a formal difference in the *process* of prediction when it is carried out actuarially (in the Sarbin-Lundberg sense of that word) and when it is carried out by the skilled clinician via the use of a hypothesis. In the actuarial case, let us suppose that the event to be predicted is a simple dichotomy, e.g., violates or does not violate parole. A finite although possibly very large set of facts is known about the individual and the particular combination of facts defines a subclass of the population of individuals for which certain relative frequencies have been determined. It may happen that the particular combination before us has never heretofore arisen, but that some inductions of a higher order have led us to the statement that certain frequencies are independent of others, some frequencies change in a specified manner with a value of certain properties, and the like. To arrive at a prediction for the case at hand we need only apply the probability calculus in a straightforward fashion and thus arrive at a number which automatically determines what we predict. While the prediction considered as a statement about the future is not a deductive consequence in the sense that it does not follow necessarily but rather in probability, the probability number *reached* is a purely deductive

consequence of the initial set of probability numbers, together with the rules of the game. If we now simply add the usual decision to predict always the more probable occurrence, the arrival at the prediction is obviously a matter of sheer deductive manipulation of a mathematical sort.

But if the prediction flows as a consequence of some sort of *structural-dynamic hypothesis* concerning the personality, the formal situation is different. For this hypothesis is not itself in any sense a formal consequence, i.e., it is not straightforwardly deducible from the facts which support it. When the hypothesis has been stated, the original data are seen as entailed *by* it, in conjunction with the general laws and the rules of inference. But someone has to state the hypothesis in the first place. It is in the initial *formulation* of the hypothesis that there occurs a genuine creative act with which the logician, as such, has no concern. There is a stage at which someone must have thought up a hypothesis which, in the context of discovery, was, to be sure, suggested by the facts, but is not a formal consequence of them. Whereas in the actuarial case, the frequency for a subclass is a formal consequence of the application of the principles of probability to a set of data.

Consider a nonpsychological analogy. Let us suppose that we have before us an opaque box, on one side of which is a row of ten buttons. A pressing of any three of these buttons constitutes a stimulus so far as the box is concerned. On the other side of the box is a row of ten colored lights, whose pattern of flashing on and off exhausts the box's potentiality of response. Let us suppose that the internal mechanism of the box involves a more or less complicated series of interrelated gears, brackets, pulleys, springs, sliding surfaces, and the like. Such a box is capable of being stimulated by one thousand distinguishable stimulus patterns. Suppose now that we permit an actuary to make a finite set of observations upon the stimulus-response connections of the box for statistical purposes. Certain rough probabilistic relations will appear. For example, he might find that when any button is pressed twice in succession, then, whatever is

done on the third pressing, 90 per cent of the time the response involves a turning on of six lights. He might also have made such observations on numerous boxes of the present sort whose mechanisms were similar to (but not identical with) the present one. If we now ask him to predict the results of a certain combination of button-pressing which he has never tried in his sampling of the present box (or possibly not in any box that he had studied), he would have to be content to make a guess on the basis of some larger class of pressing combinations of which the specific combination we mention is a member.

Suppose now that we presented a similar problem to a skilled mechanic who had dismantled many such boxes in addition to having observed properties and their frequencies. With a small number of pressings, in this case very carefully chosen, he could conceivably be led to the *formulation of a hypothesis concerning the particular structure of the internal mechanism.* It is true that this hypothesis formed by him might be erroneous. But it also might be correct; and *if* correct, would lead to definite predictions, having a very high success frequency.

It might be objected that we have subtly included actuarial information by specifying that the skilled mechanic "has taken apart many such machines." This is admitted. As I have indicated earlier, if the word "actuarial" is used as an equivalent of "experiential-inductive," then the only clinicians who would deny that they operate actuarially are those who claim to be prophets and clairvoyants. But I have tried to make clear that this use of the term "actuarial" is so broad as to remove all meaning from the issue at hand, and to take all the sting out of Sarbin's argument. Furthermore, it is one thing to say that it is *necessary* that the mechanic should have certain actuarial data to be able to formulate a good structural hypothesis, and it is quite another thing to say that such actuarial data are *sufficient* for him to arrive at the prediction, were he not skilled. By this I mean that he is enabled to invent a particular hypothesis concerning the inner workings of the present box because he has had experience with such boxes in the past; but the hypothesis at hand is not

something *derivable* as a mechanical or statistical consequence of the set of frequency statements which actually make up his previous experience. Being a skilled mechanic means that, on the basis of his actuarial experience, his brain has become capable of the creative act involved in formulating a hypothesis about the present unique box.

I think it is obvious that we could present the statistician with a table of relative frequencies concerning numbers of gears, positions of pulleys, and the like for the same sample of boxes which the skilled mechanic has dismantled, without having any assurance that the actuary would be able to invent the correct hypothesis. Whether in the long run predictions arrived at by the creation of such hypotheses are more trustworthy than those arrived at by a straightforward application of the frequency tables is an empirical question which would depend factually upon such things as the degree of complexity of the parts, the skill in hypothesis-making of the mechanic, and the size and diversity of the sample available to the actuary. I have merely tried here to indicate by a mechanical example the *kind* of situation which I feel is involved in high-level clinical activities.

A learning history begins with an organism for which the parameters occurring in the functions descriptive of the learning process are different from those parameters in other organisms. There are individual differences in initial behavior readinesses, e.g., in the susceptibility to anxiety, ease of producing crying, and the like. There are individual differences in behavior aspects which are in some degree irrelevant to satisfactions of the drive which becomes connected to them but which may later on in the history acquire significance. Such "temperamental" variables as energy expenditure and response tempo may vary widely from member to member of a response class defined by a very broad topography sufficient to guarantee the reinforcement and hence to maintain the strength. These expressive aspects of the behavior may take on a positively or negatively *adaptive* function as well—when, for example, the social stimulus value of the individual acquires a different sort of relevance for his rewards than

was the case at the time the behaviors were being firmly patterned. To the extent that secondary (derived) needs play an increasingly important role in the determination of behavior we have even greater possibilities for variations. The goal-states-of-affairs which are reinforcing are themselves rather complex stimulus configurations and sequences, the physicalistic properties of which may be difficult if not impossible to specify for the whole group of organisms, even those having a certain homogeneity of the life history as is guaranteed by a common culture. Since derived needs are so heavily stimulus-dependent, they are to be specified by the class of configurations which apparently reduce them, the usual indicators of docility, etc., being used to identify this class.

The uniqueness of the learning history brings about a uniqueness in the defining properties of the stimulus class which *constitutes* a reduction for one of these higher order needs, and *thereby brings about a uniqueness in the needs themselves.* To the extent that a very large part of human behavior is maintained on the basis of anxiety reduction and is heavily dependent upon a large and complex set of verbal and other symbolic social- and self-reinforcements, there is a perfectly legitimate sense in which we can say that the important *needs* considered by clinical psychologists are idiosyncratic in a way that the drives we study in the animal laboratory are not. I do not mean to suggest that the hunger drive of a given rat is not unique, since, as stated above, I would argue for the literal truth of Allport's view in the animal case as in the human. But the *extent* to which the sugar-hunger of a rat in one of our experiments has about the same quantitative characteristics and appears in the same role as a variable as in the case of another rat is presumably much greater than the extent to which artistic interest shows person-to-person similarity in the human adult.

It is likely that in addition to marked individual differences in primary stimulus generalization gradients, there are kinds of derived or learned generalization gradients which result in very different potentialities. We are presented not only with the differences in response dispositions, but

with differences in the disposition to acquire dispositions of various sorts. The principle of cumulative causation operates here so that the effects of relatively minor fluctuations in initial conditions may produce fantastically great complications (cf. London, **67**). We have, in the adult, variations in *needs,* not merely will their strength *but also in their defining properties*; variations in the defining properties of those habits which are cued to these idiosyncratic sets of needs; variations in the defining properties of the stimulus classes which perform both the cue and the reward function with respect to the need-habit accommodations involved; and finally variations in the functions relating some of these to others, as in the extent of generalizations from one class of needs to another or from one class of habits to another. The system of needs and associated response tendencies including the interrelations among them is often referred to as the "personality structure." The word "structure" here is perhaps not too happily employed, since it fosters some rather noncontributory imagery. Nevertheless, it seems to me there is a legitimate sense in which the properties we think of as structural apply here.

In the first place, there is an element of relative stability. Certain response dispositions may be modified by experience, but always in terms of rather permanent second-order dispositions such as referred to above. In order to predict what a given human being will learn when put in a specified situation or sequence of situations, it is not, in general, sufficient for us to have knowledge of the general laws of learning, e.g., the principle of reinforcement, generalization, the multiplicative function of drive, and the like. We ought to know, for example, the behavior readinesses (initial strengths) which he brings to the situation, since those responses with a little greater initial strength will cumulate their advantage by occurring and being given reinforcement before the alternatives have an opportunity to appear at all. Even if these readinesses were known and the organization of the environment were also known, so that we could predict the initial members of the response series and their stimulus consequences, we would still have to know the reinforcement

properties of these stimulus consequences. This in turn involves goals, that is, the rather complicated properties (or, better, dimension values) of the individual's needs. In ordinary life, for instance, we do not attempt to predict with any confidence what a civilian will learn from three years of exposure to the rather homogeneous environment of military life, without having a little knowledge of his civilian personality. Whatever more permanent second- and higher-order dispositions are involved, they constitute a sort of stable structure in which particular learnings occur.

A second sense in which the structure notion applies is that of levels or layers. We observe that A, who ordinarily approves of B, becomes touchy, moody, and irritable when he goes shopping with B. This we explain by showing that B's conduct toward salespeople makes A feel inferior, since A is unable to avoid buying things that he does not want from an aggressive salesperson. Why is this? We explain this fact in terms of a general disposition to be passive, overly compliant, and generally fearful of arousing the antagonism of others. Why does he have this characteristic? We explain this in terms of an overlearned reaction-formation against his own hostility, the strength of which is maintained by its anxiety-reducing properties. How was this learned? We look to his life history to find out why early manifestations of hostility were more anxiety-arousing in him than in other people. In terms of the historical sequence of the successive acquisition of members of this response chain, in terms of the (truncated) sequence of response dispositions at present, and finally in terms of the degree of defense against recognition ("depth") of the components of the sequence, there is a sense in which we can speak of "layers" of the personality and, hence, of a structure.

Dr. David Grant has suggested to me that even though certain predictions about human individuals may be based upon relatively complex hypotheses concerning structure in that sense, and hence are not derivable from the data by a direct application of probability calculus unmediated by the intervening step of hypothesis formation, nevertheless the formu lation of the hypothesis itself theoretically can be handled in terms of

Bayes' Theorem. I have no wish to be dogmatic on the point but I am not persuaded that this is the case. In order to apply Bayes' Theorem, it is necessary that we should have before us a set of alternative conditions, for each of which an initial probability is known, and upon each of which the probability of a certain sign or symptom is also known. How is the set to be specified in the personality case? I leave out entirely the pragmatic question, whether we have even approximations to the actual probability. It is not clear to me what this distribution of generating alternatives could consist of.

For each individual we have, in principle, a set of structural hypotheses concerning his unique organization of needs and habits. I find it difficult to imagine what hypothesis about personality, or about a segment of personality, corresponds to the Bayes urns in this case. It is true that *some* kind of inductive evidence must be the basis of deciding that a given hypothesis about the personality structure probabilistically entails some part of the evidence we have before us, but I have tried to make clear that this much of reliance on previous experiences is not precluded by Allport's views or so far as I am aware by those or anyone else. But it seems to me that the formulating or this hypothesis amounts to the hypothesizing of a new urn, with a certain distribution of marbles in it. And I fear that the formulating of this hypothesis, when it acquires any appreciable degree of complexity, is precisely that creative act which is possible only for the clinician. You cannot apply Bayes' Theorem to a problem until you have specified the initial conditions; and this means to state what are the various urns, and what are their contents. If we reject a categorical analysis and recognize that we deal not with response and stimulus and need *classes* but rather with clusters, the elements of which differ on a whole set of dimensions, we have then the continuous form of Bayes' problem.

I suppose one must admit that in principle, perhaps in a "behaviorism" stated at the micro-level, the procedure could be carried through. But we are so very far from even approaching such a situation that the direct

synthesizing of such a hypothesis which is, so to speak, merely suggested by the data is the only procedure applicable in practice. To return to the analogy of a criminal investigation, it seems to me that Dr. Grant's suggestion is like saying it is not necessary to make use of the hypothesis-forming skill of a police detective, since presumably the physical location of a person is distributed according to some as yet unknown probability and the likelihood of his behaving in a certain way in a specified situation also has some definite although unknown probability. Both of these statements are correct, of course. But in the case of a particular murder, what is required is somebody who will think up a specific hypothesis concerning an event sequence, that hypothesis not being constructible by any mechanical rules for combining the "distribution of people in space-time"; although once the hypothesis has been *conceived*, certainly nomothetic laws about behavior, properties of blunt instruments, and the like are utilized to show that it is confirmed in such and such a degree. In the same way, a statistical analysis of a distribution of frequencies of numbers of cogwheels, coefficients of friction, and the like would be of some, but insufficient, help in attempting to invent a hypothesis concerning the sealed boxes in our mechanical example. No one is denying that it is precisely these distributions of occurrences in his own past which have *eventuated* in the skill of the clinician. But this is not tantamount to saying that a nonclinician could create the same hypothesis the clinician creates, by a mechanical treatment of the distribution frequencies, even if known.

The relation of lawfulness and uniqueness to the problem of the clinician's contribution might be put in summary fashion thus: A *law,* such as the law relating habit strength to number of reinforcements, is (1) in its *form,* nomothetic for a given species (or an even larger biological group); (2) in its *parameters,* idiographic but perhaps inferable, on the basis of second-level nomothetic laws, from other parameters estimated on the given person; (3) in its *end terms,* i.e., in the defining properties or dimensional ranges that specify "S," "R," "G," "D," etc., strongly idiographic.

Since the history generating (3) is precisely what we do *not* know when confronted with the patient, the clinician must reconstruct it, *and from fragments chiefly on the dependent-variable side.*

Philosophers of science usually distinguish between "general hypotheses" and "particular hypotheses." The first are exemplified in such hypotheses as that of universal gravitation, the atomic theory, and the kinetic theory of gases. The latter refer to hypotheses concerning the state of affairs in a given space-time region; as, for example, that the American Indian came to this continent from Asia, or that Bruno Hauptmann was the murderer of the Lindbergh baby, or that the solar system was formed by a passing star. The setting up of hypotheses of the first type involves a special creative act in which the scientist has to "see" that the facts *e* at hand could be deduced from the hypothesis *h.* Presumably the difficulty of this seeing would be in considerable degree dependent upon the similarity of the hypothesized entity or process to things already familiar. From the methodological point of view, the formation of particular hypotheses is a different sort of thing; but seen *psychologically,* it might be said that where the variables are extraordinarily complicated, and knowledge relatively scanty, *the psychology of the hypothesis-forming act may be rather similar in the particular and in the general type.* What I am suggesting is that high-level clinical hypothesizing partakes to some degree of that kind of psychological process which is involved in the creation of scientific theory. It is from this point of view that one can do justice to the intuitive and nonrational element of clinical work without committing oneself to any unscientific heresy. For example, analysts have spoken of the "resonance" of the therapist's unconscious with that of the patient.

Freud says: "Expressed in a formula, he must bend his own unconscious like a receptive organ towards the emerging unconscious of the patient, be as the receiver of the telephone to the disc" ("Recommendations for Physicians on the Psycho-Analytic Method of Treatment," *Collected Papers*, II, 328). Reik has indicated a similar thing in his use of the

term "conjectures" (**83**). (See also Fenichel, **43**, p. 5.) The important point here is to realize that what these authors are discussing comes under the heading of Reichenbach's "context of discovery." Having once *conceived* a particular hypothesis concerning a patient, we must, if we are scientific (I should be inclined to say even *rational*), subject this hypothesis to the usual canons of inference. That is, we must see whether the hypothesis will entail more of the known facts than others, a greater range or diversity of the known facts, will enable us to make predictions that will square with general principles arrived at by previous inductions, can be fitted into the nomothetic scheme at the next lower level in the explanatory hierarchy, and so on. This is Reik's "comprehension." I do not see how any honest clinician can avoid answering these questions about his own hypotheses. But probably what has led some clinicians to *talk as if* they did not accept the usual principles of justification has been the failure of some nonclinical critics to do justice to the complexity and subtleties of the preliminary stage, i.e., of the events occurring in the context of discovery. As Fenichel says, the difference between psychoanalysis and the other sciences with regard to the role played by the unconscious is a quantitative one. When we ask "How did clinician A *arrive* at hypothesis *h?"* we are asking a psychological question, and we are talking about events which must not be dealt with in a simple-minded fashion if the psychology of the creative act is to be unraveled. When, on the other hand, we ask "How could the hypothesis *h* be *justified* (by clinician A or by anyone else) in rational activity?" we are asking a logical question in the context of justification. Clinicians and their scientific critics are often at loggerheads because the contexts of discovery and justification are not kept distinct in conversations on clinical activity.

I should like to mention a few analogical cases in order to minimize the discomfort that may be felt by any actuarially inclined reader. The situation in clinical psychology is in this respect similar to that in criminology, engineering, or any other applied subject matter. If we ask the opinion of an expert engineer concerning why a certain bridge has

collapsed, it is obvious that he makes use of certain general principles of mechanics, and in that sense he is proceeding actuarially. It is also obvious that he makes use of his experience with collapsed bridges which is, in a sense, also actuarial. Nevertheless, an engineer under these circumstances does not sit down with a table of relative frequencies of bridges of this and that sort, built in this and that circumstance. Having gathered the facts, he attempts to state a hypothesis which may involve the assumption of a state of affairs which has not arisen with any previous bridge that he has studied; conceivably, even, which has never existed with respect to any bridge in the world before. Admittedly, his choice of hypotheses will be determined in part by certain vaguely known initial probabilities, e.g., he may try to avoid a hypothesis which involves as one component the assumption that a certain type of metal was badly cast, because he knows that this sort of thing hardly ever happens. The clinician should be willing to admit that he could hardly fail to gain by having comparable statements of his working assumptions made numerically explicit. That *part* of the clinician's thinking which involves the use of empirical frequencies could not fail to be improved by having those frequencies objectively determined in a table rather than subjectively stored up in his skull. Where the actuarially minded critic is in danger of going astray is in inferring too much from the "obvious" superiority of explicit relative frequencies over vaguely apprehended trends, thinking that the combining of frequencies is a full account of the process of prediction even in those instances where a particular structural or historical hypothesis is utilized in making the prediction. The engineer will surely make fewer mistakes if he has a handbook giving the range of tensile strength of various alloys than if he comes to his hypothesis-creating with vague and partially erroneous judgments on these matters. But whether the distribution of tensile strength is known impressionistically or in terms of an explicit frequency table—in either case some of the available hypotheses will not occur as mechanical consequences of his data.

