American Psychologist, 1956, *11*, 263–272.

#039

Wanted—A Good Cookbook

Paul E. Meehl [1]

Once upon a time there was a young fellow who, as we say, was "vocationally maladjusted." He wasn't sure just what the trouble was, but he knew that he was not happy in his work. So, being a denizen of an urban, sophisticated, psychologically oriented culture, he concluded that what he needed was some professional guidance. He went to the counseling bureau of a large midwestern university (according to some versions of the tale, it was located on the banks of a great river), and there he was interviewed by a world-famous vocational psychologist. When the psychologist explained that it would first be necessary to take a 14-hour battery of tests, the young man hesitated a little; after all, he was still employed at his job and 14 hours seemed like quite a lot of time. "Oh, well," said the great psychologist reassuringly, "don't worry about *that.* If you're too busy, you can arrange to have my assistant take these tests *for* you. I don't care who takes them, just so long as they come out in quantitative form.

Lest I, a Minnesotan, do too great violence to your expectations by telling this story on the dust-bowl empiricism with which we Minnesotans are traditionally associated, let me now tell you a true story having the opposite animus. Back in the days when we were teaching assistants, my colleague MacCorquodale was grading a young lady's elementary laboratory report on an experiment which involved a correlation problem. At the end of an otherwise flawless report, this particular bobbysoxer had written "The correlation was seventy-five, with a standard error of ten, which is significant. However, I do not think these variables are related." MacCorquodale wrote a large red "FAIL" and added a note: "Dear Miss Fisbee: The correlation coefficient was devised expressly to relieve you of all responsibility for deciding whether these two variables are related."

If you find one of these anecdotes quite funny, and the other one rather stupid (I don't care which), you are probably suffering from a slight case of bias. Although I have not done a factor analysis with these two stories in the matrix, my clinical judgment tells me that a person's spontaneous reactions to them reflect his position in the perennial conflict between the toughminded and the tenderminded, between those for whom the proper prefix to the word "analysis" is "factor" and those for whom it is "psycho," between the groups that Lord Russell once characterized as the "simpleminded" and the "muddleheaded." In a recent book (Meehl, 1954a/1996), I have explored one major facet of this conflict, namely the controversy over the relative merits of clinical and statistical methods of *prediction.* Theoretical considerations, together with introspections as to my own mental activities as a psychotherapist, led me to conclude that the clinician has certain unique, practically unduplicable powers by virtue of being himself an organism like his client; but that the domain of straight *prediction* would not be a favorable locus for displaying these powers. Survey of a score of empirical investigations in which the actual predictive efficiency of the two

---

[1] Presidential Address, Midwestern Psychological Association, Chicago, April 29. 1955.
Reprinted 1973, with the title "Problems in the actuarial characterization of a person," in H. Feigl and M. Scriven (Eds.), *Minnesota Studies in the Philosophy of Science*, Vol. 1: *The foundations of science and the concepts of psychology and psychoanalysis* (pp. 205-222), Minneapolis: University of Minnesota Press.

methods could be compared, gave strong confirmation to this latter theoretical expectation. After reading these studies, it almost looks as if the first rule to follow in trying to predict the subsequent course of a student's or patient's behavior is carefully to avoid talking to him, and that the second rule is to avoid thinking about him!

Statisticians (and rat men) with castrative intent toward clinicians should beware of any temptation to overextend these findings to a generalization that "clinicians don't actually add anything." Apart from the clinician's therapeutic efforts—the power of which is a separate issue and also a matter of current dispute—a glance at a sample of clinical diagnostic documents, such as routine psychological reports submitted in a VA installation, shows that a kind of mixed predictive-descriptive statement predominates which is different from the type of gross prediction considered in the aforementioned survey. (I hesitate to propose a basic distinction here, having learned that proposing a distinction between two classes of concepts is a sure road to infamy.) Nevertheless, I suggest that we distinguish between: (*a*) the clinician's predictions of such gross, outcome-type, "administrative" dimensions as recovery from psychosis, survival in a training program, persistence in therapy, and the like; and (*b*) a rather more detailed and ambitious enterprise roughly characterizable as "describing the person." It might be thought that *a* always presupposes *b*, but a moment's reflection shows this to be false; since there are empirical prediction systems in which the sole property ascribed to the person *is* the disposition to a predicted gross outcome. A very considerable fraction of the typical clinical psychologist's time seems to be spent in giving tests or semitests, the intention being to come out with some kind of characterization of the individual. In part this characterization is "phenotypic," attributing such behavior-dispositions as "hostile," "relates poorly," "loss in efficiency," "manifest anxiety," or "depression"; in part it is "genotypic," inferring as the causes of the phenotype certain inner events, states, or structures, such as, "latent *n Aggression*," "oral-dependent attitudes," "severe castration anxiety," and the like. While the phenotypic-genotypic question is itself deserving of careful methodological analysis, in what follows I shall use the term "personality description" to cover both phenotypic and genotypic inferences, that is, statements of all degrees of internality or theoreticalness. I shall also assume, while recognizing that at least one group of psychologists has made an impressive case to the contrary, that the description of a person is a worthwhile stage in the total clinical process. Granted, then, that we wish to use tests as a means to securing a description of the person, how shall we go about it? Here we sit, with our Rorschach and Multiphasic results spread out before us. From this mess of data we have to emerge with a characterization of the person from whose behavior these profiles are a highly abstracted, much-reduced distillation. How to proceed?

