Psychology, Public Policy, and Law
1996, *2*, 293–323

#167

Comparative Efficiency of Informal (Subjective, Impressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical–Statistical Controversy

William M. Grove and Paul E. Meehl
University of Minnesota, Twin Cities Campus

Given a data set about an individual or group (e.g., interviewer ratings, life history or demographic facts, test results, self-descriptions), there are two modes of data combination for a predictive or diagnostic purpose. The clinical method relies on human judgment that is based on informal contemplation and, sometimes, discussion with others (e.g., case conferences). The mechanical method involves a formal, algorithmic, objective procedure (e.g., equation) to reach the decision. Empirical comparisons of the accuracy of the two methods (136 studies over a wide range of predictands) show that the mechanical method is almost invariably equal to or superior to the clinical method: Common antiactuarial arguments are rebutted, possible causes of widespread resistance to the comparative research are offered, and policy implications of the statistical method's superiority are discussed.

In 1928, the Illinois State Board of Parole published a study by sociologist Burgess of the parole outcome for 3,000 criminal offenders, an exhaustive sample of parolees in a period of years preceding. (In Meehl 1954/1996, this number is erroneously reported as 1,000, a slip probably arising from the fact that 1,000 cases came from each of three Illinois prisons.) Burgess combined 21 objective factors (e.g., nature of crime, nature of sentence, chronological age, number of previous offenses) in unweighted fashion by simply counting for each case the number of factors present that expert opinion considered favorable or unfavorable to successful parole outcome. Given such a large sample, the predetermination of a list of relevant factors (rather than elimination and selection of factors), and the absence of any attempt at optimizing weights, the usual problem of cross-validation shrinkage is of negligible importance. Subjective, impressionistic, "clinical" judgments were also made by three prison psychiatrists about probable parole success. The psychiatrists were slightly more accurate than the actuarial tally of favorable factors in predicting parole success, but they were markedly inferior in predicting failure. Furthermore, the actuarial tally made predictions for every case, whereas the psychiatrists left a sizable fraction of cases undecided. The conclusion was clear that even a crude actuarial method such as this was superior to clinical judgment in accuracy of prediction. Of course, we do not know how many of the 21 factors the psychiatrists took into account; but all were available to them; hence, if they ignored certain powerful predictive factors, this would have represented a source of error in clinical judgment. To our knowledge, this is the earliest empirical comparison of two ways of forecasting behavior. One, a formal method, employs an equation, a formula, a graph, or an actuarial table to arrive at a probability, or expected value, of some outcome; the other method relies on an informal, "in the head," impressionistic, subjective conclusion, reached (somehow) by a human clinical judge.

Correspondence concerning this article should be addressed to William M. Grove, Department of Psychology, University of Minnesota, N218 Elliott Hall, 75 East River Road, Minneapolis, Minnesota 55455-0344. Electronic mail may be sent via Internet to grove001@umn.edu.

Thanks are due to Leslie J. Yonce for editorial and bibliographical assistance.

Sarbin (1943) compared the accuracy of a group of counselors predicting college freshmen academic grades with the accuracy of a two-variable cross-validative linear equation in which the variables were college aptitude test score and high school grade record. The counselors had what was thought to be a great advantage. As well as the two variables in the mathematical equation (both known from previous research to be predictors of college academic grades), they had a good deal of additional information that one would usually consider relevant in this predictive task. This supplementary information included notes from a preliminary interviewer, scores on the Strong Vocational Interest Blank (e.g., see Harmon, Hansen, Borgen, & Hammer, 1994), scores on a four-variable personality inventory, an eight-page individual record form the student had filled out (dealing with such matters as number of siblings, hobbies, magazines, books in the home, and availability of a quiet study area), and scores on several additional aptitude and achievement tests. After seeing all this information, the counselor had an interview with the student prior to the beginning of classes. The accuracy of the counselors' predictions was approximately equal to the two-variable equation for female students, but there was a significant difference in favor of the regression equation for male students, amounting to an improvement of 8% in predicted variance over that of the counselors.

Wittman (1941) developed a prognosis scale for predicting outcome of electroshock therapy in schizophrenia, which consisted of 30 variables rated from social history and psychiatric examination. The predictors ranged from semi-objective matters (such as duration of psychosis) to highly interpretive judgments (such as anal-erotic vs. oral-erotic character). None of the predictor variables was psychometric. Numerical weights were not based on the sample statistics but were assigned judgmentally on the basis of the frequency and relative importance ascribed to them in previous studies. We may therefore presume that the weights used here were not optimal, but with 30 variables that hardly matters (unless some of them should not have been included at all). The psychiatric staff made ratings as to prognosis at a diagnostic conference prior to the beginning of therapy, and the assessment of treatment outcome was made by a therapy staff meeting after the conclusion of shock therapy. We can probably infer that some degree of contamination of this criterion rating occurred, which inflated the hits percentage for the psychiatric staff. The superiority of the actuarial method over the clinician was marked, as can be seen in Table 1. It is of qualitative interest that the "facts" entered in the equation were themselves of a somewhat vague, impressionistic sort, the kinds of first-order inferences that the psychiatric raters were in the habit of making in their clinical work.

By 1954, when Meehl published *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence* (Meehl, 1954/1996), there were, depending on some borderline classifications, about 20 such comparative studies in the literature. In every case the statistical method was equal or superior to informal clinical judgment, despite the nonoptimality of many of the equations used. In several studies the clinician, who always had whatever data were entered into the equation, also had varying amounts of further information. (One study, Hovey & Stauffacher, 1953, scored by Meehl for the clinicians, had inflated chi-squares and should have been scored as equal; see McNemar, 1955). The appearance of Meehl's book aroused considerable anxiety in the clinical community and engendered a rash of empirical comparisons over the ensuing years. As

the evidence accumulated (Goldberg, 1968; Gough, 1962; Meehl, 1965f, 1967b; Sawyer, 1966; Sines, 1970) beyond the initial batch of 20 research comparisons, it became clear that conducting an investigation in which informal clinical judgment would perform better than the equation was almost impossible. A general assessment for that period (supplanted by the meta-analysis summarized below) was that in around two fifths of studies the two methods were approximately equal in accuracy, and in around three fifths the actuarial method was significantly better. Because the actuarial method is generally less costly, it seemed fair to say that studies showing approximately equal accuracy should be tallied in favor the statistical method. For general discussion, argumentation, explanation, and extrapolation of the topic, see Dawes (1988); Dawes, Faust, and Meehl (1989, 1993); Einhorn (1986); Faust (1991); Goldberg (1991); Kleinmuntz (1990); Marchese (1992); Meehl (1956a, 1956b, 1956c, 1957b, 1967b, 1973b, 1986a); and Sarbin (1986). For contrary opinion and argument against using an actuarial procedure whenever feasible, see Holt (1978, 1986). The clinical–statistical issue is a sub-area of cognitive psychology, and there exists a large, varied research literature on the broad topic of human judgment under uncertainty (see, e.g., Arkes & Hammond, 1986; Dawes, 1988; Faust, 1984; Hogarth, 1987; Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980; Plous, 1993).