7

Remarks on Clinical Intuition

ALTHOUGH it is not the primary issue, I should like to make a few remarks here concerning clinical intuition. In discussions of statistical method in the clinical setting, we often hear that the conflict is between mathematical and so-called intuitive procedures. Although we clinicians talk a good deal about this, we do not know very much about it and there does not seem to be much American investigation of it with the exception of some of the older work of Allport and his students. Without attempting to review what experimental material we have available, it may be profitable to say a little bit from the armchair. It seems to me, from observations of my own clinical activity and that of other workers, that the phrase "clinical intuition" commonly covers two rather different situations.

The first, and the one which seems most irritating to nonclinicians, is the situation in which a clinician responds with a diagnostic, predictive, or postdictive statement about the patient and when asked for the evidence states simply that he feels intuitively that such-and-such is the case. "My third ear tells me…" or "I don't know, but I feel very strongly about this patient that…" or "He gives me kind of a schizy feeling," or "I think that if one has seen very many psychopaths of the Pd type, he cannot fail to see it in this patient," or, a somewhat more sophisticated and even apologetic statement, "I am sorry I can not make the cues explicit, but I think that this

patient is…" I am sure that most of us will admit that even when we say this sort of thing confidently, perhaps on the basis of having checked up on our guesses in the past, it is a somewhat unsatisfactory state of affairs. It would be desirable, not only from the standpoint of teaching clinical psychology, but on the basis of the general advantages of making everything explicit, to be able to verbalize the basis of one's intuitive responses. Our research ought to be directed to the making explicit of such cues by devices of slow-motion photography, the application of group judgments, the graphical and quantitative study of gesture and verbal patterns of patients correctly versus incorrectly identified intuitively, and the like. However, it is easy to make too much of a mystery of this business, whether one is contented with it or antagonistic.

I think that one of the difficulties lies in the implicit assumption that one ought "naturally" to be able to verbalize the basis of his response, and that the cases of inability to do so are rare and constitute some type of paradox. It seems to me that this is a mistake. There is no theoretical reason why the organism in responding appropriately ought automatically to be able to emit the verbalization which *characterizes* the physical situation constituting the stimulus basis. I do not mean here to distinguish between the clinician's verbal and nonverbal behavior, which is a common division in such discussion; actually the intuitive response itself is generally verbal in nature, i.e., a diagnostic or prognostic remark. The point is that some movement on the part of the patient may have become a discriminative stimulus for a certain predictive response, i.e., "this patient will lose his amiability when you begin to get into his psychological problem," and there is no principle of learning with which I am familiar that implies high strength for a verbal response such as "I make this prediction because he showed such and such a movement." The verbal responses which are themselves *descriptive* of the stimulus field are learned over and above other responses (including verbal responses) not so descriptive but appropriate in some sense or other and hence reinforced. The same is true of many responses not involving personal interaction.

It is a truism that there is a great difference between being able to tell somebody how to do something and doing it oneself. Failure to recognize this leads to a feeling that there is something unique or peculiar about clinical intuition which requires special assumptions and explanations in order to avoid sounding mystical. Thus, we hear talk of the clinician responding to the "subliminal" or "minimal" cues. I confess I find it difficult to imagine very much clinical response based upon cues which are subliminal, and I think such assumptions are quite unnecessary. How "minimal" the cues are is a matter for experimental study, but I see no reason for assuming that they are any more minimal than most of the cues which we respond to in ordinary life. When one tries to analyze his own clinical intuition, and succeeds in making explicit the basis of such responses, it frequently turns out to be nothing more than a matter of paying sufficient attention to a kind of behavior on the side of the patient which is quite gross in extent and intensity, and would not be entitled to the term "minimal" in the ordinary perceptual sense (**11**, **53**). It would be surprising if such an important set of discriminative stimuli as the expressions, gestures, inflections, and postures of other human organisms did not become very finely discriminated in their control over our behavior. But it would be equally surprising if, in the absence of explicit formal instruction of the Dale Carnegie or successful salesman type, there should be set up (in addition) a set of verbal responses *descriptive of the cue basis.* Once this is seen, there ceases to be anything special or paradoxical about the obvious fact of this sort of clinical intuition, and we have nothing to argue about except questions which are settled by specific experiments. What sensory modalities are most important, what individual differences exist among clinicians, what are the personality or historical correlates of such individual differences, what kinds of intuitive predictions are likely to have the highest validity—these are among the many questions to be investigated in detail in the experimental study of this process.

A second, although less common, use of the phrase "clinical intuition" does not involve any reference to the verbalizability of stimuli coming

from the patient, but simply confesses an inability to show in what manner a particular hypothesis was *arrived at* from the stated evidence. In this also I see nothing mysterious or paradoxical. What we seem to be asking for here is a sort of rule or recipe for the creative act of hypothesis-formation; and when we cannot formulate one but find hypotheses presenting themselves to our consciousness nevertheless, we feel somewhat disturbed. Once having conceived a hypothesis about the patient, we are not often troubled by any difficulty in showing how this hypothesis is related to certain facts. It is true that in explicating this relationship we make use of inadequately confirmed general principles, but this does not introduce anything different in principle from what occurs in the physical sciences or in the hypothesizing of ordinary life. Let me illustrate by a concrete example.

A patient tells a dream which begins as follows: "I was in the basement of my parents' house, back home. It seems that I was ironing, and a fellow whom I had not seen since junior high school, and whom I never went out with, and hardly knew, had brought some shirts over for me to iron for him. I felt vaguely resentful about this—oh, and by the way, he was dressed in a riding habit, of all things" (grinning). Now, this patient had said in the preceding interview that it would be too easy to get into the *habit* of having sexual relationships with her present boy friend, and that since she did not really care a great deal about him, she must try to avoid this. If the phrase "riding habit" is a sexual pun, we infer that the adolescent acquaintance whom she "hardly knew" represents her present friend in the dream. The remainder of the dream and her associations to it, which I will not reproduce here, confirmed this hypothesis.

Such moment-to-moment "predictions" during the course of an interview are made by all clinicians who use any sort of interpretive therapy. Of course, we know little about their success frequency, or the reliance which ought to be placed upon them in directing the interview's course. But the validity and utility of such prediction is not the point here. The

important thing is that a description of someone's clothing in a dream would only rarely constitute a pun, and that the punlike character in the present instance becomes apparent only when we have in mind the particular situation in the patient's sexual life *and her way* of *speaking about it*, from the previous interview. Since one cannot keep constantly in mind everything the patient ever said, what is required is that the verbal stimulus "riding habit" in close temporal contiguity to verbalizations of a vaguely resentful sort, and presumably the vague awareness on the part of the listener that the identity of the old acquaintance is something needing to be clarified, combine to produce an association to the phrase. As Reik has emphasized, it would be difficult to write a prescription telling anyone how to "have such associations." In the context of discovery it is not clear at all in what way the selection of preconscious material from previous interviews combines with what is now being heard to yield the association. It is our lack of information about the dynamics of this process, and what is perhaps a permanent inability to formalize the logical steps in it, which lead us to describe it as intuitive. Nevertheless, once having thought of it, we can rationally support it as a hypothesis, although, of course, at the moment of its conception it is confirmed in a very low degree. It is not impossible or particularly difficult to make explicit the way in which the facts in question can support such a hypothesis once we have thought it up. One makes use of such general principles as these: in dreams abstract notions are often represented by concrete forms and processes and in a minority of cases the mode of plastic representation involves a pun. The unusual choice of apparel in the situation of the manifest content requires explanation, as does its insertion at the place and in the manner described. *If* the patient is hostile and resentful toward her present friend, and particularly so because or an unwelcome feeling of sexual dependence upon him, but if (as is true of her in general) the expression of hostility is more difficult than the expression of eroticism, in order for the resentment to appear we may expect both its personal object and situational origin to be distorted.

Add to this the necessity of plastic representation for such an abstract notion as habit, and we have the present result.

Sarason (**84**), in an excellent article on the interpretation of the TAT, has discussed the question of intuitive inferences in the case of this instrument. On the whole, I think his treatment is admirable, but it seems to me one might still carry from it the implication that ultimately clinical procedure will be irrational unless the steps of hypothesis formation are explicated. Sarason does not actually say this, but the general tenor of his treatment might imply it to some readers. It is a mistake to equate rational predictions to mathematical-mechanical predictions, which makes, for example, scientific crime detection irrational because it does not proceed explicitly actuarially. So it seems to me it is dangerous to require that in the process of hypothesis creation, i.e., *in the context of discovery*, a set of rules or principles (recipes, for example) is a necessary condition for rationality. What should be required is that a hypothesis, once formulated, should be related to the facts in an explicit although perhaps very probabilistic way. But to *come* to the hypothesis may require special psychological dispositions on the part of the clinician which are only acquired by experience superimposed upon what may or may not be a fundamental personal talent. The teachability of such a general hypothesis-forming disposition is, of course, an important problem which has hardly been investigated at all.

Let me conclude these speculations and emphasize their outcome with an examination of Sarbin's paper, "Clinical Psychology—Art or Science?" (**85**). It must be obvious by now that I am sympathetic to Sarbin's point of view, in that I should like to see clinical psychology become as scientific as possible and am impatient with those who appear to revel in its irrational components. There are a few clinicians who pay lip-service to the *future* scientific status of clinical work, add sadly that "unfortunately" at the present time it is not in this Utopian condition, and then show by most of their off-guard behavior that if that Utopia should miraculously be brought about in our generation they would probably abandon the field

and pursue other interests more in harmony with their motivational structure. But for clinical psychologists who, in spite of possessing and respecting clinical know-how, nevertheless are genuinely committed to making the enterprise as scientific as its subject matter permits, it is important not to become impatient with the scientific because its more passionate proponents make mistakes. It is in the hope of avoiding this consequence that I am spending so much time upon a detailed analysis of the Sarbin-Lundberg position.

Sarbin says:

> The present author agrees with Lundberg in that useful diagnoses always proceed from generalizations, whether based on a rigorous statistical method or upon a crude empirical method which has been variously named intuition, insight, *verstehen*, etc. When a clinician is put to the test to defend a diagnosis, he may resort to the statement that it was "the general feel of things" in the interview that influenced him. By pushing him back, however, it is possible usually to discover the empirical basis for the diagnosis. That these inferences are informal and not made with the benefit of Hollerith cards and Monroe calculators is beside the point. They are drawn from the clinician's cumulative experience. If they are not, then the diagnostic function must be relegated to individuals with some sort of magical power. "Thus the only possible question as to the relative value of the case (or clinical) method resolves itself into a question as to whether the classification of, and generalization from, the data shall be carried on by the informal, qualitative, and subjective method…or the systematic, quantitative, and objective procedure of the statistical method" [citing Lundberg].
>
> At this point the critic will hold up his hand and bid us go no further: all that you say is true, he tells us, if you accept the postulate that clinical psychology is a science.…The clinical psychologist uses those scientific findings and techniques which are applicable to his clinical problems.…Then even while he is developing such a complete personality study, he engages upon the *genuinely artistic task* of helping the patient to solve his own problem. [italics added] …
>
> This expression, genuinely artistic task—without further definition—leads us into a morass. The possible meanings for the words art and artistic as used here are: (a) skill in the use of tools; (b) individual explorations into the unknown; (c) possession of a unique talent or gift; (d) so-called intuitive operations.

(a) If art means the skillful use of tools, then we must ask, whence come these skills? It is unnecessary to elaborate on the point that skills are acquired from experience with tools. For example, if a clinician can make ingenious predictions of social adjustment from the perusal of certain psychological tests, he would be demonstrating his skill. Such predictions are obviously made against a background of previous experience with psychological tests and social behavior. With this conception, the writer has no quarrel. It does not postulate a super-empirical method of understanding. It is not, therefore, a material departure from the proposition that clinical psychology is scientific in that predictions are made on the basis of empirical data.

(b) If art means individual explorations into the unknown, we have no way of checking on the validity of predictions formulated in the name of art. If a clinician should make a diagnosis and prescribe treatment for a case that was unique, idiosyncratic, in every conceivable way, he would be venturing into the unknown. He would be guessing. This would be an expression of personal taste. If the clinician had no experiential background, no knowledge of similar cases, then he would be making a truly individual prediction. Unless such a single prediction is ordered to a class of events, it cannot be verified and is, therefore, meaningless.

(c) If, in this context, art means the possession of a gift or talent for "making friends and influencing people," then we can look for little progress in the field of clinical psychology. If clinical psychology is an art because some clinicians possess unique traits, and if complex human problems can be solved only by these specially-gifted people, then we must agree with Rogers and "admit that we can never deal in any large way with the multitude of ills which we group together as conduct problems, since the talents of the artist can be little conveyed to his fellows." Recognition of this problem is also given in one of the most provocative books to be published recently on personnel administration. Roethlisberger and Dickson make this generalization on the basis of the outcome of a thoroughgoing research program in personnel administration:

"The skill (of diagnosing human situations) should be 'explicit' because the implicit or intuitive skills in handling human problems which successful administrators…possess are not capable of being communicated and transmitted. They are the peculiar property of the person who exercises them; they leave when the executive leaves the organization. An 'explicit' skill, on the other

hand, is capable of being refined and taught and communicated to others."…

In this connection, it should be pointed out that the so-called art of interviewing, long considered an implicit or intuitive skill, has recently been studied, refined, and communicated to others. Porter…and Bordin and Sarbin…have studies in progress which show how these so-called artistic skills may be taught and learned.

(d) If art means some super-empirical method of understanding, then we must surrender our ideas about communicating techniques and procedures in clinical psychology. If we depart from the method of logical inference, i.e., the scientific method, then we must perforce adopt some so-called intuitive approach. Not inductive, not based on logical inference, the intuitive method of understanding is described by Klein as follows:

"…(it is) the task of fathoming human motives or appreciating the entire gamut of human desires…(it) requires a knowledge of human nature. It represents the type of understanding indispensable for the development of psychology as a social science or as a Geisteswissenschaft."…

The traditional methods of science, he points out, have a place in psychology, but the intuitive approach, characterized by the quotation above, is to reap the harvest in psychology. (**85**, pp. 395-97.)

Let us consider Sarbin's fourfold classification of "genuinely artistic tasks" in the light of our previous discussion. Meaning (a), the skillful use of tools, does not produce any disagreement from Sarbin so there is little to say about it. It is, however, necessary to be aware of the fact that when Sarbin says "such predictions are obviously made against a background of previous experience," he is not proving that the prediction is actuarial in the *narrower* sense, nor does it follow from his general statement that if the clinician's prediction were made explicitly actuarial (that is, formalized in an actuarial table or a multiple regression equation) he would do a better job than he does at present. So long as the distinction between two claims—"the clinician's skill is based upon experience," and "the clinician is a second-rate substitute for a Hollerith machine"—is maintained, there seems to be no basis for disagreement in Sarbin's treatment of (a).

"(b) If art means individual explorations into the unknown, we have no way of checking on the validity of the predictions formulated in the name of art." If read literally, this assertion is simply incorrect. As I have tried to indicate in the discussion of Jones' suicide, the validity of an artistically arrived at idiographic prediction is checked in the same way that any other prediction is checked—by waiting around to see whether or not it occurs. It is somewhat surprising that Sarbin should make this mistake, since his thought is obviously heavily influenced by the work of Reichenbach, who discusses at some length the problem of prediction in a disorderly world in which there appear to be no stable relative frequencies but in which a certain clairvoyant is able to anticipate the future. The primacy of his general inductive principle is established by making clear that even in such a world there is *one* relative frequency which is not completely unlawful, i.e., the class of the clairvoyant's predictions. In such a world we would predict our futures by making use of the clairvoyant, but it is foolish to do this until we have established that such a procedure "pays off," and this means an application of the fundamental rule of induction to the clairvoyant himself.

I do not mean to suggest that the clinician is *not* behaving on the basis of previously established frequencies and complex ways of combining them, but the point here is that even were there such a thing as a clinical clairvoyant, involving a genuinely extra-mundane or supra-empirical basis of arriving at clinical predictions, the predictions of such a clinician could be confirmed or disconfirmed in the usual way, and it could also be decided what degree of confidence we should have in his predictions. Furthermore, it could be determined whether subclasses of the set of all his clinical predictions differed significantly with respect to their relative success frequencies. Such a finding would lead us to place greater faith in him when he is predicting certain kinds of events than others. It might even be shown that although his predictions do not appear to be *based* upon any of the facts available to him, so that sometimes he predicts success in the presence of an alcoholic foster father and at other times,

everything else apparently being equal, he predicts failure under similar circumstances—it might nevertheless appear that, *whatever* he predicts for the subclass of cases involving an alcoholic foster father, his success frequency is extremely high. We cannot agree with Sarbin's statement that "unless such a single prediction is ordered to a class of events, it cannot be verified and is, therefore, meaningless" (**85**, p. 396). Nor can we agree that "If a clinician should make a diagnosis and prescribe treatment for a case that was unique, idiosyncratic, in every conceivable way, he would be venturing into the unknown. He would be guessing. This would be an expression of personal taste" (**85**, pp. 395-96). I do not suppose that any clinician imagines that he deals with patients who are *completely* unique and idiosyncratic, if by "completely" is meant that there are no similarities! But even if there were such a clinician, Sarbin would not be entitled to equate such a wholly idiographic procedure to "guesswork" or "taste." Sarbin's mistake here consists in equating the nondeductive or nonformal with the *irrational.*

Suppose a clinician should come upon a fantastic organism which, although behaving lawfully, did not behave in accordance with any of the psychological laws of organisms in the clinician's experience. Given a considerable mass of material, still actuarial in Lundberg's sense of involving repeated episodes in its life history, the clinician might be led to the construction of a "theory" about this individual organism. This "theory" would be defended by the clinician on the grounds of its capacity to entail the known facts about the individual's previous and present behavior, and would be capable of entailing certain predictions about the future. If we leave open the question as to whether the prediction will be right (which will depend on whether the theory thus constructed is correct), the important point is that the clinical activity involved is in no sense of the word unrational. It is unique and idiographic; it is nonactuarial, except in the trivial sense of the word which equates it to inductive; and even the inductions do not apply to any organism except the present one. The source of Sarbin's difficulty here is his belief that to deal with *novelty* we

must either show that the novelty is merely apparent, or else we must have recourse to nonrational methods.