Some of you are no doubt wondering, "What is the fellow talking about? You look at the profiles, you call to mind what the various test dimensions mean for dynamics, you reflect on other patients you have seen with similar patterns, you think of the research literature; then you combine these considerations to make inferences. Where's the problem?" The problem is, *whether or not this is the most efficient way to do it*. We ordinarily do it this way; in fact, the practice is so universal that most clinicians find it shocking, if not somehow sinful, to imagine any other. We feed in the test data and let that rusty digital computer in our heads go to work until a paragraph of personality description emerges. It requires no systematic study, although some quantitative data have begun to appear in the literature (Dailey, 1953; Davenport, 1952; Holsopple & Phelan, 1954;

Kostlan, 1954; Little & Schneidman, 1954, 1955), to realize that there is a considerable element of vagueness, hit-or-miss, and personal judgment involved in this approach. Because explicit rules are largely lacking, and hence the clinician's personal experience, skill, and creative artistry play so great a role, I shall refer to this time-honored procedure for generating personality descriptions from tests as the *rule-of-thumb* method.

I wish now to contrast this rule-of-thumb method with what I shall call the *cookbook method.* In the cookbook method, any given configuration (holists please note—I said "configuration," not "sum"!) of psychometric data is associated with each facet (or configuration) of a personality description, and the closeness of this association is explicitly indicated by a number. This number need not be a correlation coefficient—its form will depend upon what is most appropriate to the circumstances. It may be a correlation, or merely an ordinary probability of attribution, or (as in the empirical study I shall report upon later) an average Q-sort placement. Whatever its form, the essential point is that the transition from psychometric pattern to personality description is an automatic, mechanical, "clerical" kind of task, proceeding by the use of explicit rules set forth in the cookbook. I am quite aware that the mere prospect of such a method will horrify some of you; in my weaker moments it horrifies me. All I can say is that many clinicians are also horrified by the cookbook method as applied in the crude prediction situation; whereas the studies reported to date indicate this horror to be quite groundless (Meehl, 1954a/1996, Chap. 8). As Fred Skinner once said, some men are less curious about nature than about the accuracy of their guesses (1938, p. 44). Our responsibility to our patients and to the taxpayer obliges us to decide between the rule-of-thumb and the cookbook methods on the basis of their empirically demonstrated efficiency, rather than upon which one is more exciting, more "dynamic," more like what psychiatrists do, or more harmonious with the clinical psychologist's self concept.

Let us sneak up the clinician's avoidance gradient gradually to prevent the negative therapeutic reaction. Consider a particular complex attribute, say, "strong dependency with reaction-formation." Under what conditions should we take time to give a test of moderate validity as a basis for inferring the presence or absence of this complex attribute? Putting it negatively, it appears to me pretty obvious that there are two circumstances under which we should *not* spend much skilled time on testing even with a moderately valid test, because we stand to lose if we let the test finding influence our judgments. First, when the attribute is found in almost all our patients; and second, when it is found in almost none of our patients. (A third situation, which I shall not consider here, is one in which the attribute makes no practical difference anyhow.) A disturbingly large fraction of the assertions made in routine psychometric reports or uttered by psychologists in staff conferences fall in one of these classes.

It is not difficult to show that when a given personality attribute is almost always or almost never present in a specified clinical population, rather severe demands are made upon the test's validity if it is to contribute in a practical way to our clinical decision-making. A few simple manipulations of Bayes' Rule for calculating inverse probability lead to rather surprising, and depressing, results. Let me run through some of these briefly. In what follows,

P = Incidence of a certain personality characteristic in a specified clinical population. ($Q = 1 - P$, $P > Q$)

p_1 = Proportion of "valid positives," i.e., incidence of positive test finding among cases who actually have the characteristic. ($q_1 = 1 - p_1$)

p_2 = Proportion of "false positives," i.e., incidence of positive test findings among cases who actually lack the characteristic. ($q_2 = 1 - p_2$)

1. When is a positive assertion (attribution of the characteristic) on the basis of a positive test finding more likely to be correct than incorrect?

$$\frac{P}{Q} > \frac{p_2}{p_1}.$$

Example: A test correctly identifies 80 percent of brain-damaged patients at the expense of only 15 percent false positives, in a neuropsychiatric population where one-tenth of all patients are damaged. The decision "brain damage present" on the basis of a positive test finding is more likely to be false than true, since the inequality is unsatisfied.