Table 1
Comparison of Actuarial and Clinical Predictions of Outcome of Electroshock Therapy for Schizophrenic Adults

Five-step criterion category	*n*	Percentage of hits	
		Scale	Psychiatrists
Remission	56	90	52
Much improved	66	86	41
Improved	51	75	36
Slightly improved	31	46	34
Unimproved	139	85	49

Note. Values are derived from a graph presented in Wittman (1941).

The purposes of this article are (a) to reinforce the empirical generalization of actuarial over clinical prediction with fresh meta-analytic evidence, (b) to reply to common objections to actuarial methods, (c) to provide an explanation for why actuarial prediction works better than clinical prediction, (d) to offer some explanations for why practitioners continue to resist actuarial prediction in the face of overwhelming evidence to the contrary, and (e) to conclude with policy recommendations, some of which include correcting for unethical behavior on the part of many clinicians.

Results of a Meta-Analysis

Recently, one of us (W.M.G) completed a meta-analysis of the empirical literature comparing clinical with statistical prediction. This study is described briefly here; it is reported in full, with more complete analyses, in Grove, Zald, Lebow, Snitz, and Nelson (2000). To conduct this analysis, we cast our net broadly, including any study which met the following criteria: was published in English since the 1920s; concerned the prediction

of health-related phenomena (e.g., diagnosis) or human behavior; and contained a description of the empirical outcomes of at least one human judgment-based prediction and at least one mechanical prediction. *Mechanical prediction* includes the output of optimized prediction formulas, such as multiple regression or discriminant analysis; unoptimized statistical formulas, such as unit-weighted sums of predictors; actuarial tables; and computer programs and other mechanical schemes that yield precisely reproducible (but not necessarily statistically or actuarially optimal) predictions. To find the studies, we used a wide variety of search techniques which we do not detail here; suffice it to say that although we may have missed a few studies, we think it highly unlikely that we have missed many.

We found 136 such studies, which yielded 617 distinct comparisons between the two methods of prediction. These studies concerned a wide range of predictive criteria, including medical and mental health diagnosis, prognosis, treatment recommendations, and treatment outcomes; personality description; success in training or employment; adjustment to institutional life (e.g., military, prison); socially relevant behaviors such as parole violation and violence; socially relevant behaviors in the aggregate, such as bankruptcy of firms; and many other predictive criteria. The clinicians included psychologists, psychiatrists, social workers, members of parole boards and admissions committees, and a variety of other individuals. Their educations range from an unknown lower bound that probably does not exceed a high school degree, to an upper bound of highly educated and credentialed medical subspecialists. Judges' experience levels ranged from none at all to many years of task-relevant experience. The mechanical prediction techniques ranged from the simplest imaginable (e.g., cutting a single predictor variable at a fixed point, perhaps arbitrarily chosen) to sophisticated methods involving advanced quasi-statistical techniques (e.g., artificial intelligence, pattern recognition). The data on which the predictions were based ranged from sophisticated medical tests to crude tallies of life history facts.

Certain studies were excluded because of methodological flaws or inadequate descriptions. We excluded studies in which the predictions were made on different sets of individuals. To include such studies would have left open the possibility that one method proved superior as a result of operating on cases that were easier to predict. For example, in some studies we excluded comparisons in which the clinicians were allowed to use a "reserve judgment" category for which they made no prediction at all (not even a probability of the outcome in question intermediate between *yes* and *no*), but the actuary was required to predict for all individuals. Had such studies been included, and had the clinicians' predictions proved superior, this could be due to clinicians' being allowed to avoid making predictions on the most difficult cases, the gray ones.

In some cases in which third categories were used, however, the study descriptions allowed us to conclude that the third category was being used to indicate an intermediate level of certainty. In such cases we converted the categories to a numerical scheme such as 1 = *yes*, 2 = *maybe*, and 3 = *no*, and correlated these numbers with the outcome in question. This provided us with a sense of what a clinician's performance would have been were the *maybe* cases split into *yes* and *no* in some proportions, had the clinician's hand been forced.

We excluded studies in which the predictive information available to one method of prediction was not either (a) the same as for the other method or (b) a subset of the

information available to the other method. In other words, we included studies in which a clinician had data *x*, *y*, *z*, and *w*, but the actuary has only *x* and *y*; however, we excluded studies where the clinician had *x* and *y*, whereas the actuary had *y* and *z* or *z* and *w*. The typical scenario was for clinicians to have all the information the actuary had plus some other information; this occurred in a majority of studies. The opposite possibility never occurred; no study gave the actuary more data than the clinician. Thus many of our studies had a bias in favor of the clinician. Because the bias created when more information is accessible through one method than another has a known direction, it only vitiates the validity of the comparison if the clinician is found to be superior in predictive accuracy to a mechanical method. If the clinician's predictions are found inferior to, or no better than, the mechanical predictions, even when the clinician is given more information, the disparity cannot be accounted for by such a bias.

Studies were also excluded when the results of the predictions could not be quantified as correlations between predictions and outcomes, hit rates, or some similarly functioning statistic. For example, if the study was simply reported that the two accuracy levels did not differ significantly, we excluded it because it did not provide specific accuracies for each prediction method.

What can be determined from such a heterogeneous aggregation of studies, concerning a wide array of predictands and involving such a variety of judges, mechanical combination methods, and data? Quite a lot, as it turns out. To summarize these data quantitatively for the present purpose (see Grove et al., 2000, for details omitted here), we took the median difference between all possible pairs of clinical versus mechanical predictions for a given study as the representative outcome of that study. We converted all predictive accuracy statistics to a common metric to facilitate comparison across studies (e.g., convert from hit rates to proportions and from proportions to the arcsin transformation of the proportion; we transformed correlations by means of Fisher's z_r transform—such procedures stabilize the asymptotic variances of the accuracy statistics). This yielded a study outcome that was in study effect size units, which are dimensionless. In this metric, zero corresponds to equality of predictive accuracies, independent of the absolute level of predictive accuracy shown by either prediction method; positive effect sizes represent outcomes favoring mechanical prediction, whereas negative effect sizes favor the clinical method.

Finally, we (somewhat arbitrarily) considered any study with a difference of at least ±.1 study effect size units to decisively favor one method or the other. Those outcomes lying in the interval (–.1, +.1) are considered to represent essentially equivalent accuracy. A difference of .1 effect difference units corresponds to a difference in hit rates, for example, of 50% for the clinician and 60% for the actuary, whereas it corresponds to a difference of .50 correlation with criterion for the clinician versus .57 for the actuary. Thus, we considered only differences that might arguably have some practical import.