I am sure that Sarbin does not feel this to be the case in scientific theories of a *general* sort, where from time to time in the history of science, as for example in the electromagnetic theory, whole aspects of the world began to be investigated which were genuinely novel. It is true that the *symptoms* of the magnetic field involved events which were describable in terms of mechanics, e.g., the deflection of a needle. But the laws and constructs of electromagnetic theory were of a different sort. Modern atomic physics has had occasion to introduce many objects and events which bear only the crudest analogical relationship to anything seen in macroscopic experience, and in many cases physicists have had to endow certain subatomic events and processes with characteristics that do considerable violence to our ordinary conceptions. No one doubts seriously the capacity of the human intelligence to make sense out of a fairly complex set of observations, even when the processes and laws involved are new. But here again what is involved is a capacity to *invent* such theories, and we have moved outside the province of formal logic. In the formal disciplines, the logician can tell us almost all we have to know about how to make inferences, and can make clear their logical structure. In the empirical field, he is barely beginning to reconstruct the basis of confirmation of hypotheses, using as a model even the simplest kind of world; and there is at present no hint that he will ever be able to tell us how to make up the sentences which are confirmed by certain evidence. In one restricted sense of the word, it must be admitted that *all* empirical hypothesis-making is nonrational, in the sense that explicit instructions for *creating* hypotheses cannot be stated. But this is surely not a use of the word "nonrational" which is important here.

"(c) If, in this context, art means the possession of a gift or talent for 'making friends and influencing people,' then we can look for little progress in the field of clinical psychology." I am at a loss to understand how Sarbin arrived at this statement. If high-level operations in clinical

psychology depend upon certain special human traits, then progress in clinical psychology will obviously be furthered by the use of suitable methods of selection for those trails. If we were to be forced to the conclusion that it is impossible for the actual day-by-day operations of the clinical psychologist to become explicitly scientific, we could still, as scientific personnel men, set up procedures for the selection of students on the basis of these talents. To carry the argument further, it might be discovered that only "skilled, intuitive clinicians" could detect the characteristics of "skilled, intuitive clinicians." Even this would not discourage us with respect to improving the status of clinical psychology, since these clinicians-for-selecting-clinicians can themselves be investigated as we investigate Reichenbach's clairvoyant. Somewhere along the line, in terms of *some* kind of rating, outcome, or mixture of human judgments, we will arrive at a place where everybody, intuitionists and statisticians alike, will agree we have to lay our cards on the table. Only practical difficulties, but nothing in principle, should lead to Sarbin's pessimistic conclusion from his premise. The *advantages* of being able to make certain skills explicit from the standpoint of teachability, and the other desirable consequences of having our knowledge communicable, are, of course, not to be denied.

Probably there will always be aspects of an individual's behavior which are relatively unteachable and which contribute materially to his clinical functioning. Certain talents for rapport-getting probably depend in part upon characteristics of features, size, build, voice, gesture, and choice of words, facial expression, and the like. Unless Sarbin believes in the infinite plasticity of adult human organisms, he should allow the possibility that there are combinations of personal traits in a would-be clinician which would render him inept at some kinds of clinical activity. Even if we understand all there is to know about the dynamics of the patient, we may fail unless *we* can present the right stimulus pattern *to him* at the right moment. Our ability to do this latter depends not merely upon our understanding, however complete, but also on other aspects of our

nature. Much clinical work involves *activity* in addition to *comprehension.* I may decide (even actuarially!) that the patient needs a dominant, even stern reaction from me, at this moment. Can I exhibit one?

The matter of timing is also important in this connection. Suppose, for example, that statistical studies of a factor analytic type (P technique) should show that a certain way of speaking, and even a particular choice of words, is associated with a patient's hostility to a sister. Hostility to sisters is something seen in many patients, but the tie-up between that nomothetic characteristic and the special peculiarities of this patient's language is idiographic, having arisen on the basis of a unique set of unusual experiences. In order to *prove* that this association exists, it may be necessary to carry out a very long and complex kind of statistical analysis on the verbal protocols of the individual case. There is no alternative to this, in *the context of justification;* that is to say, when the clinician says "every time he talks this way, no matter what the content of his conversation is, I know that I hear unconscious material dealing with his hostile attitudes toward his sister," he has to prove it somehow. In the therapeutic handling of the case, it is impossible for the clinician to get up in the middle of an interview, saying to the patient, "Leave yourself in suspended animation for 48 hours. Before I respond to your last remark, it is necessary for me to do some work on my calculating machine." And I do not think that this absurd illustration arises only because of our limitations of knowledge.

If I may be permitted another analogy, consider the case of the skilled baseball player. There is not much concerning the mathematical ballistics of the baseball, or the physiological and mechanical principles of locomotion, which are not understood sufficiently for all practical purposes. But no physicist, physiologist, or psychologist would argue that the writing of the differential equation of the baseball's path, and the analysis of the movements of the player's body in terms of metabolic activity producing energy to work on a complex system of third-class levers, would enable him to be at the right place, at the right time, in the right position, to perform the fielder's function.

What I am saying is that even in the Utopian stage of clinical psychology, when we have sufficient methods of selecting clinicians and have made explicit all that can be made explicit about the psychological principles we use, at the *moment* of *action* in the clinical interview the appropriateness of the behavior will depend in part upon things which are learnable only by a multiplicity of concrete experiences and not by formal didactic exposition. If Sarbin means to include this multiplication of concrete experiences under the heading of "teaching," we have no quarrel with him. But that the existence of certain kinds of behavior and discrimination are the results of such an accumulation of experiences is precisely what most of us have in mind when we refer to the artistry of the individual who is clinically skilled. Whether or not there are even biologically given individual differences of certain kinds of potentialities for clinical observation and operation we do not know. In the absence of any statistical or experimental evidence on this point, I can only say that I am appalled by the ability of some students to spend a couple of years in contact with clinical material and with constant opportunity for interchange with skilled clinicians and to retain an incredible blindness for all those clinical signs which they have not been specifically told to look for. Not only does the meaning of a behavior datum depend in most instances upon a half-formulated hypothesis concerning the case at hand, as indicated above in our discussion of the clerical worker; but all of this is impossible from the beginning unless the practitioner *notices* the behavior in question. Individual differences in such sensitivity, the source of these differences in heredity or in very early interpersonal learnings, its modifiability as a result of normal life or practicum training and the like—all are experimental questions which neither Dr. Sarbin nor I am in a position to prejudge.

With regard to (d) from Sarbin's article, what has just been said is probably sufficient. I am sure my remarks will not be interpreted to mean that I anticipate or desire that the intuitive approach will "reap the harvest" in psychology.

8

Empirical Comparisons of Clinical and Actuarial Prediction

FOR some reason the literature contains almost no carefully executed studies of the clinical-actuarial issue. Although a number of psychologists, psychiatrists, and sociologists have discussed this problem, empirical evidence concerning the relative efficacy of the two methods of prediction is largely wanting. I have been struck by the fact that both statisticians and clinicians often seem to think the answer is "obvious," the trouble being that they don't agree on what it is!

Allport (**5**) cites what he considers to be evidence for the superiority of the case-study method: "Studies should be made of the relative success of actuarial and case study predictions. If sensitive judges employing adequate documents commonly excel in their forecasting, we shall know that actuarial predictions are not the apex of scientific possibility, and shall conclude that the prevailing empirical theory is too meager to apply to the optimum level of prediction." (**5**, p. 160.) Allport adds in a footnote: "Already there seems to be considerable evidence that case study prediction excels. The experiments of Estes, F. H. Allport and N. Frederiksen, and Polansky are all relevant. To be sure these experiments are limited in scope; but they can be, and should be, extended." (**5**, p. 160.)

The three empirical investigations here cited by Allport are interesting

and have a tangential connection with the present issue. But I do not believe that these studies contribute as much to a solution of the empirical problem of clinical and actuarial prediction as Allport thinks. My analysis leads me to think that these three studies are, in fact, largely irrelevant. Any empirical study of actuarial versus nonactuarial predictive techniques should involve the making of predictions from similar or identical sets of information by the two methods, and a comparison of the success frequency arrived at in these two ways. Obviously, any investigation which does not anywhere involve the making of predictions upon an actuarial basis cannot make such an empirical comparison of predictive efficiency. None of the three studies Allport cites involve the making of predictions of an actuarial type. Hence, they can have, at most, a feebly supportive role with respect to Allport's major contention.

Let us begin by a brief consideration of Polansky's study (**80**). The essential point of Polansky's investigation was a comparison of the success frequency of predictions made by a group of judges on the basis of case histories, each of which had been written in six different ways. These six modes of writing a case history are called by Polansky *structural analysis, cultural presentation, genetic presentation, major maladjustment,* the presentation of *typical episodes,* and *individual differences* (*psychometric*). The life histories were those of three subjects, and for each subject all six types of life histories were written. Each of 36 judges made predictions twice for each of the three subjects, once by each of two methods. The judges were asked to predict, using a 12-item, 5-choice questionnaire, 12 factual items about the subject, of which the experimenter had knowledge. For example, the judges were to postdict what the subject did when he was broke, what words the subject would choose as unpleasant, how many times the subject had had sexual intercourse, his hobbies, his dress, his views on Marxism, his religious beliefs, his vocational choice among a group of stated alternatives, and so forth. The three subjects were "three friends of the experimenter…similar in age, sex, and basic cultural background." Polansky's control investigation of "cultural chance prob-

ability" utilizing Harvard College students in combination with the above brief characterization of the subjects indicates that these three subjects were Harvard College men. This homogeneity will be important in our discussion of the study.

Polansky analyzes his data in several ways, but the most important question is the relative predictive power of the six types of life histories. Analysis of variance of the percentage of "hits" made by the use of the six modes establishes that there is a significant difference among them. The most predictively effective mode is the "structural analysis," which is the kind of description of a person favored by Allport and his school. The least efficient method of presenting a life history in terms of predictive success was the mode called "major maladjustment," which is the type of personality description Allport considers typical of the usual psychiatric report. There were marked and consistent differences in the subjective responses of the judges in their willingness to make predictions on the sets of data, in their feeling of understanding, acquaintance with the subject, and the like. For purposes of the present discussion, our concern is the contrast between the most efficient mode (Allport's "structural analysis") and the mode which turned out to be next to the bottom in predictive efficiency, namely "individual differences" (psychometric mode). Of the total number of predictions made from the structural analysis mode, 47.6 per cent were objectively correct; whereas of the total number of predictions made by the psychometric mode, only 36.9 per cent were objectively correct. This difference in percentage of hits is statistically significant at the 1 per cent level. I assume that it is this comparison which Allport considers evidence on the subject of clinical and actuarial methods of prediction. I shall make several critical comments on his interpretation.

In the first place, it might be suggested that Polansky loaded the dice somewhat against the psychometric prediction by his choice of measuring instruments. The administration of the Wechsler-Bellevue, the Nebraska Inventory, and the Bernreuter Inventory to three subjects such

as Polansky's could almost be considered a waste of psychometric time. It is hardly conceivable, for instance, that three Harvard students would be sufficiently discriminated as regards intellect by a test with so little top as the Wechsler to make the obtained IQ's of any predictive significance. The evidence for validity in the case of the Nebraska Inventory, the Bernreuter, and the Pressy X-O Test is hardly impressive enough to warrant us in expecting that much of anything could be predicted from these three devices. The only two tests of the battery which I should be interested in knowing about when attempting to make predictions of the sort required in this study would be the Lentz Opinionnaire, which is relevant to only one of the twelve questions (subject's attitude to Marxism); and the Allport-Vernon Study of Values. It should be pointed out that the available psychometric devices which would be relevant to the predictions required are very limited in number and the validity of many potentially useful instruments is not definitively established. It is probable that Polansky's judges would have done somewhat better with the Psychometric mode had they been using information gleaned from such psychometric devices as the Rorschach, the Strong Vocational Interest Blank, the Kuder Preference Record, the MMPI, and, assuming that a capacity test would be fruitful in the battery at all, a measure of general intelligence more suited to the selection of subjects, e.g., a test of graduate ability such as the Miller Analogies. It should also be mentioned that there is no good reason for preventing judges in such a predictive situation from having access to the actual item responses made by the subject. The fact that there would be a specific content overlap between such responses and the facts to be predicted should not argue against such a procedure. It is evident that the same objection could be made to the predictions to the other five modes. In the case of the structural analysis, inspection of the sample history indicates that, in some cases at least, the judges have to do no more than remember the facts directly given to them in the case presentation. For instance, one of the predictions to be made involves the

question as to whether the subject has had sexual intercourse. If part of the structural analysis involves a statement concerning his virginity there is no good reason why the judge should not make use of this information. But by the same token, in attempting to predict the subject's attitude toward Marxism, it is certainly legitimate for the judge to know that when asked specifically in a verbal questionnaire whether he thought highly of Marxism the subject stated flatly that he was against it.

I do not wish to minimize in the least the terrible difficulties encountered in attempting to make concrete behavior predictions from psychometric data, even at their best. As a clinician I am fully aware of the peculiar feeling of "abstraction" which one gets in attempting to characterize a person from a set of test scores. I do not suppose that most practicing clinicians would sacrifice an hour of direct contact in an interview for any set of psychometric scores, if compelled to choose, although there are many individual cases in which the tests get at things which we do not get at in the interview. My aim here is to emphasize that Polansky's battery would not represent the power of psychometric procedures at their best. In the light of these considerations it is important to notice the fact that there was actually only about a 10 per cent difference in predictive efficiency between the psychometric mode and the structural analysis mode. It is really rather surprising that the judges were able to do as well with the psychometric data as they did! However, the crucial point is that the Polansky study *does not involve any empirical comparison of the actuarial and nonactuarial methods of combining data for predictive purposes.* Actually, all the predictions were made clinically; that is, the judges combined the information received in whatever manner seemed subjectively most appropriate, in the absence of any exact knowledge concerning the statistical relationships between this information and the to-be-predicted behavior, and with only the most scanty evidence as to the probable behavior correlates of the independent variables. In terms of the distinctions I have made previously, the Polansky study involves a comparison among *kinds of data,* not among *modes of combining data.* An

actuarial prediction, in the sense at stake in this argument, would involve the use of tables showing, for example, the distribution of intercourse frequencies in cells defined by certain complex conjunctions of the data. The table (based on empirical study of a suitable population such as male Harvard undergraduates) would contain a cell for "cases having Bernreuter B1-N scores between 30 and 50, value profiles with economic score as peak and religious low, no siblings, etc." The frequency distribution of sexual experiences within this cell would make the judge's prediction a clerical task. Note that here, as usual, the *kind-of-data* dimension cuts across the *actuarial-judgmental* dimension. Nothing remotely resembling such a procedure enters into Polansky's design. Therefore, the relation of this study to the main issue is tenuous at best.

The study by Estes (**40**) concerns the judgments of personalities from expressive behavior. The subjects were fifteen of the cases studied intensively by Murray in the *Explorations*, and the behavior sample available for judgments was moving pictures of the subjects carrying out simple tasks, such as lighting a match, putting on a coat, building a house of cards, and Indian wrestling. These findings are of considerable interest to the clinical psychologist, but again, we do not have any comparison of clinical and actuarial methods of prediction. Here it is even more obvious than in the Polansky study that *all* the predictions were made nonactuarially. Those judges who had been specifically trained in scientific and analytic methods of arriving at their opinions, e.g., experimental psychologists, students of the physical sciences and philosophy, were reliably inferior at such predictions to persons in the fields of fine arts, dramatics, and so forth. In other words, when there is no actuarial basis available to any of the judges, they have to be impressionistic; and the best "impressionists" are those who spend their time doing this kind of thing!

It is likely that we have here to deal with that sort of instantaneous Gestalt-like synthesizing operation at which the central nervous system of trained clinicians is presumably adept. If it is true that training in analytic

thinking and logical reconstruction of evidence reduces such skill, it might be desirable to avoid clinical training procedures which tend to produce such results. The all-important problem of the "clinician as instrument" is being discussed these days, and we psychologists should learn from investigations of the Estes type. But the present experimental design simply does not involve a comparison between clinical and actuarial modes of combining data, and consequently is not suited to Allport's purposes.

The study of F. H. Allport and Frederiksen (**3**) makes use of the method of correct matching. A certain dilemma involving moral decision is presented to a group of subjects, who are instructed to write a paragraph predicting the response of 5 of their friends. These 5 friends are independently presented with the same dilemma, and the problem for the judges is to match the actual responses as written by the friends with the predicted responses as written by the other subjects. Of the total of 1530 single matchings made, the investigators found 24.9 per cent correct matchings as contrasted with a chance value of 20 per cent. Because of the large N this result is statistically significant. However, when one considers that the actual results are only 4.9 per cent better than chance, it is difficult to see how any particular importance can be attached to the results. The statistical significance is obtained not because of the high degree of accuracy of the judgments but because of the very large number of responses that go into the significance test. In fact, the stability of the observed low percentage with such an N allows us to state that this sort of matching can be done very poorly! Again, no direct comparison of actuarial and nonactuarial methods of prediction is involved. For purposes of Allport's argument, this study is also more or less tangential.

The three studies cited by Allport in his footnote do not constitute evidence for the superior efficiency of nonactuarial prediction. I have managed to find twenty studies which *do* involve an empirical compareison of the two techniques, and which can perhaps shed a little light upon the problem. The ideal design is one in which the same basic set of facts is

subjected on the one hand to the skilled analysis of a trained clinician, and on the other hand is subjected to mechanical operations (table entry, multiplication by weights, or the like). The predictions arrived at by these two methods are then compared with respect to their success. In the following investigations this design is approximated to varying degrees. I do not claim that the following is a complete review of literature, but it represents everything I could locate by entering the *Psychological Abstracts* via an extended and diverse list of topic names, plus inquiry among psychologists I knew to be interested in the problem.

The first systematic investigation aimed deliberately at getting an empirical answer to our question was carried out by Sarbin (**86**). His study has not received anywhere near the attention it deserves. It was designed from the start to compare the two methods, whereas in most of the other relevant studies that comparison was incidental to some other major research aim. I have been repeatedly amazed to hear clinical workers make flat statements about the answer to Sarbin's question, only to find that they had never heard of the study.