2. When does the use of a test improve over-all decision making?

$$P < \frac{q_2}{q_1 + q_2}.$$

If $P < Q$ this has the form $Q < \frac{p_1}{p_1 + p_2}$.

Example: A test sign identifies 85 percent of "psychotics" at the expense of only 15 percent of false positives among the "nonpsychotic." It is desired to make a decision on each case, and both kinds of errors are serious.[2] Only 10 percent of the population seen in the given setting are psychotic. Hence, the use of the test yields more erroneous classifications than would proceeding without the test.

3. When does improving a sign, strengthening a scale, or shifting a cut improve decision making?

$$\frac{\Delta p_1}{\Delta p_2} > \frac{Q}{P}.$$

Example: We improve the intrinsic validity of a "schizophrenic index" so that it now detects 20 percent more schizophrenics than it formerly did, at the expense of only a 5 percent rise in the false positive rate. This surely looks encouraging. However, we work with an outpatient clientele only one-tenth of whom are actually schizophrenic. Since these values violate the inequality, "improvement" of the index will result in an increase in the proportion of erroneous diagnoses. N.B.—*Sampling errors are not involved in the above.* The values are assumed to be parameter values, and the test sign is valid (i.e., $p_1 > p_2$ in the population).

Further inequalities and a more detailed drawing out of their pragmatic implications can be found in a recent paper by Albert Rosen and myself (1955). The moral to be drawn from these considerations, which even we clinicians can follow because they involve only high-school algebra, is that a great deal of skilled psychological effort is probably being wasted in going through complex, skill-demanding, time-consuming test procedures of moderate or low validity, in order to arrive at conclusions about the patient which could often be made with high confidence without the test, and which in other cases ought not to

[2] Inequalities (2) and (3) are conditions for improvement if there is no reason to see one kind of error as worse than the other. In trait attribution this is usually true; in prognostic and diagnostic decisions it may or may not be. If one is willing to say how many errors of one kind he is prepared to tolerate in order to avoid one of the other kind, these inequalities can be readily corrected by inserting this ratio. A more general development can be found in an unpublished paper by Ward Edwards [1954].

be made (because they still tend to be wrong) even with the test indications positive. Probably most surprising is the finding that there are certain quantitative relations between the base rates and test validity parameters such that the use of a "valid" test will produce a net rise in the frequency of clinical mistakes. The first task of a good clinical cookbook would be to make explicit quantitative use of the inverse probability formulas in constructing efficient "rules of attribution" when test data are to be used in describing the personalities of patients found in various clinical populations. For example, I know of an out-patient clinic which has treated, by a variety of psychotherapies, in the course of the past eight years, approximately 5000 patients, not one of whom has committed suicide. If the clinical psychologists in this clinic have been spending much of their time scoring suicide keys on the Multiphasic or counting suicide indicators in Rorschach content, either these test indicators are close to infallible (which is absurd), or else the base rate is so close to zero that the expenditure of skilled time is of doubtful value. Suicide is an extreme case, of course (Rosen, 1954); but the point so dramatically reflected there is valid, with suitable quantitative modifications, over a wider range of base rates. To take some examples from the high end of the base-rate continuum, it is not very illuminating to say of a known psychiatric patient that he has difficulty in accepting his drives, experiences some trouble in relating emotionally to others, and may have problems with his sexuality! Many psychometric reports bear a disconcerting resemblance to what my colleague Donald G. Paterson calls "personality description after the manner of P. T. Barnum" (see Blum & Balinsky, 1951, p. 47; Dunnette, 1957, p. 223). I suggest—and I am quite serious—that we adopt the phrase *Barnum effect* to stigmatize those pseudo-successful clinical procedures in which personality descriptions from tests are made to fit the patient largely or wholly by virtue of their triviality; and in which any nontrivial, but perhaps erroneous, inferences are hidden in a context of assertions or denials which carry high confidence simply because of the population base rates, regardless of the test's validity. I think this fallacy is at least as important and frequent as others for which we have familiar labels (halo effect, leniency error, contamination, etc.). One of the best ways to increase the general sensitivity to such fallacies is to give them a name. We ought to make our clinical students as acutely aware of the Barnum effect as they are of the dangers of countertransference or the standard error of *r*.