Of the 136 studies, 64 favored the actuary by this criterion, 64 showed approximately equivalent accuracy, and 8 favored the clinician. The 8 studies favoring the clinician are not concentrated in any one predictive area, do not over-represent any one type of clinician (e.g., medical doctors), and do not in fact have any obvious characteristics in common. This is disappointing, as one of the chief goals of the meta-analysis was to identify particular areas in which the clinician might outperform the mechanical prediction method. According to the logicians' "total evidence rule," the most plausible

explanation of these deviant studies is that they arose by a combination of random sampling errors (8 deviant out of 136) and the clinicians' informational advantage in being provided with more data than the actuarial formula. (This readily available composite explanation is not excluded by the fact that the majority of meta-analyzed studies were similarly biased in the clinicians' favor, probably one factor that enabled the clinicians to match the equation in 64 studies.) One who is strongly predisposed toward informal judgment might prefer to interpret this lopsided box score as in the following way: "There are a small minority of prediction contexts where an informal procedure does better than a formal one." Alternatively, if mathematical considerations, judgment research, and cognitive science have led us to assign a strong prior probability that a formal procedure should be expected to excel, we may properly say, "Empirical research provides no clear, replicated, robust examples of the informal method's superiority."

Experience of the clinician seems to make little or no difference in predictive accuracy relative to the actuary, once the average level of success achieved by clinical and mechanical prediction in a given study is taken into account. Professional training (i.e., years in school) makes no real difference. The type of mechanical prediction used does seem to matter; the best results were obtained with weighted linear prediction (e.g., multiple linear regression). Simple schemes such as unweighted sums of raw scores do not seem to work as well. All these facts are quite consistent with the previous literature on human judgment (e.g., see Garb, 1989, on experience, training, and predictive accuracy) or with obvious mathematical facts (e.g., optimized weights should outperform unoptimized weights, though not necessarily by very much).

Configural data combination formulas (where one variable potentiates the effect of another; Meehl, 1954/1996, pp. 132-135) do better than nonconfigural ones, on the average. However, this is almost entirely due to the effect of one study by Goldberg (1965), who conducted an extremely extensive and widely cited study on the Minnesota Multiphasic Personality Inventory (MMPI) as a diagnostic tool. This study contributes quite disproportionately to the effect size distribution, because Goldberg compared two types of judges (novices and experts) with an extremely large number of mechanical combination schemes. With the Goldberg study left out of account, the difference between configural and nonconfigural mechanical prediction schemes, in terms of their superiority to clinical prediction, is very small (about two percentage points in the hit rate).

The great preponderance of studies either favor the actuary outright or indicate equivalent performance. The few exceptions are scattered and do not form a pocket of predictive excellence in which clinicians could profitably specialize. In fact, there are many fewer studies favoring the clinician than would be expected by chance, even for a sizable subset of predictands, if the two methods were statistically equivalent. We conclude that this literature is almost 100% consistent and that it reproduces and amplifies the results obtained by Meehl in 1954 (Meehl, 1954/1996). Forty years of additional research published since his review has not altered the conclusion he reached. It has only strengthened that conclusion.

Replies to Commonly Heard Objections

Despite 66 years of consistent research findings in favor of the actuarial method, most professionals continue to use a subjective, clinical judgment approach when making predictive decisions. The following sections outline some common objections to actuarial

procedures; the ordering implies nothing about the frequency with which the objections are raised or the seriousness with which any one should be taken.

"*We Do Not Use One Method or the Other— We Use Both; It Is a Needless Controversy Because the Two Methods Complement Each Other, They Do Not Conflict or Compete*"

This plausible-sounding, middle-of-the-road "compromise" attempts to liquidate a valid and socially important pragmatic issue. In the phase of discovery psychologists get their ideas from both exploratory statistics and clinical experience, and they test their ideas by both methods (although it is impossible to provide a strong test of an empirical conjecture relying on anecdotes). Whether psychologists "use both" at different times is not the question posed by Meehl in 1954 (Meehl, 1954/1996). No rational, educated mind could think that the only way we can learn or discover anything is either (a) by interviewing patients or reading case studies or (b) by computing analyses of covariance. The problem arises not in the research process of the scientist or scholarly clinician, but in the pragmatic setting, where we are faced with predictive tasks about individuals such as mental patients, dental school applicants, criminal offenders, or candidates for military pilot training. Given a data set (e.g., life history facts, interview ratings, ability test scores, MMPI profiles, nurses' notes), how is one to put these various facts (or first-order inferences) together to arrive at a prediction about the individual? In such settings, there are two pragmatic options. Most decisions made by physicians, psychologists, social workers, judges, parole boards, deans' admission committees, and others who make judgments about human behavior are made through "thinking about the evidence" and often discussing it in team meetings, case conferences, or committees. That is the way humans have made judgments for centuries, and most persons take it for granted that that is the correct way to make such judgments.

However, there is another way of combining that same data set, namely, by a mechanical or formal procedure, such as a multiple regression equation, a linear discriminant function, an actuarial table, a nomograph, or a computer algorithm. It is a fact that these two procedures for data combination do not always agree, case by case. In most predictive contexts, they disagree in a sizable percentage of the cases. That disagreement is not a theory or philosophical preference; it is an empirical fact. If an equation predicts that Jones will do well in dental school, and the dean's committee, looking at the same set of facts, predicts that Jones will do poorly, it would be absurd to say, "The methods don't compete, we use both of them." One cannot decide both to admit and to reject the applicant; one is forced by the pragmatic context to do one or the other.

Of course, one might be able to improve the committee's subsequent choices by educating them in some of the statistics from past experience; similarly, one might be able to improve the statistical formula by putting in certain kinds of data that the clinician claims to have used in past cases where the clinician did better than the formula. This occurs in the discovery phase in which one determines how each of the two procedures could be sharpened for better performance in the future. However, at a given moment in time, in a given state of knowledge (however attained), one cannot use both methods if they contradict one another in their forecasts about the instant case. Hence, the question inescapably arises, "Which one tends to do a better job?" This controversy has not been "cooked up" by those who have written on the topic. On the contrary, it is intrinsic to the pragmatic setting for any decision maker who takes the task seriously and wishes to

behave ethically. The remark regarding compromise recalls statistician Kendall's (1949) delightful passage:

> A friend of mine once remarked to me that if some people asserted that the earth rotated from East to West and others that it rotated from West to East, there would always be a few well-meaning citizens to suggest that perhaps there was something to be said for both sides and that maybe it did a little of one and a little of the other; or that the truth probably lay between the extremes and perhaps it did not rotate at all. (p. 115)