Sarbin chose as his criterion variable academic success as measured by honor-point ratio. The sample consisted of 162 freshmen (73 men and 89 women) who matriculated in the fall of 1939 in the arts college at the University of Minnesota. Honor-point ratios were calculated at the end of the first quarter of the students' freshman year. The statistical prediction was made by a clerk who simply inserted the values of the predictor variables into a two-variable regression equation. The predictor variables were high school percentile rank and score on the college aptitude test. (Note: One psychometric and one nonpsychometric variable.) The sample used was cross-validating, since the regression equation had been based upon a previous sample.

The clinical predictions were made on an eight-point scale by one of five clinical counselors in the university's Student Counseling Bureau. Four of the five counselors possessed the doctorate and all had "considerable experience" in clinical counseling of university students. The data

available to the counselors were considerably in excess of those utilized by the statistician, namely: a preliminary interviewer's notes, scores on the Strong Vocational Interest Blank, scores on a four-variable structured personality inventory, an eight-page individual record form filled out by the student, scores on several additional aptitude and achievement tests, as well as the two scores utilized by the statistician. In addition, the predicting clinician had one interview with the student prior to the beginning of fall quarter classes. At the end of the fall quarter the correlations shown in the tabulation were obtained between the two sets of predictions and the facts. There is no significant difference between the efficiency of the two methods.

	Men	*Women*
Clinical.....	.35	.69
Statistical.....	.45	.70

Even though the clinician, utilizing all this additional information, is no better at forecasting than the statistical clerk, Sarbin felt that perhaps they were hitting different cases and matters would improve if the clinician were included as a statistical variable. The increment to the multiple R given by adding his judgmental rating as a third variable in the predictive system was only .01 for men and .05 for women, neither of these improvements being significant. Some of the clinicians felt that there was no practical value in the refined eight-point prediction, and that they would do better merely being asked to predict success versus failure. Dichotomizing the continuum at an honor-point ratio of 1.00 (C average) Sarbin reanalyzed the data in these terms. For male students, the statistical and clinical methods were not significantly different in predicting this dichotomy, although there was a slight trend favoring the statistical; for women, there was borderline significance ($.01 < P < .05$) in favor of the statistical. When the data for both sexes were pooled, the statistical method was superior to the clinical at between the 1 and 2 per cent levels of confidence. The over-all magnitude of this superiority was, however,

only about 6 per cent hits (my calculation) .It was also shown that the clinicians systematically overpredicted grade average ("leniency error").

A rather surprising finding, considering the mass of additional information they bad available, was that the clinicians' predictions were significantly more correlated with the two predictor variables than the criterion variable was. That is, the clinicians overestimated the contribution of the two major predictor variables, attributing more criterion variance to them than they in fact control. As Sarbin says, "…the case-study method takes behavior segments with known weights and applies other weights which are less efficient" (**86**, p. 596) .Both methods systematically underestimate the criterion variance, although it is debatable whether this should be called "error," since if the statistical method did *not* do this the mean squared error of estimate would necessarily increase. That being so, the corresponding restriction of range by the clinician (which Sarbin calls "playing safe") is based upon a sound statistical principle. Sarbin does not present data indicating whether individual clinicians did better than the regression equation. Reliabilities for the five clinicians based on rerating after six months ranged from .64 to .88. It seems quite possible that since the clinical and statistical methods were so nearly equal when the judgments of all clinicians were pooled, one or two could well have been superior to the regression method. Wittman (**103**) developed a prognosis scale for use with schizophrenic patients, consisting of thirty variables rated on the basis of social history (and the psychiatric examination?). With the exception of marital status, all of the variables were more or less "judgmental" in character, and would involve varying requirements upon the clinical skill of the- rater. They range from semi-objective matters, such as *duration* of *psychosis*, to highly interpretive judgments, such as *anal erotic versus oral erotic*. None of the predictive variables were psychometric. Numerical weights were assigned to the values of these ratings on the basis of the "frequency…and relative importance ascribed to them in more than 50 studies by various authors" (**103**, p. 21). We may, therefore, presume the weights employed were not optimal and

hence not fair tests of the power of actuarial prognosis, since they were not determined by an actual statistical analysis of any defined sample but arose from a crude quantification of impressions found in the literature. All ratings were made by the investigator, who rated either prior to the beginning of a patient's therapy or by reading the charts (minus progress notes) of old cases. The agreement of total score between her ratings and those made by another staff member on a small sample (N = 61) was +.87, but she points out that the marked bimodality found influenced this coefficient markedly.

Independently of the scale, the psychiatric staff had made a three-step rating as to prognosis prior to beginning therapy at a "diagnostic conference." It is not clear from the report whether the final statistics cited refer to a pooled judgment or not, but it is clear that individual staff members' predictions must have been obtained for study (see below).

The criterion was a five-step rating made at a therapy staff meeting after conclusion of shock treatment. The degree of contamination of this rating by the psychiatrist's pre-treatment impression is not inferable from

Five-Step Criterion Categories	*N*	*Percentage of Hits by Scale*	*Percentage of Hits by Psychiatrists*
Remission	56	90	52
Much improved	66	86	41
Improved	51	75	36
Slightly improved	31	46	34
Unimproved	139	85	49

the report but would presumably inflate the percentage of hits for the psychiatric staff. This five-step scale was collapsed to a three-step for comparability with the staff clinical judgments, by grouping the two most favorable outcome categories as "greatly improved" and the next two as "guarded." With a "hit" defined as proper placement in this three-step scale, the results were distinctly favorable to the actuarial method, as

shown in the tabulation. The total hits by the scale were 81 per cent, as contrasted to only 44 per cent by the psychiatric staff. This difference is significant at the $P < .001$ level (my calculations). Wittman states that the individual staff members ranged from 8 per cent to 81 per cent hits, although no further details are presented. It thus appears that both the typical and the pooled staff judgment were far below the scale in predictive efficiency, and that the *best* staff member just equaled the scale.

In a second study (**104**) Wittman and Steinberg present further data based upon a larger sample of 960 patients, this time including 156 manic-depressives. The bimodality of prognosis scale scores for the schizophrenic group is again in evidence. The continuous prognosis scale was divided into three intervals as before so as to make its predictions comparable to the three-step staff ratings. Both staff judgments and prognosis scale rating were completed from eight months to three years prior to the criterion evaluation. In this follow-up study, the superiority of the scale method is still very evident but of somewhat lesser degree. Total staff hits came to 41 per cent, as contrasted to 68 per cent for the scale ($P < .001$). Both scale and staff yield highly significant chi-squares against the criterion in a nine-fold contingency table. The contingency coefficient for the scale is .61, while that for the psychiatrist is only .21 (my calculations). Wittman states that Sarbin revised and shortened the scale and that his report on the revision is contained in the same volume of the *Elgin Papers*. This is not true of the volume accessible to me, and I have been unable to locate any such research report.

Schiedt (**90**) in a doctoral dissertation in jurisprudence showed that fifteen of Burgess' factors (e.g., age, marital status, sobriety), when combined by a simple unweighted addition, were about as successful in predicting criminal recidivism in 500 Bavarian ex-prisoners during a four- to six-year follow-up period as was the judgment of a prison physician. This comparison is distinctly unfair to the actuarial method, inasmuch as the prison physician refused to predict for about one third of the cases while the success frequency for the actuarial predictions is based upon the

entire sample. An analysis of Schiedt's data shows that if the most doubtful group (i.e., cases with p near .5) is excluded from the actuarial predictions, we have left a set of cases about as numerous as those for whom the physician made a prediction. In this subset the actuarial prediction is somewhat superior. Unfortunately, we are given no information by Schiedt as to the training and skill of the clinician, who is not called a psychiatrist in Schiedt's published paper but simply a prison *Arzt*. Incidentally, it is interesting to note the great concern shown by Schiedt lest his German readers might find the actuarial technique objectionable when thus applied to a problem in the prediction of human behavior, and his struggle to satisfy himself of the genuineness of his results in the absence of any knowledge of significance tests.

Conrad and Satter (**33**) in predicting the success of naval trainees in an electrician's mate school (N = 3500) compared the predictions of interviewers of an unspecified degree of skill making use of test scores, personal history data, and an impression gained from the interview, with the predictive efficiency of a regression equation involving two objective tests (electrical knowledge and arithmetic reasoning). No cross-validation groups were studied, but the large N makes it unlikely that the coefficients would show great shrinkage in moving to new groups. The criterion was grades achieved in an electrical school to which the men were assigned. The results were slightly favorable to the actuarial method of prediction, although whether the difference would shrink to zero on cross-validation groups cannot be decided from the data presented.

Burgess (**17**) studied the outcome of 1000 cases of parole from three Illinois state prisons. Using 21 objective factors such as nature of the crime, length of sentence, nationality of father, county of indictment, size of community, type of residence, and chronological age, and combining them in unweighted fashion by simply counting the number of factors operating for or against a successful outcome of the parole, he achieved certain percentages of success in postdiction which can be compared with the percentages of two different prison psychiatrists. Again we find that

both of these clinicians employed a "doubtful" category, but Burgess' presentation makes it impossible to say how many cases were so classed. When he predicts success, each of the psychiatrists is slightly better than the statistical method (85 per cent and 80 per cent versus 76 per cent hits). When predicting failure, each psychiatrist is quite clearly inferior to the statistician (30 per cent and 51 per cent versus 69 per cent). Since these percentages are based upon a reference class of *all* cases for the statistician but upon a smaller reference class which excludes some (unknown) fraction of the "doubtfuls" for the two psychiatrists, it seems quite safe to favor the statistical method.

Dunham and Meltzer (**37**), predicting length of hospitalization of schizophrenic and manic-depressive patients, employed a weighted combination of three predictive variables (marital status, duration of psychosis, and a rating on amount of insight). In one cross-validating sample (N = 217) there was no significant difference between the success frequency of the two methods even when the clinicians left about one fourth of the cases unpredicted, and a 10 per cent difference in favor of the actuarial method if the total number of cases is taken as a base for both methods. In another cross-validating sample (N = 288) there was a 10 per cent difference in favor of the clinicians, but here their success frequency is calculated with a base of fewer than half the cases whereas the actuarial prediction is for all. If both percentages are computed on the same base there is about a 30 per cent difference in favor of the actuarial (my calculations). The data are not presented so as to make possible a separate calculation of the actuarial success-frequency with doubtful cases left unpredicted.

Lepley and Hadley made an investigation which is not generally available, and since my only familiarity with it is through several correspondents and Super's summary of it, I shall quote the latter in full:

> *The Surgeon's Classification Board* provided an opportunity for a more comprehensive clinical evaluation of cadets being considered for flying training during several months in which it was experimented with during World War II

(described in a military report by W. M. Lepley and H. D. Hadley). The board consisted of a flight surgeon and an aviation psychologist, who interviewed each cadet with stanines below the required levels for all three air crew assignments (at that time 3 for pilot and bombardier, 5 for navigator). The interviews lasted approximately eight minutes each, ranging in length from five to twenty minutes. A total of l,524 cadets were interviewed during the six months of the board's existence at this one classification center, and 285 were sent to pilot training because the board's review of the test scores and interview data led it to believe that the cadet would make a good pilot. Follow-up data were obtained for 259 of these cadets, who were test-matched with 146 cadets sent to training at a somewhat earlier date when standards were lower and without having been passed on by a board. Various analyses were made by class and time of training; in the most legitimate comparison, 68.9 per cent of the cases passed by the board failed in training, whereas 73 per cent of those with similar stanines who went automatically to training failed. The critical ratio was 0.50, showing that cadets who were clinically evaluated by a board of experts were no more likely to succeed than others who had the same stanine or psychometric index but were not clinically evaluated. Despite certain defects in the design of this real-life experiment (e.g., elimination rates were not quite the same when the two groups were in training, being slightly lower for and, therefore, favoring the board cases), Lepley and Hadley seem to have definitely put the burden of proof upon those who claim that the clinical method is superior to a comprehensive battery of objectively validated and summated tests. (**96**, pp. 545-46.)

The data most frequently mentioned by the actuarially minded are those gathered by the Army Air Force psychologists and reported in one of the AAF Research Reports (**47**). I am afraid that this research, while of the greatest intrinsic importance and interest, is largely irrelevant to the issue and for the same reason that Polansky's work is irrelevant when cited on the other side. The specific subproject commonly quoted is the one entitled "Clinical Type Procedures," treated in Chapter 24 of the report. But most of the negative findings deal with the low validity of particular *instruments*, e.g., Rorschach, and this low validity shows up when the component variables are treated statistically for predictive

purposes just as when the treatment is "global" or "clinical" (cf. p. 632, Table 24.5, and other tables in the same section). It is true that, as Dr. Super writes me, "in one sense, then, the AAF did ascertain the relative validity of clinical judgment: when clinicians used their favored techniques in their favored way, they did not do as good a job as the statisticians did when they used their favored techniques (objective tests) in their favored way (validated and weighted)." However, the comparison did involve a mixing of the question of *data* and that of *method of combining*, and hence is not strictly relevant. The subproject CE 707A, "Conference for the Interpretation of Test Scores and Occupational Background" (pp. 652-56 of the report) is misleadingly titled, since study of the original mimeographed report (**81**) and personal communications from some of the research team make it clear that the interviewers did *not* have the test *scores* available, so that the predictions were being made from other data than those which were utilized actuarially.

On the other hand, this report does contain data relevant to our main problem, although these are not the data usually referred to. Neither Lepley, predicting clinically from the scores on his Personal Audit, nor Humm doing the same from the Humm-Wadsworth profile, were able to improve upon the validity of a straight actuarial (regression) technique. Humm actually did worse, since his ratings had no validity while two of his test scores showed significant (although very low) correlations with the criterion (pp. 583-88 of the published report). It was also shown that the clinical use of the Rorschach had the same lack of validity for this criterion as did a regression combination of Rorschach variables. A similar result obtained for a small number of traits scored on the TAT. The probative significance of all these failures on the part of the clinical method is greatly reduced by the low correlations shown between the basic variables themselves and the particular criterion involved in these investigations.

The reports of Kelly and Fiske (**62**, **63**, **64**) on the clinical psychology trainee assessment study are sometimes cited as showing the superiority

or at least equality of statistical to clinical procedures. In the summary chapter of their most complete publication, these authors make the following remark: "At this point, readers are reminded of the overall findings of the project with respect to the relative accuracy of statistical and clinical predictions of future behavior; *in this situation both approaches worked equally well* [italics theirs]." (**64**, p. 199.)

I have made no effort to review this monumental investigation, careful study of which is obligatory on all clinicians. Such study makes it clear that in the quotation above the authors are not using the terms "statistical" and "clinical" in precisely the way I have proposed, so that their summary statement must be taken to refer to some mixture of the two dichotomies, *kind of information* and *mode of combination.* For this reason, I hold that those who cite the Kelly-Fiske study on the actuarial side are making the same mistake Allport makes in citing the Polansky study on the other side. Some of the most arresting findings of Kelly and Fiske, such as the insignificant contribution of either a "preliminary" or "intensive" (two-hour) interview to the validity coefficient, are really tangential to the problem as we have posed it. Such data are valuable in their sobering effect on clinical enthusiasm, and thus indirectly affect one's orientation to the whole controversy. But the predictions and ratings *before* the interview were already truly clinical in our sense, i.e., judgmental, nonclerical inferences from the documentary information. The same is true of certain other fascinating findings in the Kelly-Fiske investigation, such as the failure of pooled, post-conference judgments by skilled clinicians to improve over the original predictions based on the same data. On the other hand, the fact that these pooled judgments, based upon an integrative *conference*, were no more valid than an arithmetical combination of pre-conference ratings does bear at least obliquely upon our central question (**64**, p.177). Taking the original ratings as the raw data, we may ask what is the best way to use them? The evidence suggests that a clerk can pool them as well as a skilled staff. But even this comparison is not quite in accord with our paradigm, since the staff con-

ference presumably involves a re-examination of the individual judges' ratings in *the context of a group discussion of the data the judges had used separately.*

There are, however, some findings in the Kelly-Fiske report which can be made to bear directly upon our problem. Because of the lack of a suitable cross-validating group, the authors did not report multiple correlations (**64**, p. 157). The correlations of their criteria with individual tests (e.g., Miller Analogies, psychology key of the Strong, certain of Gough's Multiphasic keys) rather definitely suggest a multiple R at least as good as the judgment made by assessing clinicians (*pre-interview*!) from these same test and documentary data. Thus, the median validity coefficient for the prediction of a clinician based upon objective test scores *plus* a credential file (blueprint, letters of recommendation, Civil Service Form 57) was only .29, ranging from .04 to .51 for the various criteria (**64**, p. 168). We may surely assume that these would shrink if the nonpsychometric data in the credentials file had been excluded from the information presented to the clinician for judgment. Inspection of the tables on pages 158-59 of the report indicates it very likely that a suitable combination of the best test variables in a regression system could hardly fail to do better than this. Consequently, this limited portion of the Kelly-Fiske study can presumably be included, with suitable reservations, as evidence leaning toward the actuarial side.

A study by Dunlap and Wantman (**38**) has sometimes been cited in discussions as adverse to clinical methods of prediction (cf. **35**, p. 57). In my view this study is only indirectly relevant, since it again confounds the question of kind of information with that of mode of combination. Using several criteria of pilot success (pass-fail, ground school grades, time in learning, flight instructor ratings and check lists, and camera records of instrument readings during flight), Dunlap and Wantman made a comparison between the predictive efficiency of an objective, paper-pencil test battery and judgments on several sorts of variables made by interviewers, including an over-all rating on "fitness for flight training." The interviewers worked as a three-man team of psychologist, personnel man, and

aviator. A semistandardized interview was used and a check list and rating form was available to aid the interviewers in making and recording their judgments. The interviews lasted for 25 minutes. The actual interviews were preceded by a training period which included a critical discussion of two recorded practice interviews with each board. It was shown that the reliabilities of mean ratings were fairly good (Spearman-Brown estimates .81 to .87 for the nine rated variables). The study was done using interview boards at four different universities and considerable variation among teams appeared. The total sample consisted of 208 pilot trainees. Most of the validity coefficients (10 rated variables and 9 criteria) are either insignificant or too low to be practically useful.