The preceding mathematical considerations, while they should serve as a check upon some widespread contemporary forms of tea-leaf reading, are unfortunately not very "positive" by way of writing a good cookbook. "Almost anything needs a little salt for flavor" or "It is rarely appropriate to put ketchup on the dessert" would be sound advice but largely negative and not very helpful to an average cook. I wish now to describe briefly a piece of empirical research, reported in a thesis just completed at Minnesota by Charles C. Halbower, which takes the cookbook method 100 percent seriously; and which seems to show, at least in one clinical context, what can be done in a more constructive way by means of a cookbook of even moderate trustworthiness.[3] By some geographical coincidence, the psychometric device used in this research was a structured test consisting of a set of 550 items, commonly known as MMPI. Let me emphasize that the MMPI is not here being compared with anything else, and that the research does not aim to investigate Multiphasic validity (although the general order of magnitude of the obtained correlations

[3] I am indebted to Dr. Halbower for permission to present this summary of his thesis data in advance of his own more complete publication [Halbower, 1955].

does give some incidental information in that respect). What Dr. Halbower asked was this: given a Multiphasic profile, how does one arrive at a personality description from it? Using the rule-of-thumb method, a clinician familiar with MMPI interpretation looks at the profile, thinks awhile, and proceeds to describe the patient he imagines would have produced such a pattern. Using the cookbook method, we don't need a clinician; instead, a $230-per-month clerk-typist in the outer office simply reads the numbers on the profile, enters the cookbook, locates the page on which is found some kind of "modal description" for patients with such a profile, and this description is then taken as the best available approximation to the patient. We know, of course, that every patient is unique—absolutely, unqualifiedly unique. Therefore, the application of a cookbook description will inevitably make errors, some of them perhaps serious ones. If we knew *which* facets of the cookbook sketch needed modification as applied to the present unique patient, we would, of course, depart from the cookbook at these points; but we don't know this. If we start monkeying with the cookbook recipe in the hope of avoiding or reducing these errors, we will in all likelihood improve on the cookbook in some respects but, unfortunately, will worsen our approximation in others. Given a finite body of information, such as the 13 two-digit numbers of a Multiphasic profile, there is obviously *in fact* (whether we have yet succeeded in *finding* it or not) a "most probable" value for any personality facet, and also for any configuration of facets, however complex or "patterned" (Meehl, 1954a/1996, pp. 131-134). It is easy to prove that a method of characterization which departs from consistent adherence to this "best guess" stands to lose. Keep in mind, then, that the raw data from which a personality description was to be inferred consisted of an MMPI profile. In other words, the Halbower study was essentially a comparison of the rule-of-thumb versus the cookbook method where each method was, however, functioning upon the same information—an MMPI. We are in effect contrasting the validity of two methods of "reading" Multiphasics.

In order to standardize the domain to be covered, and to yield a reasonably sensitive quantification of the goodness of description, Dr. Halbower utilized Q sorts. From a variety of sources he constructed a Q pool of 154 items, the majority being phenotypic or intermediate and a minority being genotypic. Since these items were intended for clinically expert sorters employing an "external" frame of reference, many of them were in technical language. Some sample items from his pool are: "Reacts against his dependency needs with hostility"; "manifests reality distortions"; "takes a dominant, ascendant role in interactions with others"; "is rebellious toward authority figures, rules, and other constraints"; "is counteractive in the face of frustration"; "gets appreciable secondary gain from his symptoms"; "is experiencing pain"; "is naive"; "is impunitive"; "utilizes intellectualization as a defense mechanism"; "shows evidence of latent hostility"; "manifests inappropriate affect." The first step was to construct a cookbook based upon these 154 items as the ingredients; the recipes were to be in the form of directions as to the optimal Q-sort placement of each item.

How many distinguishable recipes will the cookbook contain? If we had infallible criterion Q sorts on millions of cases, there would be as many recipes as there are possible MMPI profiles. Since we don't have this ideal situation, and never will, we have to compromise by introducing coarser grouping. Fortunately, we know that the validity of our test is poor enough so that this coarseness will not result in the sacrifice of much, if any, information. How coarsely we group, that is, how different two Multiphasic curves have to be before we refuse to call them "similar" enough to be coordinated with the same recipe, is a

very complicated matter involving both theoretical and practical considerations. Operating within the limits of a doctoral dissertation, Halbower confined his study to four profile "types." These curve types were specified by the first two digits of the Hathaway code plus certain additional requirements based upon clinical experience. The four MMPI codes used were those beginning 123′, 13′, 27′, and 87′ (Hathaway, 1947). The first three of these codes are the most frequently occurring in the Minneapolis VA Mental Hygiene Clinic population, and the fourth code, which is actually fifth in frequency of occurrence, was chosen in order to have a quasi-psychotic type in the study. It is worth noting that these four codes constitute 58 percent of all MMPI curves seen in the given population; so that Halbower's gross recipe categories already cover the majority of such outpatients. The nature of the further stipulations, refining the curve criteria within each two-digit code class, is illustrated by the following specifications for code 13′, the "hysteroid valley" or "conversion V" type:

1. *Hs* and *Hy* ≥ 70.
2. *D* < (*Hs* and *Hy*) by at least one sigma.
3. *K* or *L* > *?* and *F*.
4. *F* ≤ 65.
5. Scales 4,5,6,7,8,9,0 all ≤ 70.

For each of these MMPI curve types, the names of nine patients were then randomly chosen from the list of those meeting the curve specifications. If the patient was still in therapy, his therapist was asked to do a Q sort (eleven steps, normal distribution) on him. The MMPI had been withheld from these therapists. If the patient had been terminated, a clinician (other than Halbower) did a Q sort based upon study of the case folder, including therapist's notes and any available psychometrics (except, of course, the Multiphasic). This yields Q sorts for nine patients of a given curve type. These nine sorts were then pairwise intercorrelated, and by inspection of the resulting 36 coefficients, a subset of five patients was chosen as most representative of the curve type. The Q sorts on these five "representative" patients were then averaged, and this average Q sort was taken as the cookbook recipe to be used in describing future cases having the given MMPI curve. Thus, this modal, crystallized, "distilled-essence" personality description was obtained by eliminating patients with atypical sortings and pooling sortings on the more typical, hoping to reduce both errors of patient sampling and of clinical judgment. This rather complicated sequence of procedures may be summarized thus:

Deriving cookbook recipe for a specified curve type, such as the "conversion V" above:

1. Sample of *N* = nine patients currently or recently in therapy and meeting the MMPI specifications for conversion V curve.
2. 154-item Q sort done on each patient by therapist or from therapist notes and case folder. (These sorts MMPI-uncontaminated.)
3. Pairwise Q correlations of these nine patients yields 36 intercorrelations.
4. Selection of subset *N*′ = five "modal" patients from this matrix by inspectional cluster method.
5. Mean of Q sorts on these five "core" patients is the cookbook recipe for the MMPI curve type in question.

Having constructed one recipe, he started all over again with a random sample of nine patients whose Multiphasics met the second curve-type specifications, and carried out these cluster-and-pooling processes upon them. This was done for each of the four curve types which were to compose the cookbook. If you have reservations about any of the steps in constructing this miniature cookbook, let me remind you that this is all preliminary, that is, *it is the means of arriving at the cookbook recipe*. The proof of the pudding will be in the eating, and any poor choices of tactics or patients up to this point should merely make the cookbook less trustworthy than it would otherwise be.

Having thus written a miniature cookbook consisting of only four recipes, Halbower then proceeded to cook some dishes to see how they would taste. For cross validation he chose at random four new Mental Hygiene Clinic patients meeting the four curve specifications and who had been seen in therapy for a minimum of ten hours. With an eye to validity generalization to a somewhat different clinical population, with different base rates, he also chose four patients who were being seen as inpatients at the Minneapolis VA Hospital. None of the therapists involved had knowledge of the patients' Multiphasics. For purposes of his study, Halbower took the therapist's Q sort, based upon all of the case folder data (minus MMPI) plus his therapeutic contacts, as the best available criterion; although this "criterion" is acceptable only in the sense of construct validity (Cronbach & Meehl, 1955). An estimate of its absolute level of trustworthiness is not important since it is being used as the common reference basis for a comparison of two methods of test reading.

Given the eight criterion therapist Q sorts (2 patients for each MMPI curve type), the task of the cookbook is to predict these descriptions. Thus, for each of the two patients having MMPI code 123′, we simply assign the Q-sort recipe found in the cookbook as the best available description. How accurate this description is can be estimated (in the sense of construct validity) by Q correlating it with the criterion therapist's description. These eight "validity" coefficients varied from .36 to .88 with a median of .69. As would be expected, the hospital inpatients yielded the lower correlations. The Mental Hygiene Clinic cases, for whom the cookbook was really intended, gave validities of .68, .69, .84, and .88 (see Table 1).