"*Pro-Actuarial Psychologists Assume That Psychometric Instruments (Mental Tests) Have More Validity Than Nonpsychometric Findings, Such as We Get From Mental Status Interviewing, Informants, and Life History Documents, but Nobody Has Proved That Is True*"

This argument confuses the character of data and the optimal mode of combining them for a predictive purpose. Psychometric data may be combined impressionistically, as when we informally interpret a Rorschach or MMPI profile, or they may be combined formally, as when we put the scores into a multiple regression equation. Nonpsychometric data may be combined informally, as when we make inferences from a social case work history in a team meeting, but they may also be combined formally, as in the actuarial tables used by Sheldon and Eleanor T. Glueck (see Thompson, 1952), and by some parole boards, to predict delinquency. Meehl (1954/1996) was careful to make the distinction between kind of data and mode of combination, illustrating each of the possibilities and pointing out that the most common mode of prediction is informal, nonactuarial combining of psychometric and nonpsychometric data. (The erroneous notion that nonpsychometric data, being "qualitative," preclude formal data combination is treated below.)

There are interesting questions about the relative reliability and validity of first-, second-, and third-level inferences from nonpsychometric raw facts. It is surely permissible for an actuarial procedure to include a skilled clinician's rating on a scale or a nurse's chart note using a nonquantitative adjectival descriptor, such as "withdrawn" or "uncooperative." The most efficacious level of analysis for aggregating discrete behavior items into trait names of increasing generality and increasing theoretical inferentiality is itself an important and conceptually fascinating issue, still not adequately researched; yet it has nothing to do with the clinical versus statistical issue because, in whatever form our information arrives, we are still presented with the unavoidable question, "In what manner should these data be combined to make the prediction that our clinical or administrative task sets for us?" When Wittman (1941) predicted response to electroshock therapy, most of the variables involved clinical judgments, some of them of a high order of theoreticity (e.g., a psychiatrist's rating as to whether a schizophrenic had an anal or an oral character). One may ask, and cannot answer from the armchair, whether the Wittman scale would have done even better at excelling over the clinicians (see Table 1 above) if the three basic facets of the anal character had been separately rated instead of anality being used as a mediating construct. However, without answering that question, and given simply the psychiatrist's subjective impressionistic clinical judgment, "more anal than oral," that is still an item like any other "fact" that is a candidate for combination in the prediction system.

"*Even if Actuarial Prediction Is More Accurate, Less Expensive, or Both, as Alleged, That Method Does Not Do Most Practitioners Any Good Because in Practice We Do Not Have a Regression Equation or Actuarial Table*"

This is hardly an argument for or against actuarial or impressionistic prediction; one cannot use something one does not have, so the debate is irrelevant for those who (accurately) make this objection. We could stop at that, but there is something more to be said, important especially for administrators, policymakers, and all persons who spend taxpayer or other monies on predictive tasks. Prediction equations, tables, nomograms, and computer programs have been developed in various clinical settings by empirical methods, and this objection presupposes that such an actuarial procedure could not safely be generalized to another clinic. This brings us to the following closely related objection.

"*I Cannot Use Actuarial Prediction Because the Available (Published or Unpublished) Code Books, Tables, and Regression Equations May Not Apply to My Clinic Population*"

The force of this argument hinges on the notion that the slight nonoptimality of beta coefficients or other statistical parameters due to validity generalization (as distinguished from cross-validation, which draws a new sample from the identical clinical population) would liquidate the superiority of the actuarial over the impressionistic method. We do not know of any evidence suggesting that, and it does not make mathematical sense for those predictive tasks where the actuarial method's superiority is rather strong. If a discriminant function or an actuarial table predicts something with 20% greater accuracy than clinicians in several research studies around the world, and one has no affirmative reason for thinking that one's patient group is extremely unlike all the other psychiatric outpatients (something that can be checked, at least with respect to incidence of demographics and formal diagnostic categories), it is improbable that the clinicians in one's clinic are so superior that a decrement of, say, 10% for the actuarial method will reduce its efficacy to the level of the clinicians. There is, of course, no warrant for assuming that the clinicians in one's facility are better than the clinicians who have been employed as predictors in clinical versus statistical comparisons in other clinics or hospitals. This objection is especially weak if it relies upon readjustments that would be required for optimal beta weights or precise probabilities in the cells of an actuarial table, because there is now a sizable body of analytical derivations and empirical examples, explained by powerful theoretical arguments, that equal weights or even randomly scrambled weights do remarkably well (see extended discussion in Meehl 1992a, pp. 380-387; cf. Bloch & Moses, 1988; Burt, 1950; Dawes, 1979, 1988, chapter 10; Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975; Gulliksen, 1950; Laughlin, 1978; Richardson, 1941; Tukey, 1948; Wainer, 1976, 1978; Wilks, 1938). (However, McCormack, 1956, has shown that validities, especially when in the high range, may differ appreciably despite high correlation between two differently weighted composites). If optimal weights (neglecting pure cross-validation shrinkage in resampling from one population) for the two clinical populations differ considerably, an unweighted composite will usually do better than either will alone when applied to the other population (validity generalization shrinkage). It cannot simply be assumed that if an actuarial formula works in several outpatient psychiatric populations, and each of them does as well as the local clinicians or better, the formula will not work well in one's own clinic. The turnover in clinic professional personnel, and with more recently trained staff having received their

training in different academic and field settings, under supervisors with different theoretical and practical orientations, entails that the "subjective equation" in each practitioner's head is subject to the same validity generalization concern and may be more so than formal equations.

It may be thought unethical to apply someone else's predictive system to one's clientele without having validated it, but this is a strange argument from persons who are daily relying on anecdotal evidence in making decisions fraught with grave consequences for the patient, the criminal defendant, the taxpayer, or the future victim of a rapist or armed robber, given the sizable body of research as to the untrustworthiness of anecdotal evidence and informal empirical generalizations. *Clinical experience* is only a prestigious synonym for *anecdotal evidence* when the anecdotes are told by somebody with a professional degree and a license to practice a healing art. Nobody familiar with the history of medicine can rationally maintain that whereas it is ethical to come to major decisions about patients, delinquents, or law school applicants without validating one's judgments by keeping track of their success rate, it would be immoral to apply a prediction formula which has been validated in a different but similar subject population.