Some of the criteria were probably too unreliable to be predictable, although others (e.g., Ohio State Flight Inventory and Pennsylvania Camera Criterion) were very satisfactory, and even the less trustworthy were shown to be predictable to some extent, both by certain of the interview ratings and by objective tests. The report indicates that the interview boards did *not* have access to the two most powerful psychometric predictors, which were scores on the Biographical Inventory and the Mechanical Comprehension tests. That is, the clinical predictions were not based on the same information as the statistical. The third paper-pencil test, a Personal History Inventory covering several areas of relatively objective facts of the subject's life history, *was* available to each interviewer and he was supposed to study the responses prior to the interview as a partial basis for guiding the questioning. Each interviewer was to make his own "subjective scoring stencil" as an aid in interpreting the written responses to this questionnaire. Strictly speaking, the most meaningful comparison for our present purposes is between the interviewers' over-all prediction (based on the subjects' responses to the personal history questionnaire as followed up in the interview questioning) and the prediction yielded by a strictly mechanical scoring of the questionnaire to yield a single score. This latter scoring was an empirical scoring based on item analysis of the records of 1427 previous trainees

against success in primary flight training. Examination of the tables with this comparison in mind is difficult because the criteria available varied among the four schools and the sample sizes vary even within the tables for a single school. Rough estimates (mine) from the maximum N's indicated suggest that there is no significant difference between the validity of the P-H score and the interviewers' estimates, although among significant correlations the preponderance favors the interviewer. In one sample there is a difference of .40 in favor of the latter (from .15 to .55, Table XIV on p. 24 of the report) which would be at about the 5 per cent level.

The authors do not concern themselves with this comparison, but stress the fact that the inclusion of the interviewers' ratings in the multiple regression system does not materially improve the multiple correlation over what is yielded by the objective tests. But as pointed out above, of the three tests considered only one was available to the clinicians when making their judgments. It is presumably worth noting that the interview was not justified as a procedure when its time and cost are considered in relation to the negligible increment it gives to predictive efficiency, even though that is not the most relevant comparison to make for our purposes. None of the multiple coefficients were cross-validative. To the extent that the study is germane to our topic, it seems to indicate approximate equality between clinical and actuarial methods of combining the same data for predictive purposes.

Bobbitt and Newman (**14**) studied the predictive efficiency of an unweighted battery of aptitude, achievement, and personality tests against a criterion of success or failure in the training program for cadets in the United States Coast Guard Academy. The criterion was uncontaminated by the test data. Two short (ten- to twenty-minute) interviews were also held by two independent interviewers (a psychologist and a psychiatrist). The interviewer had all the test data at hand and attempted to combine the scores with his interview impressions to arrive at an over-all numerical rating. The interview ratings were standardized and the standard scores were summed to yield an interview score. Although the

data were not analyzed with the present comparison in mind, detailed inspectional study of the cumulative percentages suggests that the interviewer's final clinical judgment tends to run from 2 to 7 per cent superior to the test battery at most levels. Significance tests are not given, and it is impossible to know whether this slight advantage would be lost had the test battery been weighted optimally. Rough estimates of the standard errors suggest that whether the improvement is significant would depend on the success-failure split, since with a split yielding minimum standard error the percentage difference would apparently be of borderline statistical stability. These authors also studied the efficiency of a score obtained by an (unweighted) addition of the interview *and* test results, a procedure which added another 2 to 3 per cent to the hits. It is worth mentioning that this study, which seems to give a slight edge to the clinician, utilized apparently very skilled interviewers whose judgments were of extremely high reliability (**78**). Davis (**35**, p. 57) cites the study as showing "no improvement" yielded by the clinical procedure, presumably because of the small size of the increment. The authors emphasize the homogeneity of the group in respect to some of the tested capacity variables. In subsequent work by these researchers it appears that the interview has been eliminated from the selection procedures of the academy, so apparently its contribution was not considered to be sufficient to justify the additional effort (**77**, p. 249).

Borden (**15**) studied the prediction of parole violation in 261 ex-reformatory inmates. He began with 28 factors, about 22 of which were relatively objective indicators obtainable from history material such as legal documents. Pearson correlations were computed (uncorrected for the extreme coarseness of grouping) against a five-step objective criterion of parole success based on status one year after release on parole. All the relationships were very low, the highest being only .20 (number of previous commitments). "Psychologist's prognosis," the clinical prediction, correlated .16 with the criterion, as did the diagnosis of intellectual level (four steps). The multiple R on all three of these predictors was .41, not

cross-validated. The tabular data and the partial betas indicate that the optimal combination of intellectual level and previous commitments would be more efficient than the clinical prediction by the psychologist. I have examined Borden s raw data in another way, reducing the criterion to a (more meaningful) dichotomy and locating an optimal cut for each of his three most powerful predictors. The number of previous commitments gives 62 per cent hits on the sample, as compared with 58 per cent for intellect and 58 per cent for the psychologist's prognosis. It should be noted that the psychologist did not predict for 7 per cent of the cases, a fact not pointed out by Borden. Since the other two variables correlated only –.10, it seems quite safe to conclude that in combination they would be superior to the clinical estimate. All these comparisons lose much of their meaning when it is seen that a blind guess of success on parole will succeed 58 per cent of the time, this being the base frequency for the entire sample. All things considered, this study can presumably be tallied on the actuarial side.

Hamlin (**48**) studied 501 consecutively admitted reformatory inmates using a composite criterion of adjustment within the institution, which included such items as number of times in guardhouse, shop demerits, shop and school grades, and discipline marks. The prediction problem was to estimate this composite criterion over a four- to ten-month period following admission. More than 100 items, chiefly objective or semi-objective facts routinely obtained as part of the history on all cases, were tabulated. On the basis of zero-order correlations with the criterion, a subset of 68 items was chosen to yield an adjustment prediction score. Since this score included several clinical estimates, and was not cross-validated, it is not particularly relevant for our purposes. However, the author also presents correlations of the 20 most powerful items with the criterion, these contingency coefficients ranging from .25 to .35. Significance tests of this rather small variation of the coefficients are not given. The item of direct interest to us here, "Prognosis for institutional adjustment, psychiatrist's estimate," ranks thirteenth in efficiency among the

twenty (C = .28). A similar clinical item, although not aimed directly at the criterion studied, "Prognosis for future behavior, psychiatrist's estimate," ranks second. In fairness to clinical judgment, it should be added that the most powerful predictor was "Original assignment in reformatory," which is presumably based on some kind of human judgment by an administrator, but the author does not explain it. The criterion correlation of a prediction score based on a linear combination of fifteen nonoverlapping items (*not* optimally weighted) was .55. This excludes the psychiatric prediction of institutional adjustment, but it includes the psychiatrist's prediction of a different criterion (future behavior)! Although the author does not compute a multiple R based on the eleven or twelve purely objective factors alone, inspection of the table, together with the above cited .55 figure for all fifteen, surely justifies us in saying that the actuarial prediction would be at least as efficient as any of the clinical or administrative estimates, and very probably more efficient.

Bloom and Brundage (**13**, p. 251) report on the validity of interviewers' quality classification ratings against a criterion of success in training. The study involved a sample of 37,862 naval enlisted men who were subsequently sent to naval training schools for one of nine types of specialized training. The interviewers had test scores before them during the interview, but the correlation between interviewer evaluations and success in training was actually lower than that yielded by the same test scores used alone.

Melton (**73**) studied the efficiency of fourteen counselors in forecasting the honor-point ratios earned by 543 entering arts college freshmen in their first year's work. The actuarial prediction was based upon a two-variable regression equation (ACE and high school rank) with betas derived from a previous sample. The counselors made their predictions immediately after an interview of 45 minutes to one hour duration, and had available the two regression variables plus scores on the Cooperative English Test, the Mooney Problem Check List, and a four-page personal inventory form. The counselors were graduate students in psychology or

educational psychology in their second to final year of graduate study. He found that the mean absolute error of the actuarial prediction was significantly less than that of the counselors; the counselors overestimated honor-point ratio; there were significant differences among the counselors in their average error; eleven counselors were less accurate than the regression equation, while three were more accurate, but not significantly; when a counselor predicts *knowing* the actuarial prediction, his result tends to be less accurate than the actuarial prediction itself, i.e., the addition of clinical judgment reduces predictive power (borderline significance); and, finally, if counselors who are poor predictors are allowed to use the actuarial table in making predictions, they then predict as well as the good predictors.

Barron (**8**), in a carefully executed study of test correlates of therapeutic outcome, was able to compare the efficiency of clinical and mechanical sortings of MMPI profiles. Thirty-three adult psychoneurotics received intensive outpatient psychotherapy (one hour weekly for six months) and were judged as to improvement by two independent experts other than the therapists. These criterion judgments had a reliability of .91 and on a two-category sorting yielded disagreement on only 2 of the 33 patients. Judgments were uncontaminated by any knowledge of MMPI profiles. Eight clinicians skilled in MMPI interpretation were asked to predict this outcome criterion knowing only the patients' age, sex, and MMPI curve. Total pooled hits ($N = 264 = 8 \times 33$) came to 62 per cent, and the three best clinical sorters averaged 69.7 per cent hits. While these subjective sortings are reliably above chance ($P < .01$), they are inferior to the results obtained by applying any of three a priori mechanical systems to the MMPI profiles, which yielded hit frequencies of 73 per cent and "between 75 and 80 per cent" (p. 239) .Thus, the over-all efficiency for the skilled clinicians is between 11 and 18 per cent less than mechanical combination of the same psychometric data, the best mechanical rule being superior to the over-all clinical rate at the 2 per cent level of confidence. Even allowing for the chance variation over eight clinicians, we find the top three still slightly behind the weakest mechanical rule,

although not significantly so. Similar results were obtained with the Rorschach, but the significance of this comparison is reduced by the fact that neither a mechanical use of the usually mentioned signs nor a subjective sorting by four Rorschach experts (with access to the protocols) showed any correlation with therapeutic outcome. Barron's findings should probably be classed as slightly in favor of the statistical method, but since the differences are of borderline significance and there is variation over eight clinicians and three mechanical systems, I shall lean over backwards to call it a draw.

Blenkner (**12**) studied predictive factors in family casework, and she reports some incidental data which are highly relevant to our problem. Two skilled judges evaluated the movement of casework clients by reading the entire case records from initial contact to closing. These movement judgments had a reliability of .86. Three other judges (p. 73) "who had had considerable experience in casework, supervision, and/or teaching and who were not members of the agency staff" (p. 67), after studying the initial interview data only, filled out a ten-page schedule which had been pre-tested on similar material and in the use of which they had been trained. Five factors from this initial-contact schedule were found to be significantly associated with movement in a criterion sample of 63 cases, each of whom had experienced at least two interviews (median approximately five). These five factors were (1) referral source, (2) problem area, (3) insight, (4) resistance, and (5) degree to which client was overwhelmed. It is evident that the rating on these five variables as exhibited in the records of the initial contact already involves some considerable degree of clinical judgment by the skilled case reader. Treating each factor as either favorable or adverse, and arbitrarily assigning a score of 2 to each favorable factor, a prediction score defined as the product of the five predictor values was set up (i.e., ranging from 0 to 32 by powers of 2). Such a formula can hardly be considered a best fitting statistical function, for obvious reasons; for example, a client with four favorable factors and one adverse gets the same prognosis score as a client with all five adverse!

So the study is not a fair test of the actuarial method. However, this very crude prediction index showed a point-biserial of .62 with (dichotomized) movement ratings on the derivation sample, which shrank to .52 on the cross-validation sample (N = 47). The same skilled judges were also asked to make a dichotomous clinical prediction of movement on the basis of their reading of the same initial interview data; these predictions, as made by each of the three judges, had no validity, and the judges did not agree with one another. Apparently these skilled case readers can rate relatively more specific but still fairly complex factors reliably enough so that an inefficient mathematical formula combining them can predict the criterion; whereas the *same* judges cannot combine the *same* data "impressionistically" to yield results above chance.

Hovey and Stauffacher (**59**) compared what they characterized as "intuitive" and "objective" prediction from a test. As in Barron's study, the data available to the clinical judge were identical with those used mechanically—namely, the MMPI profile, alone. On the basis of previous empirical research on a student nursing population (N = 97 plus cross-validation on N = 40) a collection of 35 personality traits (as judged by nursing supervisors) was known to be significantly associated with the various MMPI scales (considered *singly*). The task set was to predict supervisor ratings on a third sample of 47 student nurses by utilizing the MMPI profile in two ways. In the "objective" (mechanical, actuarial) approach, a trait from the list would be attributed to the subject if the subject showed a deviation (amount not indicated) on an MMPI scale which had been associated with that trait in the original study. From four to six MMPI scores, deviating high or low, were utilized in each case, depending upon the magnitude of deviations. Special rules were set up, more or less arbitrarily, to deal with instances of scales "in opposition"—e.g., if three high scores argued *for* the same trait but one low score argued *against* it, the trait would be automatically attributed; whenever the peak score exceeded all others by five T-score points, the traits correlated with it were attributed; and so on. In this manner, "present" or "absent"

predictions were made on from 9 to 24 of the 35 traits for each of the subjects. Using the other, "intuitive" approach, an experienced clinician examined the profile and decided for each of the mechanically made trait predictions whether it would be rated "present" or "absent" by the criterion supervisor. That is, the skilled judge knew which traits (among the available pool of 35) had been mechanically predicted for each profile, although he did not know in which direction. He was required to force a judgment on these subsets only, as predetermined by the mechanical method for each case. A total of 663 single-trait predictions were made by each method. An uncontaminated rating by three supervisors, each having observed each student nurse for a month, was the criterion, attributing a trait (or its absence) if two of the three supervisors checked it and the third did not check the opposite as being true. Only 328 of the 663 test predictions "could be compared" with plus and minus evaluations by the supervisors. The mechanical method yielded a hit-miss ratio of 1.7:1, as compared with 2.8:1 for the clinical method ($P < .01$).

Since this is the sole study found in which the mechanical method was significantly inferior, it deserves careful study by way of interpretation. The most obvious caveat arises in connection with the rules used in making the mechanical predictions. Without invoking some ideal mathematical function which would, *by definition,* yield the optimal trait attributions for every MMPI pattern, one may fairly ask what guarantee there is that the mechanical method used was a fairly reasonable approximation to the best fit of even a rather simple type of prediction equation? There seems to be no reason for assuming this. Furthermore, all sophisticated MMPI workers operate with profile patterns, as the clinicians in this study quite consciously did (personal communication; see also Hovey's note 3 on p. 144 of the original study, **58**). One has no way of knowing to what extent even such crude patterning methods as the Hathaway code (**51**) are approximated by the mechanical rules used. So what we have is actually a comparison of the predictive efficiency of

skilled MMPI readers with that of a *linear, nonconfigural function of non-optimal weighs.* I would argue that an empirical determination of weights is a legitimate part of the very *definition* of the actuarial method. In the case of multivariable devices, such as the MMPI, where profile form is known (or even thought) to be highly relevant, I think it fair to go farther and to say that if only a linear function is tried, and no test is made for significant interaction effects, the statistical method has still not been given its innings, even though the optimal linear weights had been fitted.

On the other hand, one must beware of any temptation to settle the issue verbalistically in favor of the statisticians. It is tautologous that the "best rule" will excel any less-than-best rule, but this nonspecific truism of decision theory does not help us in *formulating* the "best rule." As Dr. Hovey points out (in a personal communication), "…using peak and valley scores in individual profiles would be superior to using high and low scores. But even with 137 cases…there would have been too few cases of various combinations to make it worthwhile." If a many-variable prediction system is highly "configurated" (see next chapter) then determination of the function form and, a fortiori, estimation of the constants, require a very large N. For obvious reasons, shrinkage on cross-validation due to excessive capitalization on sampling errors tends to be greater as the prediction function becomes more complex (e.g., involving higher powers, cross-products, numerous constants). Certain patterns may not appear at all in a limited statistical experience, or too infrequently to permit "statistical discovery."

For example, the MMPI *Atlas* (**52**) reports only 1 per cent of male psychiatric patients as exhibiting a profile with a 68′ code. This weak characterization of a twelve-variable pattern extends only to two of the nine clinical scales. Suppose I wish to investigate whether the trait "acting out" in such cases is differentially expected depending on whether some other score, say Hy,is high or low. In order to locate enough cases to have even 10 high and 10 low Hy profiles among those with the required Pa, Sc peaks ≥ 70, the actuary would need a patient pool of approximately 2000!

The obvious retort is, "But where does the *clinician* get his experience of this particular pattern? Isn't he subject to the same sampling problem, *plus* the errors of human recall and weighting?" To the extent that the clinician is doing nothing but generalizing statistical experience, I think this objection is unanswerable. The answer, if any, has to be that the clinician in some cases synthesizes his personality description without specific experience with the particular pattern, utilizing his vague theoretical causal model as a means of extrapolating to regions of the profile space not hitherto sampled. His theory, poor as it is, as to the psychodynamics of the Hy scale may lead him to conclude that Hy elevations argue against acting-out in paranoid-schizoid individuals. It is surely possible for clinicians to think this way in the absence of direct statistical experience with 68′3 profiles. How *successful* such causal-theory-mediated extrapolations are likely to be is an empirical question. The main point here is this: "The best mathematical function will excel any other rule for combining the same data" is a tautology; but it is *not* a tautology that "The best mathematical function which can be appropriately fitted on the basis of a medium-sized statistical experience will excel the judgment-mediated decisions of a clinician who utilizes a causal-dynamic theory respecting the same scores and traits and who has the same limited statistical experience." This second proposition may or may not appeal to one's prejudices, and we have examined it at length in the preceding sections. But it is obviously not, like the first one, a purely mathematical or "logical" truth.

A further statistical difficulty in assessing this study concerns the inverse probability problem. From the fact that the trait "shy" is significantly correlated with the Pt scale (**58**, p. 143) it does not, of course, follow that when Pt is somewhat elevated, or even the peak score, "shy" should be attributed. Whether or not this is wise policy depends not only upon the shyness-Pt correlation but also upon the base rate of "shy" in the population under study. If 10 per cent of nurses-in-general are described as shy and a significantly greater fraction, say 40 per cent, of nurses with

Pt > 70 are so described, the best bet for either a high or low Pt case is still *"not* shy." Since the mechanical procedure attributed traits on the basis of four to six of the MMPI scales being classed as deviate (high or low), and anywhere from 9 to 24 of the traits were thereby attributed to each nurse, it seems very likely that traits were being attributed in individual cases which should not have been *from purely actuarial considerations.* If the clinician was being more sensible in this respect, he had an advantage, because the actuarial method was not being applied according to its own recognized rules. Finally, it is not clear what was the effect of presenting to the clinician only those traits which the mechanical rule scored for the particular case. In a way, this is letting the clinician correct the actuary, after first screening out certain potential clinical errors. Either the clinician accepts the trait, or he reverses the mechanical prediction. Presumably the reversals are very often cases where configural thinking enters the picture, otherwise he lets it alone. I think these several considerations show that this study must be interpreted with extreme caution, and that it indicates at most the superiority of a skilled MMPI reader to an undoubtedly non-optimal linear function. Out of the kindness of my heart, and to prevent the scoreboard from absolute asymmetry, I shall score this study for the clinician.