How does the rule-of-thumb method show up in competition with the cookbook? Here we run into the problem of differences in clinical skill, so Halbower had each MMPI profile read blind by more than one clinician. The task was to interpret the profile by doing a Q sort. From two to five clinicians thus "read" each of the eight individual profiles, and the resulting 25 sorts were Q correlated with the appropriate therapist criterion sorts. These validity coefficients run from .29 to .63 with a median of .46. The clinicians were all Minnesota trained and varied in their experience with MMPI from less than a year (first-year VA trainees) through all training levels to PhD staff psychologists with six years' experience. The more experienced clinicians had probably seen over two thousand MMPI profiles in relation to varying amounts of other clinical data, including intensive psychotherapy. Yet not one of the 25 rule-of-thumb readings was as valid as the cookbook reading. Of the 25 comparisons which can be made between the validity of a single clinician's rule-of-thumb reading and that of the corresponding cookbook reading of the same patient's profile, eighteen are significant in favor of the cookbook at the .01 level of confidence and four at the .05 level. The remaining three are also in favor of the cookbook but not significantly so.

Confining our attention to the more appropriate outpatient population, for (and upon) which the cookbook was developed, the mean *r* (estimated through *z* transformation) is .78 for the cookbook method, as contrasted with a mean (for seventeen rule-of-thumb descriptions) of only .48, a difference of 30 points of correlation, which in this region amounts to a difference of 38 percent in the validly predicted variance! The cookbook seems to be superior to the rule-of-thumb not merely in the sense of statistical significance but by an amount which is of very practical importance. It is also remarkable that even when the cookbook recipes are applied to patients from a quite different kind of population, their validity still excels that of rule-of-thumb MMPI readers who are in daily clinical contact with that other population. The improvement in valid variance in the hospital sample averages 19 percent (see item 5 in Table 1).

TABLE 1

Validation of the Four Cookbook Descriptions on New Cases and Comparative Validities of the Cookbook MMPI Readings and Rule-of-Thumb Readings by Clinicians

1. Four patients currently in therapy Q-described by the therapist (10 hours or more therapy plus case folder minus MMPI). This is taken as best available criterion description of each patient.
2. MMPI cookbook recipe Q-correlated with this criterion description.
3. For each patient, 4 or 5 clinicians "read" his MMPI in usual rule-of-thumb way, doing Q-sorts.
4. These rule-of-thumb Q-sorts also Q-correlated with criterion description.
5. Cross-validation results in outpatient sample.

Validities	MMPI Curve Type			
	Code 123′	Code 27′	Code 13′	Code 87′
Cookbook	.88	.69	.84	.68
Rule-of-thumb (mean)	.75	.50	.50	.58
Range (4–5 readers)	.55–.63	.29–.54	.37–.52	.34–.58

Mean of 4 cookbook validities, through z_r =.78
Mean of 17 rule-of-thumb validities, through z_r =.48
Cookbook's superiority in validly predicted variance = 38%

6. Validity generalization to inpatients (psychiatric hospital) sample with different base rates; hence, an "unfair" test of cookbook.

Validities	MMPI Curve Type			
	Code 123′	Code 27′	Code 13′	Code 87′
Cookbook	.63	.29	.30	.50
Rule-of-thumb (2 readers)	.37, .49	.29, .42	.30, .30	.50, .50

Mean of 4 cookbook validities, through z_r = .60
Mean of 8 rule-of-thumb validities, through z_r = .41
Cookbook's superiority in validly predicted variance = 19%

A shrewd critic may be thinking, "Perhaps this is because all kinds of psychiatric patients are more or less alike, and the cookbook has simply taken advantage of this rather trivial fact." In answer to this objection, let me say first that to the extent the cookbook's

superiority did arise from its actuarially determined tendency to "follow the base rates," that would be a perfectly sound application of the inverse probability considerations I at first advanced. For example, most psychiatric patients are in some degree depressed. Let us suppose the mean Q-sort placement given by therapists to the item "depressed" is seven. "Hysteroid" patients, who characteristically exhibit the so-called "conversion V" on their MMPI profiles (Halbower's cookbook code 13), are less depressed than most neurotics. The clinician, seeing such a conversion valley on the Multiphasic, takes this relation into account by attributing "lack of depression" to the patient. But maybe he over-interprets, giving undue weight to the psychometric finding and understressing the base rate. So his rule-of-thumb placement is far down at the nondepressed end, say at position three. The cookbook, on the other hand, "knows" (actuarially) that the mean Q placement for the item "depressed" is at five in patients with such profiles—lower than the over-all mean seven but not displaced as much in the conversion subgroup as the clinician thinks. If patients are so homogeneous with respect to a certain characteristic that the psychometrics ought not to influence greatly our attribution or placement in defiance of the over-all actuarial trend, then the clinician's tendency to be unduly influenced is a source of erroneous clinical decisions and a valid argument in favor of the cookbook.