If for some reason it is deemed necessary to revalidate a predictor equation or table in one's own setting, to do so requires only a small amount of professional time. Monitoring the success of someone else's discriminant function over a couple of years' experience in a mental hygiene clinic is a task that could be turned over to a first-year clinical psychology trainee or even a supervised clerk. Because clinical predictive decisions are being routinely made in the course of practice, one need only keep track and observe how successful they are after a few hundred cases have accumulated. To validate a prediction system in one's clinic, one does not have to do anything differently from what one is doing daily as part of the clinical work, except to have someone tallying the hits and misses. If a predictor system does not work well, a new one can be constructed locally. This could be done by the Delphi method (see, e.g., Linstone & Turoff, 1975), which combines mutually modified expert opinions in a way that takes a small amount of time per expert. Under the assumption that the local clinical experts have been using practical clinical wisdom without doing formal statistical studies of their own judgments, a formal procedure based on a crystallization of their pooled judgments will almost certainly do as well as they are doing and probably somewhat better. If the clinical director is slightly more ambitious, or if some personnel have designated research time, it does not take a research grant to tape record remarks made in team meetings and case conferences to collect the kinds of facts and first-level inferences clinicians advance when arguing for or against some decision (e.g., to treat with antidepressant drugs or with group therapy, to see someone on an inpatient basis because of suicide risk, or to give certain advice to a probate judge). A notion seems to exist that developing actuarial prediction methods involves a huge amount of extra work of a sort that one would not ordinarily be doing in daily clinical decision making and that it then requires some fancy mathematics to analyze the data; neither of these things is true.

"*The Results of These Comparative Studies Just Do Not Apply to Me as an Individual Clinician*"

What can one say about this objection, except that it betrays a considerable professional narcissism? If, over a batch of, say, 20 studies in a given predictive domain, the

typical clinician does a little worse than the formula, and the best clinician in each study—not cross-validated as "best"—does about equal to the formula or slightly better, what except pride would entitle a clinician, absent an actuarial study of one's own predictive powers in competition with a formula, to think that one is at the top of the heap? Given 20 studies, with, on average, each of them involving, say, five clinicians, and only 1 or 2 out of the total 100 clinicians beating the formula, what would entitle a particular clinician to assert, absent empirical evidence of one's truly remarkable superiority to other practitioners, that one is in the top 1%? One need not be an orthodox Bayesian to say that has a rather low prior and therefore requires strong support. The clinician is not entitled to assert such superiority without collecting track record data.

"*I Cannot Use Actuarial Prediction Because It Is More Expensive Than Clinical Prediction*"

This objection is obviously in need of large scale, diversified empirical investigation. If I apply a formula developed in another clinic, the cost is negligible compared with the cost of a team meeting or case conference. The cost of developing a tailor-made formula in one's own clinic by assigning a graduate student to do some simple statistics is also less costly than usual clinical procedures for decision making. One of us (P.E.M.) computed years ago the cost in personnel hours of a Veterans Administration case conference and estimated conservatively that to reach decisions about the patient in that way cost the taxpayer at least 12 times as much as it would cost to have a clerk apply a formula under supervision by a doctoral-level psychologist. On the one hand, for predictive tasks in which there is a significant superiority of the formula, utility and ethical considerations enter the picture, sometimes decisively. On the other hand, proprietary actuarial–mechanical prediction services are not free. For example, the cost of the Minnesota Report (Butcher, 1986), an automated MMPI–2 interpretation service, is currently about $30 per case. If clinicians are paid $30 per hour ($60,000 per year) and can do as well as the automated report, they are cheaper as MMPI–2 interpreters if they take less than one 1 hour per case; most clinicians we have observed take 10–40 minutes per profile.

"*Clinicians Want Not Merely to Predict but to Change Behavior From What Would Be Predicted Without Intervention*"

The fallacy here is to suppose that one can select an intervention aimed toward changing behavior without implicitly relying on a prediction. From the decision theory standpoint, not doing anything is, of course, a form of action; therefore, this may be included as one of the options among which one chooses. If one intends to do anything, it is because one hopes and expects that doing some action to, for, or with a patient will reduce the probability of an undesirable outcome, O_U, or raise the probability of a desirable outcome O_D. Generalizing, one can imagine a set of envisaged outcomes (e.g., failure in air crew training, earning a PhD in 5 years, recovering from a depression, committing another rape) associated with certain dispositions that the individual has and kinds of intervention (e.g., psychological, social, chemical, legal) that will alter the distribution of outcome probabilities. No matter how inaccurately one does this, no matter how great or little faith one has in the process, if there were no such background hope and expectation, the whole enterprise would be feckless and certainly not a

justifiable expenditure of the taxpayers' money. Therefore, the argument that we do not want only to predict behavior but to change it is based on the simple mistake of not seeing that the selection of an intervention is predicated on the belief—sound or unsound, warranted or unwarranted—that the intervention will redistribute the outcome probabilities in the desired direction. This line of reasoning applies at various levels of description and analysis, both to long-term socially defined consequences of numerous behavioral events (e.g., student X will succeed in dental school) and to narrowly specified individual dispositions (depressed patient X will attempt suicide). The basic logic and statistics of the situation have not changed.

The reasoning holds even for the expected outcome of a therapist's single action during psychotherapy (e.g., remaining silent vs. a Rogerian reflection vs. a psychoanalytic interpretation vs. a rational-emotive therapy philosophical challenge). One does not think of that decision process as proceeding actuarially, but experienced therapists, when asked why they do (or avoid) a certain kind of thing, will typically claim that their clinical experience leads them to think that a certain kind of remark usually (or rarely) works. A computerized rapid moment-to-moment analysis of the patient's discourse as a signaler to the therapist is something that, to our knowledge, has not been tried; however, given the speed of the modern computer, it would be foolish to reject such a science fiction idea out of hand. Yet that is not the predictive context that we are addressing here. If one does anything, including both refraining from action and intervening, the justification for it—economic, scientific, ethical, educational—always lies in some set of empirical beliefs (or at least hopes) regarding empirical probabilities and their susceptibility to influence by the set of interventions available.

"*Statistical Predictionists Aggregate, Whereas We Seek to Make Predictions for the Individual, so the Actuarial Figures Are Irrelevant in Dealing With the Unique Person*"

This complaint, unlike most, at least has some slight philosophical interest because the precise "logic" of how one properly applies an empirical relative frequency to the individual case has deep epistemological components. Unfortunately, space does not permit us to develop those in detail, and it would be undesirable to treat them superficially. The short, forceful reply proceeds like this: Suppose you are suffering from a distressing illness, painful or incapacitating, and your physician says that it would be a good idea to have surgeon X perform a certain radical operation in the hope of curing you. You would naturally inquire whether this operation works for this disease and how risky it is. The physician might say, "Well, it doesn't always work, but it's a pretty good operation. It does have some risk. There are people who die on the operating table, but not usually." You would ask, "Well, what percentage of times does it work? Does it work over half the time, or 90%, or what? And how many people die under the knife? One in a thousand? If it were five in a hundred, I don't know that I'd want to take the chance, even though this illness is irksome to me." How would you react if your physician replied, "Why are you asking me about statistics? We are talking about *you*—an individual patient. You are unique. Nobody is exactly like you. Do you want to be a mere statistic? What differences do those percentages make, anyway?" We do not think a person should be pleased if the doctor replied in that evasive fashion. Why not? Because, as Bishop

Butler (1736) said, probability is the guide of life. The statistics furnish us with the probabilities so far as anything can.