A final relevant study I have from personal correspondence with Henry Chauncey, president of the Educational Testing Service. In 1936 he undertook a comparison of predictive methods, the criterion being college grades (end of freshman year) and the subjects being a random sample of 100 Harvard entering freshmen. Statistical predictions were made on the basis of high school rank and College Board Examination scores. These were genuine predictions, i.e., made at the start of the freshman year. The clinical predictions were made by three members of the freshman dean's office, working independently on each case. These clinical predictions were based upon the same two quantitative items as the regression equation, *plus* letters of recommendation, information on extracurricular activities, and a statement by the student as to his reasons for coming to

Harvard. All four of the resulting validity correlations were in the .60's, the statistical validity ranking second. The difference between the statistical coefficient and that of the "best" clinician would not be significant with an N of 100.

Schneider, Lagrone, Glueck, and Glueck (**91**) studied the utility of the Glueck prediction tables in the military situation. The subjects of the investigation were 200 army general prisoners who had been delinquent in civilian life prior to their entrance into the army, and who were confined at a rehabilitation center at the time of study for having committed offenses while in the army. The clinical predictions (which were actually postdictions) were psychiatric diagnoses made at the rehabilitation center by army psychiatrists. These diagnoses were based upon many sources of data, such as FBI and police reports, data from service records, questionnaires filled out by employers, teachers, parents, relatives, former army associates, hospital reports, Red Cross social histories, and interviews by the psychiatrists and psychologists. In other words, the information available to the diagnosing psychiatrist included all the facts which were employed in the actuarial predictions, and more. If it is assumed that the diagnoses would have been the same had the psychiatric examination preceded the military arrest, and that a diagnosis of psychopathic personality, psychoneurosis, psychosis, or post-traumatic syndrome would have been considered at the induction center as predictive of failure in a military situation, 168 of the 200 cases, or about 84 per cent, would have been predicted to fail by the psychiatrist's diagnosis. These assumptions are certainly false, but they err in a direction which strongly favors the psychiatrist. That is, it can hardly be supposed that the percentage of correct diagnoses *after* failure would be less than the percentage at the time of induction. The actuarial prediction was based upon a mechanical combination of the five factors in Glueck's tables, namely, parental education, intelligence, age at first delinquency, age of beginning work, and industrial skill. The data needed in order to enter this prediction table would be easily available at induction, and in the present case were in fact

obtained independently of the information collected at the rehabilitation center. In other words, the prediction based Upon the Glueck tables is a prediction in a somewhat more genuine sense than that based upon psychiatric diagnosis. For these same 200 cases, the mechanical application of the Glueck tables would have resulted in identification of 84.5 per cent of the group of 200 soldiers. Here, then, the two methods are of equal predictive efficiency, if we ignore the fact that the study favors the clinical method by the nature of the time relationships involved. No indication of the false positive rate is given for either method. Since general experience with statistical screening devices in such situations suggests that many are screened who are not considered diagnosable upon closer psychiatric study, it is particularly important to know the false positive rates for the Glueck tables versus the psychiatrist. Since we lack this information, the present study can only be classed as indeterminate, and I have discussed it for the sake of completeness but have not included it in the summary tally.

In the interpretation of these studies, there are several complicating factors which must be kept in mind. In the first place, we know too little about the skill and qualifications of the clinicians who were making the predictions. For instance, there is no reason to assume that the guesses of an otherwise undescribed Bavarian physician will be based upon sufficient psychiatric insight so that they ought to be taken as fair samples of the outcome of clinical judgment.

Secondly, some of the studies have involved the comparison of clinical predictions with the predictions of regression equations in which the statistical weights were determined by the data of the group to be predicted, not cross-validated. Partly counterbalancing this is the fact that only seven of the multivariable studies used empirical weights assigned by efficient methods; the remaining eight assigned weights judgmentally or by other non-optimal methods.

Only five of these studies evaluate predictive efficiencies for the several clinicians separately. The clinician is a shadowy figure, and while it is important to know what the *average* clinician can do in competition

with the actuary, it is also important, and of even greater theoretical interest, to know whether there are some clinicians who can (consistently) do better than the regression equation. However, it is difficult to evaluate the argument sometimes offered that the *best* clinician is the appropriate representative for comparison with the statistical technique. Actually, if we judge from the studies reviewed, even this standard of evaluation would probably not do much to change the box score as between the two methods. But is the proposal a sound one anyhow? Statistical considerations make it clear that to apply such a principle in the interpretation of a single cross-sectional study of one set of clinical and actuarial predictions would be to introduce a serious pro-clinical bias, because the observed variation in hit rates is generated by whatever stable individual differences exist among the clinicians plus random errors, the relative contribution of these components being unknown and not accurately assignable in such a design. The seriousness of this problem is greater as we deal with more clinicians, and of course appears also on the side of the mechanical methods when more than one is tried (as in the Barron study). Presumably some kind of longitudinal study is needed to find out whether and to what degree the "good" clinician is stably such, rather than being merely the momentarily luckiest fellow among a crew of equal or near-equal mediocre guessers. Even a clear proof of stable differences among clinicians would still leave us with a serious practical problem.

Suppose that one in ten clinicians, sampled randomly from some national population, can do consistently better than the statistical formula in a specified type of prediction problem. In attempting to utilize such published information administratively in a different clinical installation, we definitely stand to lose if we treat all, or most, or a randomly chosen subset, of our clinical staff as if they were among that one tenth. Unless we have an accurate method of identifying this local representative of clinical genius, it is not of any practical value to know that he exists (or, more accurately, that there is a certain probability that we have one such). In fact, if we have fewer than seven clinicians on our staff, the odds are that

he is not among us at all! And, of course, even if he is, an over-all clinic policy of acting on clinical judgment when it contradicts the actuarial prediction will not pay off unless the unidentified expert is so markedly superior to the actuary that he can counterbalance the deficit accruing from our concurrent reliance upon the less efficient guesses of his colleagues. A little arithmetic will convince the reader that if only a small minority of clinicians excel the actuarial method, it would take an impossibly high superiority to justify a blanket shift to the clinical mode. For some values of the rates involved, algebraic constraints make it simply impossible for a few such deviates to be good enough (i.e., more than perfect!) to make up for the losses.

The only way to get around this problem is to identify the better than actuarial staff in each clinic. This in turn means either doing almost the whole comparison study over again in every installation, or developing highly accurate indirect methods (e.g., personality tests) for detection of such personnel. Evidence to date on the generality of such traits, as well as the general drift of selection studies in other areas, can hardly make us optimistic about this approach, although it should be thoroughly explored. Finally, unless there is greater generality over clinical skills than we have any reason to expect, not only the personnel involved but also the *prediction problem* cannot be changed without raising the issue of generality in predictive skill. And even if the generality were very high in correlational terms, we need to know how the *absolute* predictive efficiencies in the new problem compare with those of the actuarial method. The difficulties and complexities involved in the practical use of a finding that some subset of clinicians can excel the actuary are tremendous.

It is apparently hard for some clinicians to assimilate the kind of thinking represented in the preceding lines—chiefly, I gather, because they cannot help concentrating on the unfortunate case who *would* have been handled better *had* we followed the advice of super-clinician X, in defiance of the actuary. But what such objectors do not see is that in order to save that case, they must lay down and abide by a decision-policy

which will misclassify some other patient by defying the statistics. Presumably it hurts me as a patient just as much to be misevaluated regardless of whether the final mistaken judgment is made by a Ph.D. or by a clerk. A clinic's departure from the optimal method merely effects an exchange of some cases for others—but doesn't quite break even on the exchange. I do not quite know how to alleviate the horror some clinicians seem to experience when they envisage a treatable case being denied treatment because a "blind mechanical" equation misclassifies him, except to reiterate that the only way we could have prevented this happening to him would be to have employed a strategy which, while saving him, would systematically have guaranteed that the same error would be made with respect to somebody else whom we have in fact saved. This other error is, of course, just as blind as one made by the blindest cut-and-dried formula, since the plain fact is that the clinician, with wide-open eyes (and supernumerary ear) nevertheless did not see the world rightly. So it is the *number* of errors by the two methods that is all-important. In this connection see the excellent book by Bross (**16**) and the paper by Duncan *et al.* (**36**).

It seems to me from these considerations that our decision as to the economical and ethical thing to do cannot validly be influenced by the possibility of the best clinicians being somewhat better than the statistician; and that the burden of proof lies upon those who advance this argument to show empirically that such deviates show sufficient generality over the major predictive tasks and can be accurately identified by feasible methods. Until this is done the argument is on very shaky ground.

These studies do not tell us much about the kind and amount of clinical study that is competing with the actuarial method. On a priori grounds, one might expect that mediocre or poor clinical methods would be inferior to the actuarial, since the latter is always as good as the sample can make it, but that superior clinical methods might be better than the actuarial.

The studies did not involve collection of clinicians' judgments of their own confidence, and it is important to know whether a subset of the

clinically made predictions can be identified as better guesses by means of the associated feelings of assurance. This would imply, if the clinician and actuary are equal on the whole, another subset in which the clinician actually runs worse, which, of course, is quite possible.

Since the question being considered is the relative efficiency of actuarial and nonactuarial methods of combining the same data to yield a prediction, one might ask how we know that the clinician is actually making use of all the information at his disposal, or at least is employing to the best of his ability that fraction of the information which is being combined by regression techniques. In the studies cited we do not know how much of the relevant information the clinician is combining. There is some evidence in Sarbin's study that the clinician actually makes *less* use of some parts of the information than he should, or, more precisely, that he weights certain parts of the information too heavily. But I do not believe this should be prevented in such comparative studies, since so long as the data are made available to him, the clinician should be permitted to assign the predictive weights. There is no reason to exclude artificially the special case in which he assigns a zero weight to a factor and thus fails to use it in prediction. The ability of the clinician to weight the information is precisely what is being studied in such comparisons, and the decision as to whether $\beta = 0$ is a special case of this general problem of weight assignment. Consequently, all that the experimenter should do is to assure himself that any information used actuarially is *available* to the clinician, regardless of whether the latter sees fit to use it.

Future studies of the relative efficiency of the two methods should either require the clinician to predict for all cases, or else it should be arranged beforehand to assign a "doubtful" prediction to the categories or bands of values of the predictor variables for which the actuarial prediction is least trustworthy, so that the latter method will also be permitted to avoid prediction in a stated fraction of cases. Otherwise, meaningful comparison of the efficiency of the two methods is difficult.

In spite of the defects and ambiguities present, let me emphasize the brute fact that we have here, depending upon one's standards for admission as relevant, from 16 to 20 studies involving a comparison of clinical and actuarial methods, *in all but one of which the predictions made actuarially were either approximately equal or superior to those made by a clinician.* Further investigation is in order to eliminate the defects mentioned, and to establish the classes of situations in which each method is more efficient. I do not feel that such a strong generalization as that made by Sarbin is warranted as yet. Note that in terms of the *kind* of thing being predicted, there is not much heterogeneity, Essentially three *sorts* of things are being predicted in all but one of these studies, namely: (1) success in some kind of training or schooling; (2) recidivism; (3) recovery from a major psychosis, Studies of prognosis for outpatient psychotherapy of neurotics, probably the most important situation in terms of current predictive practices, are represented only by the work of Barron (**8**), in which the information available for clinical prediction was confined to age, sex, and the thirteen scores on the MMPI—hardly a typical setup as clinicians operate. Nevertheless, it is clear that the dogmatic, complacent assertion sometimes heard from clinicians that "naturally," clinical prediction, being based on "real understanding," is superior, is simply not justified by the facts collected to date. In about half of the studies, the two methods are equal; in the other half, the clinician is definitely inferior. No definitely interpretable, fully acceptable study puts him clearly ahead. In the theoretical section preceding we found it hard to show rigorously why the clinician *ought* to do better than the actuary; it turns out to be even harder to document the common claim that he in fact does!

Perhaps I ought to be embarrassed by this latter point, having devoted so much time to a theoretical discussion of how the clinician's operations *could* transcend the limitations of the clerical worker. Now I cite a mass of empirical studies indicating that as a matter of fact they do not. I imagine that most clinicians will feel themselves still persuaded of *something* about

clinical methods by the examples given in the theoretical section, and available from their own interview experience, in spite of the present studies. I have to admit that I share this weakness. At the risk of seeming to defend the clinician's special talents at any price, let me suggest some differences between the situations that convince clinicians of their powers and the situations dealt with by the studies I have cited. In suggesting these differences I am not trying to escape the burden of the nineteen studies. I believe they should be taken very seriously and that clinicians should be humbled by them. My purpose in the following remarks is rather to "explain" to myself as clinician *what* it is that these studies show, and to find out, if possible, how we could have been so mistaken in our expectations as clinicians about the outcome of such studies. Essentially I shall argue this: The kind of episode *during therapy* which gives us a conviction of our own predictive power may be quite legitimate, but the transition to the straight *prediction problem* involves features which seriously impair an analogy between the two sorts of situation. In other words, even if the clinician is right in believing that his "third ear" activity could not be duplicated by a clerk, *this should still not lead him to expect other results than those in the studies cited.*

In the first place, there is a major pragmatic difference between the predictive demands made upon the clinician during therapy and those made in the purely prognostic setting. All of us expect a certain amount of blind-alley hypothesizing to occur in the course of a therapeutic series. Therapists form transitory hypotheses of extreme tentativity and often may not follow them up by so much as a leading question unless additional support subsequently appears in the client's spontaneous productions. In interpretive therapy of moderate duration (say, 25 to 50 hours) one ordinarily expects some fraction of the total conversation to be devoted to the exploration of possibilities that turn out to be of minor significance or (more rarely) totally irrelevant. Nobody knows what the payoff rate is for these moment-to-moment guesses that come to therapists; but the over-all success frequency *might* be considerably less than

50 per cent and still justify the guessing, for unless the therapist is clumsy or the client unusually impatient, the time spent on exploration of poor guesses need not greatly detract from the positive contribution of successful ones. Presumably even the unsuccessful paths are rarely pure waste, since they contribute to such diverse concurrent aims as further getting acquainted, general desensitization, and incidental support for quite unrelated constructions (e.g., how free is the client to suggest that something is getting nowhere? Is he too docile in following the therapist into this blind alley? What does he do with the flat interpretations that are likely to emerge in these doldrums?). But even if the to-be-discarded hypotheses were *pure* filler, they would not impede the therapy except as they consumed time. The natural tendency of therapists to "mark where they hit and never where they miss" may lead them to have more confidence in their hypothesizing than is justified; but it is not clear just how they should be asked to alter their behavior if this is true. If the success frequency of intercurrent therapeutic hypotheses is .35 and a therapist thinks it is .60, knowledge of this discrepancy might lower his morale but it would be a complicated matter to say precisely how it should affect his interviewing procedures. The point is that Reik may be justified in the kind of thinking shown in the abortion example *even if, for any single guess, the odds are against him.*

When we move over into the straight prediction situation, all this is radically changed. Here, no erroneous weighting is filler in the above harmless sense, because statistical filler is error variance. Every time a clinician pays attention to a factor in predicting a single case, he is betting that the factor is *not* filler. Further, if he corrects his tentative prediction by a certain amount because of this factor, he is assigning a definite weight to it in the prediction equation. Error in this weighting (which no clinician will deny is practically inevitable) contributes to error in prediction. It is well known that if a variable *X* has zero beta in the true regression system, assignment of any non-zero beta will necessarily increase the mean squared error. So that in the straight prediction setting, *all bad ideas tend to*

subtract from the power of good ones. The realization of this difference between prognosis and therapy should help clinicians to accept the rather unfavorable findings of prediction studies, to the extent that their initial incredulity flows from a conviction that the third ear *does* payoff therapeutically. It may do so, and I myself believe that it does. But this is not good counter-evidence against Sarbin's claim that it does *not* payoff in the straight prediction situation.

Another difference between the Reik type of example and the quantitative studies cited is that the latter all involve the prediction of a somewhat heterogeneous, crude, socially defined behavior outcome. In Sarbin's study, for instance, we are not concerned to predict individual reaction-forms or even a specified disjunction of response classes, but rather a socially defined *outcome*, namely, satisfactory grades on the student's record. In the case of academic success, and even more strikingly in the case of failure, there are many alternative paths. The student's honor-point ratio is itself a statistical quantity, related only by great indirection to the multifarious individual reactions and decisions that were made over a long period of time, and which contribute by small increments to the determination of his particular number. It might seem that the case of parole violations constitutes an exception to this generalization, since one usually violates his parole by being detected in a single forbidden act. But this would be a superficial analysis of the case. While it is usually a particular action which comes to the attention of the parole officer, everybody knows that a large number of criminal occurrences do not come to the attention of the law; and of those which do, a considerable number never become legally attached to a particular criminal. It is very probable that those parolees whose parole is revoked because they have violated one of the rules of parole have been detected in *one* of a considerable number of criminally forbidden acts, the rest of which went undetected. This means that parole violators, as a group, are those who over a period of time have responded on a much larger number of occasions in forbidden ways than nonviolators, so that the *probability* of detection is considerably raised for the former group.

With this explanation, I think we can say that the kinds of prediction investigated in the studies cited have the common property mentioned. If we consider now the examples cited in my theoretical discussion of clinical activity, we see that they involve relatively specific and concrete predictions and postdictions, e.g., that the patient had an abortion performed, that on a certain night in walking home the patient was having the unconscious fantasy that she would leave her husband and make him feel sorry, in turn entailing the prediction that she would talk about material of this type in the succeeding moments of the interview. I think that if a clinician asks himself what kind of evidence causes him to remain unconvinced by the statistical studies I have cited, he will find himself thinking of individual predictions and postdictions of this concrete type.