TABLE 2

Validation of the Four Clinicians' Description of "Average Patient," of the Mean of These Stereotypes and of the Cookbook Recipe (Outpatient Cases Only)

MMPI Curve Type	Validities of "Average Patient" Descriptions by 4 Clinicians	Validity of Mean of These 4 "Average Patient" Stereotypes	Validity of Cookbook Recipe
Code 123′	.63 to .69	.74	.88
Code 27′	–.03 to .20	.09	.69
Code 13′	.25 to .37	.32	.84
Code 87′	.25 to .35	.31	.68

However, if this were the chief explanation of Halbower's findings, the obvious conclusion would be merely that MMPI was not differentiating, since any test-induced departure from a description of the "average patient" would tend to be more wrong than right. Our original question would then be rephrased, "What is the comparative efficiency of the cookbook and the rule-of-thumb method *when each is applied to psychometric information having some degree of intrinsic validity*?" Time permits me only brief mention of the several lines of evidence in Halbower's study which eliminate the Barnum effect as an explanation. First of all, Halbower had selected his 154 items from a much larger initial Q pool by a preliminary study of therapist sortings on a heterogeneous sample of patients in which items were eliminated if they showed low interpatient dispersal. Second, study of the placements given an item over the four cookbook recipes reveals little similarity (e.g., only two items recur in the top quartile of all four recipes; 60 percent of the items occur in the top quartile of only one recipe). Third, several additional correlational findings combine to show that the cookbook was not succeeding merely by describing an "average patient" four times over. For example, the clinicians' Q description of their conception of the "average

patient" gave very low validity for three of the four codes, and a "mean average patient" description constructed by pooling these clinicians' stereotypes was not much better (see Table 2). For Code 123′ (interestingly enough, the commonest code among therapy cases in this clinic) the pooled stereotype was actually more valid than rule-of-thumb Multiphasic readings. (This is Bayes' Theorem with a vengeance!) Nevertheless, I am happy to report that this "average patient" description was still inferior to the Multiphasic cookbook (significant at the .001 level).

Let me ruminate about the implications of this study, supposing it should prove to be essentially generalizable to other populations and to other psychometric instruments. From a theoretical point of view, the trend is hardly surprising. It amounts to the obvious fact that the human brain is an inefficient recording and computing device. The cookbook method has an advantage over the rule-of-thumb method because it (*a*) samples more representatively, (*b*) records and stores information better, and (*c*) computes statistical weights which are closer to the optimal. We can perhaps learn more by putting the theoretical question negatively: when should we *expect* the cookbook to be inferior to the brain? The answer to this question presumably lies in the highly technical field of computing machine theory, which I am not competent to discuss. As I understand it, the use of these machines requires that certain rules of data combination be fed initially into the machine, followed by the insertion of suitably selected and coded information. Putting it crudely, the machine can "remember" and can "think routinely," but it cannot "spontaneously notice what is relevant" nor can it "think" in the more high-powered, creative sense (e.g., it cannot invent theories). To be sure, noticing what is relevant must involve the exemplification of some rule, perhaps of a very complex form. But it is a truism of behavior science that organisms can *exemplify* rules without *formulating* them. To take a noncontroversial example outside the clinical field, no one today knows how to state fully the rules of "similarity" or "stimulus equivalence" for patterned visual perception or verbal generalization; but of course we all exemplify daily these undiscovered rules. This suggests that as long as psychology cannot give a complete, explicit, quantitative account of the "dimensions of relevance" in behavior connections, the cookbook will not completely duplicate the clinician (Meehl, 1954b). The clinician *here* acts as an inefficient computer, but that is better than a computer with certain major rules completely left out (because we can't build them in until we have learned how to formulate them). The use of the therapist's own unconscious in perceiving verbal and imaginal relations during dream interpretation is, I think, the clearest example of this. But I believe the exemplification of currently unformulable rules is a widespread phenomenon in most clinical inference. However, you will note that these considerations apply chiefly (if not wholly) to matters of *content*, in which a rich, highly varied, hard-to-classify content (such as free associations) is the input information. The problem of "stimulus equivalence" or "noticing the relevant" does not arise when the input data are in the form of preclassified responses, such as a Multiphasic profile or a Rorschach psychogram. I have elsewhere (1954a/1996, pp. 110-111) suggested that even in the case of such prequantified patterns there arises the possibility of causal-theory-mediated idiographic extrapolations into regions of the profile space in which we lack adequate statistical experience; but I am now inclined to view that suggestion as a mistake. The underlying theory must itself involve some hypothesized function, however crudely quantified; otherwise, how is the alleged "extrapolation" possible? I can think of no reason why the estimation of the parameters in this underlying theoretical function should constitute an exception to the cookbook's

superiority. If I am right in this, my "extrapolation" argument applies strictly only when a clinician literally *invents new theoretical relations or variables* in thinking about the individual patient. In spite of some clinicians' claims along this line, I must say I think it very rarely happens in daily clinical practice. Furthermore, even when it does happen, Bayes' Rule still applies. The *joint* probability of the theory's correctness, and of the attribute's presence (granting the theory but remembering nuisance variables) must be high enough to satisfy the inequalities I have presented, otherwise use of the theory will not pay off.