Claiming concern with the unique person rather than an aggregate receives illegitimate, fallacious weight from an assumption that the antiactuarial objector would not dare to assert explicitly: that the statistics give mere probabilities, average results, or aggregate proportions, whereas in dealing with the unique individual one will know exactly what will befall that person. Of course, such a claim can almost never be made. If the proposed operation does invariably cure all patients with the disease, and if nobody ever dies on the operating table, then the physician's proper (statistical) answer is that it is 100% successful and it has 0% risk. If the physician cannot claim that, it means that there are other percentages involved, both for the cure rate and for the risk of death. Those numbers are there, they are objective facts about the world, whether or not the physician can readily state what they are, and it is rational for you to demand at least a rough estimate of them. But the physician cannot tell you beforehand into which group—success or failure—*you* will surely fall.

Alternatively, suppose you are a political opponent held in custody by a mad dictator. Two revolvers are put on the table and you are informed that one of them has five live rounds with one empty chamber, the other has five empty chambers and one live cartridge, and you are required to play Russian roulette. If you live, you will go free. Which revolver would you choose? Unless you have a death wish, you would choose the one with the five empty chambers. Why? Because you would know that the odds are five to one that you will survive if you pick that revolver, whereas the odds are five to one you will be dead if you choose the other one. Would you seriously think, "Well, it doesn't make any difference what the odds are. Inasmuch as I'm only going to do this once, there is no aggregate involved, so I might as well pick either one of these two revolvers; it doesn't matter which"?

There is a real problem, not a fallacious objection, about uniqueness versus aggregates in defining what the statisticians call the reference class for computing a particular probability in coming to a decision about an individual case. We may hold that there is a real probability that attaches to the individual patient Jones as regards the individual behavior event, but we do not know what that real probability is. We could assign Jones to various patient categories and get the probability of the event (e.g., suicide or recovery); the resulting proportions would differ depending on which reference class we used. We might, for example, know of a good study indicating 80% success with depressed patients having symptom combination *x*, *y*, *z* and another study that does not tell us about symptoms *y* and *z* but only *x* and also disaggregates the class with regard to age or number of previous episodes. Here the situation is the same as that faced by an insurance actuary. To assign the probability of Meehl's death in the following year, we would start with his being a Caucasian male, age 75. There is a huge mass of statistical data assigning that *p* value. If we add the fact that he has a mitral valve lesion from rheumatic fever, the probability of death rises somewhat. If we add the fact that he is not overweight, takes a 5-mile (8.0 km) walk daily, and has quit smoking, the probability of death goes down again. If we now add the fact that he has some degree of left ventricular hypertrophy, the death probability goes up, and so forth. Each of these probabilities is an objectively correct relative frequency for the reference class on which it was computed. (We are here neglecting sampling error in proportions, which is not relevant to the

present issue.) It is important to note that there are as many probabilities as there are reference classes. Which reference class should we choose? Reichenbach's (1938) answer was to choose the narrowest reference class (richest intension, smallest extension) for which the number of cases is large enough to provide stable relative frequencies. That is not satisfactory as it stands, because the stability of a proportion is not a yes–no matter but changes continuously with changes in sample size. The insurance company's examining physician provides the data on which a recommendation is made, but if the physician's recommendation goes against a strong actuarial finding, the latter will be followed in deciding whether to insure or to assign a special premium rate.

The empirical—some would say metaphysical—question as to whether complete nomological determinism holds for human behavior fortunately does not need to be answered in this context. There are hardly any clinical, counseling, or personnel decisions made by either formal or informal procedures that informed persons claim to be absolutely certain. (To find any such, you would have to imagine bizarre situations, such as predicting that a person with IQ 75 on repeated testings and mentally retarded by other social criteria could not achieve a PhD in theoretical physics.) The insurance actuary knows that many facts could be added in defining more and more restrictive reference classes, but it does not pay to attempt to work out life tables which take account of all possible configurations. The number of reference classes rises exponentially with the number of factual or inferential predictors used (e.g., 10 dichotomous factors yield 1,024 subcategories).

This application of aggregate statistics to a decision about an individual case does give rise to one of the few intellectually interesting concerns of antistatistical clinicians. Suppose there are certain facts about the individual that are so rare that researchers setting up prediction systems have not seen fit to include them in the actuarial formula but that are so important when they do occur that they should be permitted to countervail even the strongest actuarial probability. It is not satisfactory to say that if they are all that rare, it does not matter. For a particular patient it matters if we guess wrong, and in that sense we are surely concerned about this individual. Second, while a particular fact may have a low probability of being present in our data for a class of patients, there may be a large number of such (different) particular facts, each of which is rarely seen but that in aggregate define a sizable subset of patients for whom the actuarial equation should be countermanded. As the statistician's joke has it: "An improbable event is one that almost never happens, but improbable events happen every day." Meehl (1954/1996) explicitly addressed this. He considered the situation of a sociologist studying leisure time activities who has worked out a regression equation for predicting whether people will go to the movies on a certain night. The data indicate that Professor X has a probability $p = .84$ of going to a movie on Friday night, with the equation including demographic information such as academic occupation, age, and ethnicity, and ideally some previous statistics on this individual. (It is, of course, a mistake to assume that all statistics must be cross-sectional and never longitudinal as to their database.) Suppose that the researcher then learns that Professor X has a fractured femur from an accident of a few days ago and is immobilized in a hip cast. Obviously, it would be absurd to rely on the actuarial prediction in the face of this overwhelmingly prepotent fact. Among the proactuarial psychologists, this example has come to be known as "the broken leg case." We think

that research on this kind of situation is one of the most important areas of study for clinical psychologists.

The obvious, undisputed desirability of countervailing the equation in the broken leg example cannot automatically be employed antiactuarially when we move to the usual prediction tasks of social and medical science, where physically possible human behavior is the predictand. What is the bearing of the empirical comparative studies on this plausible, seductive extrapolation from a clear-cut "physical" case? Consider the whole class of predictions made by a clinician, in which an actuarial prediction on the same set of subjects exists (whether available to the clinician and, if so, whether employed or not). For simplicity, let the predictand be dichotomous, although the argument does not depend on that. In a subset of the cases, the clinical and actuarial prediction are the same; among those, the hit rates will be identical. In another subset, the clinician countermands the equation in the light of what is perceived to be a broken leg countervailer. We must then ask whether, in these cases, the clinician tends to be right more often than not. If that is the actuality, then in this subset of cases, the clinician will outperform the equation. Because in the first subset the hit rates are identical and in the countermanded subset of psychological or social "broken legs" the clinician does better than the equation, it follows by simple arithmetic that the clinician must do better on the whole group (both subsets combined) than does the equation. However, because the empirical comparative studies show this consequence to be factually false, it follows necessarily that clinicians' broken leg countermandings tend to be incorrect.