If I am correct in this, we might think about it in the following way. Since a very large number of concrete situations, in relation to specific and general psychological complexes, determine jointly the long-time social outcome in the case of such a question as surviving in school, in order to predict this outcome by clinical understanding it would be necessary to formulate an extremely detailed conceptual model of personality structure. This model, each of the components of which would have to be highly confirmed, would then have to be combined with an extremely detailed account of the situations which the subject would meet during his first academic year. These two together would then be used to arrive at concrete predictions of many single episodes, or at most restricted classes of episodes—for example, whether the subject would attend a certain musical comedy two nights before his midquarter examination in differential equations. These concrete predictions would then be cumulated in some complicated way to arrive at the prediction of the honor-point ratio. Now it is obvious that in none of the studies cited did the clinician have an opportunity to "formulate the personality" or to determine the *press* in anything like the detail indicated. Under these circumstances the appropriate attitude would be something like this:

"In order to predict the broad social outcome of *getting through college*, I would have to know a very great deal about the individual, which I cannot learn in a matter of an hour's interview, or even in several hours. I would also have to know beforehand a great deal about the kinds (and dates!) of situations to which he was going to be subjected, which would not be possible even with an army of social workers at my disposal. Therefore, I shall abandon the effort to mediate my predictions by means of actual hypothesis formations concerning the personality structure. Instead, I shall fall back on the well-known psychological rule that the best way to predict the way a person is going to act is to find out how he has acted in the past. I know that there are a great many ways of behaving which contribute to academic success, and their relationship to the need structure of an individual is so complicated that only a very intensive study would enable me to make use of this need structure in prediction. But if I take more of an 'empty organism' attitude, I can simply ask, 'Have the complex, heterogeneous behaviors of this individual in the past been on *the whole such that he has achieved academically*?' If they have, I shall make the assumption (which would be true for the great majority of people in my sample) that his behavior dispositions, *whatever they are*, will remain relatively stable during his first year in the arts college. Those students who have, in their diverse ways, behaved so as to get good marks in high school and good marks on my entrance examination, will usually continue to behave in the same sorts of ways. To attempt to *characterize* these sorts of ways in any detail, upon the evidence available to me, would be fruitless."

I am inclined to think that the same sort of consideration would apply in the case of a more strictly clinical domain, such as the prediction of response to psychotherapy. If I wanted to know whether a certain event had occurred in the patient's past, or whether a circumscribed topic would recur in the associations in the next few interviews, I think I would prefer to proceed nonactuarially. If, on the other hand, I was asked to predict whether a patient would respond well to psychotherapy which, after all,

involves a socially defined outcome achieved by a very large number of individual learnings, I would trust such statistics as the duration of his illness, the number of previous therapeutic efforts made upon him, and the prognosis associated with his psychiatric diagnosis, more than any clinician's judgment based upon an estimation of his dynamics.

One very striking difference between the empirical studies and the clinical examples lies in the *form* in which the prediction problem is couched. The task facing the actuary is rather like that presented on a multiple-choice test, in the sense that the actuary (and the clinician competing with him in these studies) is initially presented with the available alternatives. Thus, we are told that the students in Sarbin's study must either fail or pass, that the schizophrenics in Wittman's study either recover or remain ill, and so on. The class of possibilities is indicated for us, and the predictive task we face for each individual case is to assign him to one of these. Even in the continuous case, such as predicting freshman honor-point ratio, we are still informed that the variable being predicted is a number from –1 to 3.0, and we are aware of a good deal of qualitative matter about this dimension. Now contrast with this the clinical examples we have discussed, Here the clinician has in a sense to create the prediction, not merely to say "Yes" or "No" to certain indicated alternatives.

I am not here talking about whether, in some philosophical sense, the actual set of meaningful possibilities is finite. What I wish to stress is the concrete psychology of the task as presented *externally*. Sarbin's clinicians, as well as his statistical clerk, were told the nature and range of the continuum to be predicted. They did not have to call upon their previous experience, hoping that by synthesis and recombination their brains would, so to say, form and then riffle through all the possible states of affairs. By contrast, in, say, Reik's postdiction of the abortion, the population of alternatives is as large as that of human thought and life. Any restriction of the hypotheses to a narrow class, however plausible, has come from the clinician. No one has listed for Reik a set of 30,000 latent

thoughts (which could yield silence followed by "There's a book on its head"), one of these being that of an abortion. Somehow this response must be *emitted* by Reik in the presence of the ambiguous stimulus. In a very real sense it is the difference between taking an individual Rorschach and filling out a true-false inventory. Or, to take the example of the raven, no one has ever seriously proposed providing clinicians with a master list which would contain all possible fantasies regarding sleeping husbands. Yet, from the standpoint of mere *mechanics*, how else, lacking such a preposterous table, can we enable the nonclinical clerk to think of them? Statistical weights enable us to assign probabilities to values of variables (including the disjuncts in a class of named alternative situations). They do not, quite obviously, enable us to fantasy the situations or to list their names!

If the Polansky study were repeated using actuarial methods, but with the prediction problem remaining as specific as it was in his investigation, it would be interesting to see whether the advantage of the actuarial method would appear. These considerations are, of course, entirely speculative, and it is of the greatest importance that suitable experimental designs should be worked out for the actuarial study of such moment-to-moment clinical predictions as are discussed in our theoretical section. The possibility that the choice of more suitable prediction problems, in which the advantage of structural-dynamic hypotheses would have more chance to show itself, might lead to superiority of the clinical method, should not detract from the practical significance of such empirical studies as those by Sarbin and Wittman. Prediction of length of hospitalization, response to certain kinds of treatment, and perhaps even exacerbation of illness, resemble the generic socially defined sort of case more than they do a specific event or content prediction. On the present evidence, I think it would be wise for a psychologist who is asked by the psychiatric staff whether or not a patient will benefit from shock treatment to put his reliance on actuarial rather than nonactuarial procedures.

All the above comparisons have treated efficiency solely in terms of

predictive success. This method of evaluation, while illuminating for theoretical purposes, actually gives the clinician a considerable advantage. For practical purposes, the concept of efficiency must include some reference to the amount and level of work required to arrive at a given degree of predictive success. Once some sort of statistical backlog has been collected (and this takes no more time than is needed for the clinician to get experience), the actuarial method almost invariably takes less time, less effort, and—no minor point—can be entrusted to lower paid personnel possessing much less skill. Any realistic assessment of the comparative efficiencies of the two methods must give very heavy weight to these considerations. A concrete feeling for this point may be readily gained by reading the report of Kelly and Fiske (**63**) in which the mechanical use of the appropriate Strong scores obtainable by mail at low cost is more effective than seven man-days of skilled clinical time and a cost of $300 per case.

The several hours of highly skilled work sometimes expended in arriving at a dynamic formulation of the patient by an ingenious extrapolation of test results could very possibly be spent much better in added hours of psychotherapy. Whether the patient is seen in private practice or in a charity hospital, the skilled clinician is being paid, and someone is footing the bill. It has often struck me as paradoxical to find a near-routine battery of complex, skill-demanding tests being administered in a clinical setting where the median number of therapeutic hours per case is not appreciably in excess of the total skilled time expended by the psychologist on the case in making often dubious dynamic and prognostic inferences from the test data. A really honest examination of this sort of question contains, needless to say, a great deal of dynamite for the profession. Sooner or later it must be done; and the socially significant meaning of the phrase "predictive efficiency" will have to be employed rather than the theoretical meaning we have used throughout the present discussion.

Although I can present no statistics on the matter, I have a distinct impression that the amount of time expended by a psychologist in the

administration of multiple and fancy diagnostic devices, and in the dynamic formulation based upon an alleged integration of them, progressively declines as this factor of economics is increasingly highlighted by the character of his practice. Psychologist X gives four or five projective devices, a Shipley, a Wechsler-Bellevue, and an MMPI, and chews over the results *ad nauseam* when he is functioning in an institutional context. He picks up his semimonthly check in any event, so the value of time per case, while willingly conceded as important *in abstracto*, is not strikingly called to his attention. We find the battery, and the time spent on interpreting it, undergoes a suspicious shrinkage when our psychologist acts as consultant to a privately practicing psychiatrist with the latter collecting his fee for him. Lastly, behold him in his own private therapeutic practice, where he *himself* is the evaluator of the therapeutic power conferred by his armamentarium and he himself has to put the financial bite on his client. His enthusiasm for "advance knowledge through dynamic integration" has now so flagged that we find him slipping the client a quick Bender and sending him home with a group form MMPI to be filled out between sessions. I have a hunch that some profound and terrifying truths are discernible in this psychometric devolution but I shall not press the point.

9

General Remarks on Quantification of Clinical Material

IN DISCUSSING the problem of actuarial prediction one often comes across certain misconceptions held by the more tender-minded clinicians which prevent clear thinking. For instance, there is still the misconception that mathematical descriptions of persons in terms of scores *require* that persons achieving identical scores should be identical or indistinguishable with respect to the traits so quantified. We sometimes hear this view expressed by such statements as "A human being is more than just a set of numbers." It is pointed out that two persons who achieve a score of 1.5 sigmas above the mean on an introversion test do not manifest their introversion in precisely the same way, and that they did not arrive at it via the same sequence of experiences. The first thing to see about such a statement is that it is true. But this indubitable uniqueness of the single case is no more fatal to psychology than it is to physics. To see it as fatal to psychological quantification is to forget that the class character of concepts and dimensions is found in all descriptive enterprises. As Cattell says, "It seems that one must subscribe to the extreme sense of Allport's argument and admit that *all* traits are in some way unique" (**28**, p. 61). No two individuals are exactly alike, and no verbal or mathematical characterization can do complete justice to their individuality. No two explosions are identical nor can any system of equations give a description of any of

them which is exhaustive. As Thurstone has pointed out, those who object to assigning the same score to two introverts because their introversion is distinguishable should in all consistency object to saying that two men have the same income since one of them works and the other steals (**100**, p. 54). A cannon ball falling through the air is "more than" the equation $S = \frac{1}{2} g t^2$, but this has not prevented the development of a rather satisfactory science of mechanics. The exhaustive description of an individual event is not aimed for in the scientific analysis of the world *nor can it be hoped for in any descriptive enterprise* (**54**, **76**). All macroscopic events are absolutely unique. It is a further mistake to exaggerate the degree to which this lack of concreteness reflects a special failing of the scientist, since there is *no* kind of human knowledge which exhaustively characterizes direct experience by a set of propositions. No set of percentile ranks, no graphical representation of personality components, *and no paragraph of characterological description* can contain all the richness of our immediate experience. The abstractive or summarizing character of descriptions is shared by differential equations, maps, gossip, and novels alike. So-called scientific description, however, abstracts those things which are most relevant in terms of causal-analytic and predictive aims; and, secondly, employs a language (mathematical when possible, but not always!) which minimizes ambiguity.

Further objections associated with the one just mentioned imply that quantitative descriptions cannot yield a unique person. Stouffer (**95**) has pointed out that with only ten traits, each of which may take on only four values, there are somewhat over a trillion possible unique individuals. It is well known that the science of fingerprinting makes use of a small number of dimensions and is nevertheless capable of identifying the unique case. In the quantitative case of continuous variables this is even more obvious.

It is sometimes suggested that mathematical description assumes that equal amounts of a component must always mean the same thing psycho-

For instance, the difference between zero M and 4 M in the Rorschach of a bright adult "means" more than the difference between 7 M and 11 M. What actuarial assumption denies this? We know that a change from 98.6° to 99.6° is more significant of pathology than one from 101° to 102°. The confusion present in this argument is perhaps partly the fault of the psychological statisticians who have confined themselves largely to the study of linear relations such as are used in multiple regression; but these arbitrary restrictions are, of course, not a necessary consequence of the application of mathematical methods. Incidentally, there is a lot of loose talk around these days about nonlinear relations. I do not doubt that there are a large number of such in the behavior domain, but we ought not to browbeat the statisticians with this phrase until we know more about where these nonlinear dependencies occur and how much they payoff predictively over and above the much-maligned linear regression system! Clinicians intoxicated with the abstract idea of nonlinearity and interaction of variables might contribute to their historical perspective by reading a short paper of Thorndike's published in 1918 (**99**).

A further confusion is involved in the frequent claim that mathematical description or prediction involves the assumption of simple additive relations among the variables and is inadequate to deal with dynamic interactions. ("Additive" is a favorite pejorative epithet with some clinicians.) For example, the occurrence of several M with 7P and 88 per cent F+ on the Rorschach is healthy. The same number of M with only 1P and 50 per cent F+ suggests a malignant break of fantasy from reality, such as in a paranoid schizophrenia. Zero M with 1P and 50 per cent F+ might be suggestive of low intellect. The only remark to be made on this score is that mathematical analysis does not in any way exclude such possibilities. The mathematical treatment of nonadditive situations is found in great profusion among the formulations of Hullian learning theory (a system which I have heard clinicians confusedly stigmatize as "atomistic"; Hull's composite expression for reaction potential involves a *product* of several

part functions of independent variables, and even these part functions are not simple linear relations). Probably much of the talk about "patterning" as something to be contrasted with "statistics" would not occur but for the fantastic mathematical ignorance of most clinicians. I have heard clinicians discuss the topic in such a fashion that the only possible inference a listener could draw was that they had never heard of the interaction term of the analysis of variance!

It is difficult to attempt a precise discussion of the problem of patterning, since we generally use the term with a certain looseness, to characterize what may be a very heterogeneous collection of types of functional dependency. However, a perusal of the clinical literature leads to the identification of one kind of situation as the commonest to which the term is applied. That is the situation in which *the indication of a given variable with respect to the criterion is not constant,* but the weight, and possibly even the *direction* (sign) of contribution of that variable, are functions of the values which the other predictor variables have taken on. How important such a refinement actually is remains an empirical problem which I shall not consider here; but the clinician can be pardoned for his irritation when a nonclinician academic psychologist with some statistical interest informs him on the one hand that whatever the clinician does is essentially statistical; and on the other hand, fails to present him with convenient tools for expressing such a state of affairs. It is sometimes said that the differential weights of the regression equation or of the discriminant function were devised specifically to take account of pattern relationships, It is obvious that the kind of patterning which we are considering here is *not* adequately dealt with by such procedures.

Probably the most striking instance of patterning is a situation in which *neither* variable is related to the criterion, and yet the criterion is predictable to some degree from a knowledge of the values the *two* variables take on. I am not talking here about the familiar cases which appear paradoxical only because the Pearson *r* is used as the indicator of a relationship to which it is unsuited. I mean unrelated in the strict, general

mathematical sense of *independence* as it is defined in probability theory. If variables x and y are both totally independent of a criterion z, is it possible to predict z from a knowledge of x and y alone? I have discussed the possibility of such an extreme case elsewhere (**72**), and Horst (**56**) has recently presented a generalized mathematical treatment. In order for patterning effects to occur, it is, of course, not necessary that the unpatterned variables have zero validity, although that is the most striking case for getting the point across. In the case of continuous predictive functions, what makes the system patterned is that the rate of change of the criterion estimate with respect to one of the predictor variables depends upon one or more of the other predictor variables. This is a stronger claim than mere nonlinearity, of which it is one, but not the only, form. A predictive function $y = \log \sin x_1 + x_2^3$ is not linear and would be rather poorly approximated by the usual multiple regression methods. But neither is it patterned, because the mode of dependence of y upon x_1 is invariant with respect to the values taken on by x_2, and conversely. On the other hand, a predictive function such as $y = x_1 + x_2x_3$ is patterned, because the effect of an increment in x_2 depends upon the value of x_3. Similarly, $y = x_1 + x_2^{(1 - ax_3)}$ is patterned, in a more complex manner. If the values of the x-variables are grouped and thus divided into discontinuous *levels*, what we have is simply a significant interaction term in the analysis of variance.

Although it does not help us in hitting upon a configured predictive function or in determining its parameters, I should like to present an abstract definition of patterning for the continuous case. I offer this mainly for the benefit of statisticians who have wondered what clinicians could reasonably mean by their talk of patterning, over and above (1) differential weighting and (2) nonlinearity. The nature of the variables—i.e., whether phenotypic or genotypic, measured or judged, present or future, outer or inner, psychometric or case history, etc.—is irrelevant. Nor is the appropriateness of the metric relevant. Consider a predictive function $y = f(x_1, x_2, x_3, \ldots x_m)$. Differentiating partially with respect to x_i and then

repeating this with respect to x_j, we examine the second-order mixed partial derivative

$$\frac{\delta^2 y}{\delta x_i \delta x_j}.$$

If it is not identically zero (or at least equals zero for all values of x_i, x_j within the empirically realized range), we say that the predictor variables x_i and x_j are *patterned* with respect to the criterion. If this derivative vanishes over the range, they are *unpatterned.* If all the m(m – 1)/2 mixed partials of this sort vanish, the prediction *system* is unpatterned (or, if it makes anyone happy to call it that, "atomistically" related to the criterion). If all these partials are non-zero, the function may be said to be *patterned pairwise.* This is what is being claimed if we say "You can't interpret any Rorschach variable independently of any of the others." If there exists a *k*th-order mixed partial derivative

$$\frac{\delta^k y}{\delta x_1 \delta x_2 \delta x_3 \ldots \delta x_k} \neq 0,$$

but all the $(k+1)$th order mixed partials vanish, the system may be described as *patterned of order k.* Finally, if we partially differentiate successively with respect to all *m* of the variables and find that the *m*th order mixed partial derivative

$$\frac{\delta^m y}{\delta x_1 \delta x_2 \delta x_3 \ldots \delta x_m}$$

is nonvanishing, we say the system is *totally configurated* with respect to the criterion.

This last, very strong condition is a precise formulation of the common claim for multivariable tests that the interpretation of any variable depends upon the *interrelation* of *all* the others. I have never been so fortunate as to see anyone actually perform this feat with any clinical instrument, and I doubt very much that it is possible for the finite mind. But if it ever does occur, the preceding is a mathematical rendition of it. It is worth noting that Q correlations, sometimes said to represent the

personality pattern, are not patterned in the sense defined above; since, while the Q correlation is expressible as a function of a sum of squared differences (and hence it is nonlinear in the predictors), if one were to use it predictively, i.e., to predict the salience of a given trait in a self-sort from knowledge of the therapist's other-sort, the usual unpatterned (and linear!) regression equation would be employed. It goes without saying that any *practical* application of genuinely patterned systems, particularly those of high order, will require the development of powerful searching methods for choosing the functions, and for estimating the constants. It seems unlikely, however, that any mathematical progress will free us from the necessity of a large N, to counteract the excessive sampling instability alluded to above in our discussion of the Hovey-Stauffacher investigation.