What are the pragmatic implications of the preceding analysis? Putting it bluntly, it suggests that for a rather wide range of clinical problems involving personality description from tests, the clinical interpreter is a costly middleman who might better be eliminated. An initial layout of research time could result in a cookbook whose recipes would encompass the great majority of psychometric configurations seen in daily work. I am fully aware that the prospect of a "clinical clerk" simply looking up Rorschach pattern number 73 J 10-5 or Multiphasic curve "Halbower Verzeichnis 626" seems very odd and even dangerous. I reassure myself by recalling that the number of phenotypic and genotypic attributes is, after all, finite; and that the number which are ordinarily found attributed or denied even in an extensive sample of psychological reports on patients is actually very limited. A best estimate of a Q-sort placement is surely more informative than a crude "Yes-or-No" decision of low objective confidence. I honestly cannot see, in the case of a *determinate trait domain* and a *specified clinical population*, that there is a serious intellectual problem underlying one's uneasiness. I invite you to consider the possibility that the emotional block we all experience in connection with the cookbook approach could be dissolved simply by trying it out until our daily successes finally get us accustomed to the idea.

Admittedly this would take some of the "fun" out of psychodiagnostic activity. But I suspect that most of the clinicians who put a high value on this kind of fun would have even more fun doing intensive psychotherapy. The great personnel needs today, and for the next generation or more, are for psychotherapists and researchers. (If you don't believe much in the efficacy of therapy, this is the more reason for research.) If all the thousands of clinical hours currently being expended in concocting clever and flowery personality sketches from test data could be devoted instead to scientific investigation (assuming we are still selecting and training clinicians to be scientists), it would probably mean a marked improvement in our net social contribution. If a reasonably good cookbook could help bring about this result, the achievement would repay tenfold the expensive and tedious effort required in its construction.

REFERENCES

Blum, M.L., & Balinsky, B. (1951). *Counseling and psychology*. New York: Prentice-Hall.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Dailey, C.A. (1953). The practical utility of the clinical report. *Journal of Consulting Psychology*, *17*, 297-302.

Davenport, B.F. (1952). The semantic validity of TAT interpretations. *Journal of Consulting Psychology*, *16*, 171-175.

Dunnette, M.D. (1957). Use of the sugar pill by industrial psychologists. *American Psychologist*, *12*, 223-225.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*, 380-417.

Halbower, C.C. (1955). *A comparison of actuarial versus clinical prediction to classes discriminated by MMPI*. Unpublished Ph.D. thesis, University of Minnesota.

Hathaway, S.R. (1947). A coding system for MMPI profiles. *Journal of Consulting Psychology*, *11*, 334-337.

Holsopple, J.Q., & Phelan, J. G. (1954). The skills of clinicians in analysis of projective tests. *Journal of Clinical Psychology*, *10*, 307-320.

Kostlan, A. (1954). A method for the empirical study of psychodiagnosis, *Journal of Consulting Psychology*, *18*, 83-88.

Little, K.B., & Shneidman, E.S. (1954). The validity of MMPI interpretations. *Journal of Consulting Psychology*, *18*, 425-428.

Little, K.B., & Shneidman, E.S. (1955). The validity of thematic projective technique interpretations. *Journal of Personality*, *23*, 285-294.

Meehl, P.E. (1954a/1996). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press. Reprinted with new Preface, 1996, by Jason Aronson, Northvale, NJ.

Meehl, P.E. (1954b). Comment [on C. McArthur "Analyzing the clinical process"]. *Journal of Counseling Psychology*, *1*, 203-208.

Meehl, P.E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194-216.

Paterson, D.G. (Unpublished, mimeographed). Character reading at sight of Mr. X according to the system of Mr. P.T. Barnum. First printed in Blum & Balinsky (1951); reprinted in Dunnette (1957).

Rosen, A. (1954). Detection of suicidal patients: an example of some limitations in the prediction of infrequent events. *Journal of Consulting Psychology*, *18*, 397-403.

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

pdf by LJY, November 2019