The problem that antiactuarial clinicians have with this simple reasoning is that they focus their attention on the cases in which they could have saved an actuarial mistake, neglecting the obvious point that any such decision policy, unless infallible, will also involve making some mistakes in the opposite direction. It is the same old error of "men mark where they hit, and not where they miss," as Jevons (1874/1958) put it. This is not a complicated problem in epistemology or higher mathematics; it is simply the ineradicable tendency of the human mind to select instances for generalizations that it favors. It is the chief source of superstitions.

What is wrong with the analogy between the broken leg case and countervailing a regression equation because of an alleged special circumstance in the environment or rare attribute of the individual, when done by a parole board, psychotherapist, or dean's selection committee? The answer is obvious. In the broken leg example, there are two near certainties relied on, which are so blindingly clear from universal human experience that no formal statistical study is needed to warrant our having faith in them. First, a broken leg in a hip cast is a highly objective fact about the individual's condition, ascertainable by inspection with quasi-perfect reliability. Second, the immobilizing consequence of such a condition accords with universal experience, not tied to particular questions, such as whether a person in such circumstances will go to the movies. The physiological-mechanical "law" relied on is perfectly clear, universally agreed on, not a matter of dispute based on different theories or ideologies or engendered by different kinds of training or clinical experience. We have here an almost perfectly reliable ascertainment of a fact and an almost perfect correlation between that fact and the kind of fact being predicted. Neither one of these delightful conditions obtains in the usual kind of social science prediction of behavior from probabilistic inferences regarding probable environmental influences and probabilistic inferences regarding the individual's behavior

dispositions. Neither the "fact" being relied on to countervail the equation nor the correlation between that kind of fact and the outcome is usually known with high accuracy. Those behavior predictors who reject the comparative accuracy research or deny its practical implications by invoking the broken leg paradigm are deceiving themselves and the policy makers they persuade.

There is a more interesting conceptual breakdown of the total class of cases that deserves theoretical analysis and empirical study. The existence of this interesting problem does not contradict our general line of reasoning, which is a straightforward application of the principle that if something pays off—a question of fact—then we should use it, but not otherwise. Not all disagreements between the clinician and the actuarial formula represent conscious countervailings on the basis of alleged broken leg situations. Some of the deviations—perhaps most of them, inasmuch as we know—do not involve the clinician's thinking about something special, but simply (a) the assignment of nonoptimal weights to the same facts that the actuary is using and (b) unreliability (inconsistency) in informally applying these subjective weights (see discussion of the Goldberg paradox below). It could be that the (rare?) countervailings by the clinician induced by sociopsychological "broken leg" situations or attributes does pay off more often than not. However, because the total class of disagreements includes these (together with the unreliable application of nonoptimal weights), the adverse influence of this second set produces a statistical swamping of the smaller subset of valid broken leg countervailings. This complex situation still leaves the clinician's judgment equal or behind the formula overall (as the studies show) and, hence, would not warrant our preferring informal predictions to the actuarial ones. However, it is possible that by dissuading the clinician from broken leg countervailings, we would be reducing the overall success rate, even below what it now is, because ceasing to make broken leg countervailings does not automatically do anything to improve the other subset where disagreement is based not upon imputed broken legs but merely upon unreliable application of nonoptimal weights.

If research should succeed in showing this, the ameliorative prescription would be educating sophisticated (and rational!) clinicians to realize that, in general, they do not do as well as the equation and then to realize how they can improve upon the equation once in a while by clear-cut "broken leg" countervailings but that they should set a high threshold for countervailing the equation (cf. Meehl, 1957b). This is a very important question for research and we are unaware of even a single study that addresses it.

"*Understanding an Individual Patient (Client, Applicant, Criminal Offender) Is an Idiographic Rather Than a Nomothetic Undertaking, Hence, Statistics—a Kind of Nomothetic Information—Do Not Apply*"

The distinction between the idiographic and the nomothetic approaches to "understanding something," introduced by the German philosopher Wilhelm Windelband in the last century, was emphasized for psychologists and other social scientists by Gordon Allport (1937). It is related to, but not identical with, the German scholars' distinction between two sorts of substantive disciplines, *Geisteswissenschaften* and *Naturwissenschaften*, the former dealing with mind and society and the latter with the inorganic and the nonmental biological sciences. Some have held, and others have vigorously denied, that what British terminology calls the moral sciences—history,

sociology, psychology, political science, economics—have a peculiar method, emphasizing details of the particular sequence of events rather than emphasizing the search for, or even the application of, general laws. That is a deep question involving logical, epistemological, and perhaps metaphysical issues beyond the scope of the present article; what we say here must unavoidably have a certain appearance of dogmatism about matters still in dispute among scholars.

The short answer to this antiactuarial argument for the policy maker is that even supposing the distinction between disciplines were a fundamental, qualitative one (rather than, as most social scientists would hold today, a matter of degree and predominant interest), the pragmatic question must nevertheless be faced: whether the peculiarly idiographic method tends to lead to successful judgments more than the nomothetic one. That is clearly an empirical question rather than a purely philosophical one decidable from the armchair, and the empirical evidence is, as described above, massive, varied, and consistent. In the present context, that pragmatic finding could suffice, but we will offer a few additional comments by way of clarification.

In making the nomothetic–idiographic distinction, one must be clear about whether it is a matter of one's scholarly interest or of generic principles of method that cut across interests. A historian who studies the state documents of countries involved in the outbreak of World War I has an inherently idiographic interest, a desire to get an accurate narration of what took place and, within limits, of how and why it took place as it did. The historian pursuing this scholarly interest cannot be faulted for not trying to come up with general laws of economics, political science, or history. On the other hand, in ascertaining the idiographic "facts," the historian unavoidably makes use of some general principles, and these are, by definition, nomothetic in character. One reason a philosophical idiographer may mistakenly miss this crucial point is that the identification of the nomothetic with the natural sciences (e.g., physics, chemistry, astronomy) generates a mental set that formal, rigorous, exceptionless regularities—laws of nature—expressible in mathematical equations comprise the only kind of nomothetic information in a knowledge domain. That is incorrect, because the historian or biographer makes use of rough general commonsense principles about human conduct. An example would be Aristotle's practical syllogism: If one desires a certain goal and believes that a particular action will tend to bring about that goal, then one ought (in an appropriate instrumental means–end rather than moral sense) to perform that action. This syllogism is similar to Kant's hypothetical (as distinguished from categorical) imperative. This special kind of inference can be reformulated, not as a criterion of rationality for a decision maker but as a (statistical) generalization about human conduct (which tends to be rational in this respect): *ceteris paribus*, an agent who believes that a certain action will produce a particular goal, and who wants to realize that goal, will in fact have a strong disposition (propensity) to perform the action. That people act this way, and that most people who are sane and who take their choices seriously believe certain means–end relations hold empirically need not be matters of technical decision theory, or psychology of motivation and learning, or cognitive science generally, but are based on our commonsense observations, known long before the rise of modern social or natural science. Thus, it is erroneous to think that if one's interest is idiographic (i.e., the narration and, so far as obtainable, comprehension of a sequence of events of a particular historical or personal sort), therefore, nothing nomothetic can or should be relied on.