A major research need is further empirical comparison of the two methods of prediction, with the elimination of the disturbing factors mentioned previously. On the formal side, we shall have to wait for the logicians to achieve a clarification of the nature of the concept of probability, especially the probability of hypotheses, and the general formulation of inductive logic. Systematic studies should be undertaken of the success frequency of certain *subsets* of the clinician's predictions. For instance, at what type of prediction is he best? What importance should be assigned to his own subjective degree of confidence? When the clinician and the actuary are in disagreement, to whom should we listen? This latter is important because one commonly hears it said by psychiatrists that they are predicting for the individual case, so that the greater success frequency of the actuary, even if clearly established, is treated as of no importance in practice. This thinking is, of course, thoroughly muddled. In any given instance, we must decide on whom to place our bets; and there is no rational answer to this question *except* in terms of relative frequencies. If, when the clinician disagrees with the statistics, he tends to be wrong, then, if we put our bets *in individual instances* upon him, we will tend to be wrong also.

10

A Final Word: Unavoidability of Statistics

ALL clinicians should make up their minds that of the two uses of statistics (structural and validating), the validating use is unavoidable. Regardless of one's theory about personality and regardless of one's choice of data, whether Rorschach, MMPI, Bender, age, marital status; regardless of how these data are fused for predictive purposes—by intuition, table, equation, or rational hypotheses developed in a case conference—the honest clinician cannot avoid the question "Am I doing better than I could do by flipping pennies?" In answer to a demand for validation, one sometimes hears it stated that, whereas a certain clinical device or method has not been proved valid "in the usual sense," such formal validation is not required, since the instrument has been validated by its "clinical usefulness." When we hear this from a clinician, all we can say is that he *thinks* he is using it to advantage. Out of the welter of diverse cases, with mixed data and complex judgments, you simply cannot tell whether your use of a procedure is paying off or not. Consider almost any clinical instrument; there are many people, neither fools nor knaves, who are willing to stand up in defense of it. Others, equally competent, invoke the same kind of evidence—clinical experience—as a basis for discarding it as useless. The untrustworthiness of clinical impressions is by no means confined to the behavior disorders, of course. Over a period of years patients suffering from multiple sclerosis were treated by the use of vitamins, diathermy,

oral administration of spinal cords, high dairy diets, potassium iodide, quinine bisulphate, and now we have histamine. All these treatments found support with certain clinicians on the grounds that they were proving themselves useful in clinical practice. Most of them were subsequently abandoned when people began to keep systematic records of the ultimate results. Among his many virtues, the characteristic vice of our colleague the psychiatrist is his tendency to draw conclusions before graphs, and some detect a growing tendency for clinical psychologists to be cheerfully infected by this vice.

What can clinical validation legitimately mean? Let us admit that validity is to be established by the application of a technique in the real life situation. Not all human motives are readily transplantable to the laboratory. Nevertheless, we must keep track of our guesses. "Leaving the laboratory" is not equivalent to "scrapping the rules." It is a common error to group the terms "quantitative," "statistical," and "experimental" together, setting them into opposition with "qualitative," "clinical," "non-experimental." I have even heard psychologists use the terms "quantitative" and "documentary" in such opposition, whereas it is obvious that the *quantitative* study of *documents* is a rapidly growing and powerful science. I would defend simultaneously (and, I hope, consistently) the two propositions that (1) there are some behavior phenomena which cannot be best studied in the laboratory, at least with any confidence in one's extrapolations, and (2) until some quantification, at least frequency counts and contingency measures, is applied to clinical evidence, we can have very little confidence in our claims.

Is any clinician infallible? No one claims to be. Hence, sometimes he is wrong. If he is sometimes wrong, why should we pay any attention to him? There is only one possible reply to this "silly" question. It is simply that he *tends* (read: "is likely") to be right. "Tending" to be right means just one thing—"being right in the long run." Can we take the clinician's word for this? Certainly not. As psychologists we do not trust our memories, and have no recourse except to record our predictions at the time,

allow them to accumulate, and ultimately tally them up. We do not do this because we have a scientific obsession, but simply because we know there is a difference between veridical knowledge and purported knowledge, between knowledge which brings its credentials with it and that which does not. After we tally our predictions, the question of success (hits) must be decided upon. If we remember that we are psychologists, this must be done, either by some objective criterion, or by some disinterested judge who is not aware of the predictions. When as clinicians we have done all these things, and thus provided a secure basis for deciding how much trust we *can* put in ourselves, what have we done? We have carried out a validation study of the traditional kind! I am led by this reasoning to the conclusion, in complete agreement with Sarbin, that the introduction of some special "clinical utility" as a surrogate for validation is inadmissable. If the clinical utility is really established and not merely proclaimed, it will have been established by procedures which have all the earmarks of an acceptable validation study. If not, it is a weasel phrase and we ought not to get by with it.

If a clinician says "This one is different" or "It's not like the ones in your table," "This time I'm surer," the obvious question is, "Why should we care whether you think this one is different or whether you are surer?" Again, there is only one rational reply to such a question. We have now to study the success frequency of the clinician's guesses when he asserts that he feels this way. If we have already done so, and found him still behind the hit frequency of the table, we would be well advised to ignore him. Always, we might as well face it, the shadow of the statistician hovers in the background; *always* the actuary will have the final word.

References

1. Abel, T. (1948) The operation called *Verstehen. American Journal of Sociology,* 54:211-18.
2. Alexander, F. (1934) Evaluation of statistical and analytical method in psychiatry and psychology. *American Journal of Orthopsychiatry,* 4:433-38.
3. Allport, F. H., and N. Frederiksen. (1941) Personality as a pattern of teleonomic trends. *Journal of Social Psychology,* 13:141-82.
4. Allport, G. W. (1937) *Personality.* New York: Holt.
5. Allport, G. W. (1942) *The use of personal documents in psychological science.* S.S.R.C. Bulletin No. 49.
6. Ash, Philip. (1949) The reliability of psychiatric diagnoses. *Journal of Abnormal and Social Psychology,* 44:272-76.
7. Baldwin, A. L. (1942) Personal structure analysis: A statistical method for investigating the single personality. *Journal of Abnormal and Social Psychology,* 37:163-83.
8. Barron, Frank. (1953) Some test correlates of response to psychotherapy. *Journal of Consulting Psychology,* 17:235-41.
9. Beck, S. J. (1944) *Rorschach's test. I. Basic processes.* New York: Grune and Stratton.
10. Bergmann, G. (1944) Holism, historicism, and emergence. *Philosophy of Science,* 11:209-21.
11. Berne, E. (1949) The nature of intuition. *Psychiatric Quarterly,* 23:203-26.
12. Blenkner, M. (1954) Predictive factors in the initial interview in family casework. *Social Service Review,* 28:65-73.
13. Bloom, R. F., & E. G. Brundage. (1947) Prediction of success in elementary schools for enlisted personnel. In D. B. Stuit (ed.), *Personnel research and test development in the Bureau of Naval Personnel.* Princeton, NJ: Princeton University Press. Pp. 233-61.
14. Bobbitt. J. M., & S. H. Newman. (1944) Psychological activities at the United States Coast Guard Academy. *Psychological Bulletin,* 41:568-79.
15. Borden. H. G. (1928) Factors for predicting parole success. *Journal of American Institute of Criminal Law and Criminology.* 19:328-36.
16. Bross, I. D. J. (1953) *Design for decision.* New York: Macmillan.
17. Burgess, E. W. (1928) Factors determining success or failure on parole. In A. A. Bruce, (ed.), *The workings of the indeterminate sentence law and the parole system in Illinois.* Springfield, Illinois.

18. Burgess, E. W. (1941) An experiment in the standardization of the case-study method. *Sociometry,* 4:329-48.
19. Burgess, E. W. (1942) Rejoinder, *American Journal of Sociology.* 48:84-86.
20. Burgess, E. W., & L. S. Cottrell. (1939) *Predicting success or failure in marriage.* New York: Prentice-Hall.
21. Burt, C. (1941) *The factors of the mind*. New York: Macmillan.
22. Carnap, R. (1945) The two concepts of probability. *Philosophy and Phenomenological Research,* 5:513-32.
23. Carnap, R. (1945) On inductive logic. *Philosophy of Science,* 12:72-97.
24. Carnap, R. (1946) Remarks on induction and truth. *Philosophy and Phenomenological Research,* 6:590-602.
25. Carnap, R. (1947) Probability as a guide in life. *Journal of Philosophy,* 44:141-48.
26. Carnap, R. (1947) On the application of inductive logic. *Philosophy and Phenomenological Research,* 8:133-48.
27. Carnap, R. (1948) Reply to Felix Kaufman. *Philosophy and Phenomenological Research,* 9:300-4.
28. Cattell, R. B. (1946) *Description and measurement of personality.* Yonkers, NY: World Book.
29. Cattell, R. B. (1947) P-technique demonstrated in the determination of psycho-physiological source traits in a normal individual. *Psychometrika,* 12:267-88.
30. Chapman, D. W. (1934) The statistics of the method of correct matchings. *American Journal of Psychology,* 46:287-98.
31. Chauncey, Henry. Personal communication.
32. Chein, I. (1945) The logic of prediction: some observations on Dr. Sarbin's exposition. *Psychological Review,* 53:175-79.
33. Conrad, H. S, & G. A. Satter. (1945, Sept. 3) Use of test scores and quality classification ratings in predicting success in Electrician's Mates School, OSRD Report No 5667.
34. Cottrell, L. S. (1941) The case-study method in prediction. *Sociometry,* 4:358-70.
35. Davis, F. B. (1947) *Utilizing human talent.* Washington, D.C.: American Council on Education.
36. Duncan, O. D., L. E. Ohlin, A J. Reiss, & H. R. Stanton. (1953) Formal devices for making selection decisions. *American Journal of Sociology,* 58:573-84.
37. Dunham, H. W., & B. N. Meltzer. (1946) Predicting length of hospitalization of mental patients. *American Journal of Sociology,* 52:123-31.
38. Dunlap. J. W., & M. J. Wantman. (1944) *An Investigation* of *the interview as a Technique for Selecting Aircraft Pilots.* Civil Aeronautics Administration Report No. 33. Washington, D.C.
39. Elkin, F. (1947) Specialists interpret the case of Harold Holzer. *Journal of Abnormal and Social Psychology,* 42:99-111.

40. Estes. S. G. (1938) Judging personality from expressive behavior. *Journal of Abnormal and Social Psychology*, 33:217-36.
41. Feigl, H. (1950) Existential hypotheses. *Philosophy of Science*, 17:35-62.
42. Felix, R H., D. C. Cameron, J. M. Bobbitt, & S. H. Newman. (1945) An integrated medico-psychological program at the United States Coast Guard Academy. *American Journal of Psychiatry,* 101:635-42.
43. Fenichel, O. (1941) *Problems of psychoanalytic technique.* New York: Psychoanalytic Quarterly.
44. Foreman, P. B. (1948) The theory of case studies. *Social Forces.* 26:408-19.
45. Freud, S. (1924, 1950) *Collected papers.* Vols. I-V. London: Hogarth.
46. Goodenough, F. L. (1932) Expression of the emotions in a deaf-blind child. *Journal of Abnormal and Social Psychology,* 27:328-33.
47. Guilford, J. P. (ed.) (1947) *Printed classification tests.* AAF Aviation Psychology Program Research Report No. 5. Washington, D.C.: U.S. Government Printing Office.
48. Hamlin, R. (1934) Predictability of institutional adjustment of reformatory inmates. *Journal of Juvenile Research,* 18:179-84.
49. Hanks, L. M. (1936) Prediction of case material from personality tests. *Archives of Psychology,* No. 207.
50. Harrison, R. (1943) The TAT and Rorschach methods of personality investigation in clinical practice. *Journal* of *Psychology,* 15:49-74.
51. Hathaway, S. R. (1947) A coding system for MMPI profiles. *Journal of Consulting Psychology,* 11:334-37.
52. Hathaway, S. R., & P. E. Meehl. (1952) *An atlas for the clinical use of the MMPI.* Minneapolis: University of Minnesota Press.
53. Hebb, D. O. (1946) Emotion in man and animal: An analysis of the intuitive processes of recognition. *Psychological Review,* 53:88-106.
54. Hempel, C. G. (1949) The function of general laws in history. In H. Feigl, and W. Sellars (eds.), *Readings in philosophical analysis.* New York: Appleton-Century-Crofts.
55. Hollingworth, H. L. (1923) *Judging human character.* New York: Appleton-Century.
56. Horst, P. (1954) Pattern analysis and configural scoring. *Journal* of *Clinical Psychology,* 10:3-11.
57. Horst, P. (1941) *Prediction of personal adjustment.* S.S.R.C. Bulletin No. 48.
58. Hovey, H. B. (1953) MMPI profiles and personality characteristics. *Journal* of *Consulting Psychology,* 17:142-46.
59. Hovey, H. B., and J. C. Stauffacher. (1953) Intuitive versus objective prediction from a test. *Journal of Clinical Psychology,* 9:349-51.
60. Hull, C. L. (1943) *Principles of behavior.* New York: Appleton-Century.
61. Jenkins, R. L. (1945) The relationship between scientific and pre-scientific methods in psychiatry and mental hygiene. *Mental Hygiene,* 29:78-94.

62. Kelly, E. L., & D. W. Fiske. (1948) *The selection of clinical psychologists: Progress report and preliminary findings*. Ann Arbor, MI.: Edwards Letter Shop.
63. Kelly, E. L., & D. W. Fiske. (1950) The prediction of success in the V.A. Training Program in Clinical Psychology. *American Psychologist*, 5:395-406.
64. Kelly, E. L., & D. W. Fiske. (1951) *The prediction of performance in clinical psychology.* Ann Arbor, MI: University of Michigan Press.
65. Klopfer, B., & D. M. Kelley. (1946) *The Rorschach technique*. Yonkers, NY: World Book.
66. London, I. D. (1945) Psychology and Heisenberg's Principle of Indeterminacy. *Psychological Review*, 52:162-68.
67. London, I. D. (1946) Some consequences for history and psychology of Langmuir's concept of convergence and divergence of phenomena. *Psychological Review*, 53:170-88.
68. Luft, J. (1950) Implicit hypotheses and clinical predictions. *Journal of Abnormal and Social Psychology*, 45:756-59.
69. Luft, J. (1951) Differences in prediction based on hearing versus reading verbatim clinical interviews. *Journal of Consulting Psychology*, 15:115-19.
70. Lundberg, G. A. (1941) Case-studies vs. statistical methods—An issue based on misunderstanding. *Sociometry*, 4:379-83.
71. MacCorquodale, K., & P. E. Meehl. (1948) On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55: 95-107.
72. Meehl, P. E. (1950) Configural Scoring. *Journal of Consulting Psychology*, 14:165-71.
73. Melton, R. S. (1952) A comparison of clinical and actuarial methods of prediction with an assessment of the relative accuracy of different clinicians. Unpublished Ph.D. thesis, University of Minnesota.
74. Munroe, R. (1942) An experiment in large scale testing by a modification of the Rorschach method. *Journal of Psychology*, 13:229-63.
75. Murray, H. A. (1938) *Explorations in personality.* New York: Oxford University Press.
76. Nagel, E. (1952) Some issues in the logic of historical analysis. *Scientific Monthly*, 74:162-69.
77. Newman, S. H., & J. M. Bobbitt. (1948) The development of entrance tests for the United Stales Coast Guard Academy. *Journal of Applied Psychology*, 32:248-54.
78. Newman, S. H., J. M. Bobbitt, & D. C. Cameron. (1946) The reliability of the interview method in an officer candidate evaluation program. *American Psychologist*, 1:103-9.
79. Piotrowski, Z. (1946) Differences between cases giving valid and invalid personality inventory responses. *Annals of the New York Academy of Science*, 46:633-38.
80. Polansky, N. (1941) How shall a life-history be written? *Character and Personality*, 9:188-207.
81. Psychological Research Unit No.1, Army Air Forces. (1943-44) *Clinical techniques project*. Mimeographed bulletin. Restricted, unavailable.
82. Reichenbach, H. (1938) *Experience and prediction.* Chicago: University of Chicago Press.

83. Reik, T. (1948) *Listening with the third ear.* New York: Farrar, Straus.
84. Sarason, S. (1948) The TAT and subjective interpretation. *Journal of Consulting Psychology,* 12:285-99.
85. Sarbin, T. R. (1941) Clinical psychology—Art or science? *Psychometrika,* 6:391-400.
86. Sarbin, T. R. (1943) A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology,* 48:593-602.
87. Sarbin, T. R. (1944) The logic of prediction in psychology. *Psychological Review,* 51:210-28.
88. Sarbin, T. R., & R. Taft. (1952) *An essay on inference in the psychological sciences.* Berkeley, CA: Garden Library.
89. Sargent, H. (1945) Projective methods: Their origins, theory, and application in personality research. *Psychological Bulletin,* 42:257-93.
90. Schiedt, R. (1936) *Ein Beitrag zum Problem der Rückfallsprognose.* Ph.D. thesis. Munich: Münchner-Zeitungs-Verlag.
91. Schneider, A. J. N., C. W. Lagrone, E. T. Glueck, & S. Glueck. (1944) Prediction of behavior of civilian delinquents in the Armed Forces. *Mental Hygiene,* 28:456-75.
92. Sorokin, P. A. (1942) A criticism of "*The prediction of personal adjustment.*" *American Journal of Sociology,* 48:76-80.
93. Spence, K. W. (1944) The nature of theory construction in contemporary psychology. *Psychological Review,* 51:47-68.
94. Spence, K. W. (1948) The postulates and methods of behaviorism. *Psychological Review,* 55:67-78.
95. Stouffer, S. A. (1941) Notes on the case study and the unique case. *Sociometry,* 4:349-57.
96. Super, D. E. (1949) *Appraising vocational fitness.* New York: Harper.
97. Taft, Ronald. (1950) Some correlates of the ability to make accurate social judgments. Unpublished Ph.D. thesis, University of California.
98. Taylor, D. W. (1947) An analysis of predictions of delinquency based on case studies. *Journal of Abnormal and Social Psychology,* 43:45-56.
99. Thorndike, E. L. (1918) Fundamental theorems in judging men. *Journal of Applied Psychology.* 2:67-76.
100. Thurstone, L. L. (1947) *Multiple factor analysis.* Chicago: University of Chicago Press.
101. Viteles, M. S. (1925) The clinical viewpoint in vocational selection. *Journal of Applied Psychology,* 9:131-38.
102. Vold, G. B. (1941) Comment on crucial problems in methods of predicting social adjustment. *Sociometry,* 4:374-78.
103. Wittman, M. P. (1941) A scale for measuring prognosis in schizophrenic patients. *Elgin Papers,* 4:20-33.
104. Wittman, M. P., & L. Steinberg. (1944) Follow-up of an objective evaluation of prognosis in dementia praecox and manic-depressive psychoses. *Elgin Papers,* 5:216-27

Index

pdf by LJY, December 2004