Second, although there is clearly a correlation between the idiographic–nomothetic distinction and the social science–natural science distinction, it is not a one-to-one correspondence. The physical sciences, despite their predominantly nomothetic concerns, do sometimes include important idiographic conjectures. The big bang theory in cosmology, theories about the origin of our solar system, Lyell's uniformitarianism versus Buffon's castrophism in the earth's history, Wegener's hypothesis of continental drift (long laughed at but now accepted in the modern form of plate tectonics), the various explanations of individual earthquakes or of how Yellowstone's Old Faithful works, are all clearly idiographic matters belonging to the inorganic realm. The theory of evolution is idiographic in biology, although the explanatory mechanisms purport to be nomothetic. Most "laws" in biological and social science are, from the strict logicians' viewpoint, more like accidental universals (e.g., all coins in my pocket are silver) than true nomologicals—laws of nature, such as Maxwell's equations or quantum mechanics. This is because the biological laws are structure dependent, hinging on the existence of certain kinds of organisms, which could have been otherwise without the basic nomologicals being different but with slightly altered initial conditions or perhaps quantum-indeterminate crucial events. There could have been unicorns but no hippopotamus. "All mammals have young born alive" was thought to be a biological law, until we explored Australia; likewise, "all swans are white" and many other examples. (For further discussion, see Meehl, 1970a, pp. 385-391, references in footnote 11, pp. 385-387, the Carnap quotation in footnote 14, p. 391, and footnote 18, pp. 395-396). Precise formulation of the distinction between nomologicals and accidental universals is a highly technical problem in philosophy of science and one that we believe is still unsolved. In medicine, it would be quite wrong to say that, because pathology seeks out general laws about what disordered tissue conditions give rise to which clinical syndromes, a pathologist's report concerning an individual patient is not scientifically allowable. Thus, although we readily accept the distinction of aim involved between idiographic and nomothetic research and allow for the obvious administrative distinctions between social, biological, and physical sciences in the academy, we reject the implication of a near-perfect correlation between these two dichotomies.

Finally, the "uniqueness" of a particular event can never be used as a ground for rejecting nomothetic formulations, whether they are strictly nomological or only, as in the subject matter of this article, stochastological (Meehl, 1978c). With the exception of elementary atomic processes, all events in the physical, biological, and social world are, taken literally, utterly unique. Every explosion is unique, but each takes place in accordance with the laws of chemistry. Every fatal coronary attack is unique, although it fits into the general laws of pathophysiology. Every epidemic of a disease is unique, but the general principles of microbiology and epidemiology obtain. The short answer to the objection to nomothetic study of persons because of the uniqueness of each was provided by Allport (1937), namely, the nomothetic science of personality can be the study of how uniqueness comes about. As Thurstone (1947) put it, to object to the statistical method of factor analysis on the grounds that each person, whatever the particular test scores or inferred factor scores, is unique would lead one to reject theoretical economics or the accounting practices of banks on the ground that the statement "Smith and Jones have the same income" is inadmissible as Smith works for a living, whereas Jones steals.

"*The Important Clinical Data Are Qqualitative (e.g., Informants' Adjectives, Criminal Conviction Record, Narrative Case Studies, Nurses' Chart Notes), Not Numerical, so One Cannot Use Mathematics on Them*"

It has been known for many years in social science that anything capable of being recorded in a document can be encoded. Thus, for instance, a whole class of adjectives that might be used to describe a juvenile delinquent (taken from the dictionary or ramified searching of a thesaurus, entered with the generic *aggressive*) can be tallied by encoding each textual appearance with the numeral 1. In this way, we can count occurrences of the verbal class in statements by informants, mental status examination, intake social workers' notes, and so forth. The common error is to think that any actuarial system must necessarily be in the form of a weighted composite of quantitative variables, such as a linear discriminant function or a multiple regression equation. An actuarial table with proportions is, of course, a formal objective procedure for combining encoded data. "Formal" does not mean numerical, involving quantitative dimensions (scales, a metric, such as psychometric test scores), although it includes them, as well as rank orderings. If the implication is that formalized encoding eliminates the distinctive advantages of the usual narrative summary and hence loses subtle aspects of the flavor of the personality being appraised, that is doubtless true. However, the factual question is then whether those allegedly uncodable configural features contribute to successful prediction, which again comes back to the negative findings of the studies.

This is as good a place as any to stress that mere encoding alone does not make an actuarial prediction formula or prediction table. All actuarial procedures are mechanical (formal, algorithmic, automated), but not all mechanical procedures are actuarial. A computer-printed interpretation is not de facto a statistical prediction procedure.

"*The Relationship Between Me and My Patient or Client Is an "I–Thou" Relationship, Not a Cold Mechanical One, and Statistical Prediction Treats the Individual as an Object, Like a White Rat or a Coin Being Flipped Rather Than as a Person; Hence, It Is Inhumane and Degrading, Depriving the Person of Dignity*"

First, advice or decision or communication to an empowered third party (e.g., judge, dental school dean) as arrived at by the most efficacious method (i.e., the one that results in better probability of successful prediction) is not the same phase of the case handling that occurs in the face-to-face interaction with the patient, client, candidate, or offender. It would be absurd to say that if a physician rationally prefers penicillin to sulfadiazine in treating a strep throat because objectively that is what the studies show works better on the average (and is therefore what has a better chance to be good for the patient), then in listening to and so advising the patient, the physician must be cold, unfeeling, unempathic, or tactless. Practitioners of all professions differ in their personal dispositions and talents of empathy, compassion, warmth, and tactfulness, and it is the task of selection and training to prevent cold, unfeeling, or hostile persons from going into the helping professions.

Second, the I–thou relationship objection has a spurious appearance of humaneness but is in fact inhumane. When, to use traditional theological language, does one have *caritas* (love, in the moral, not romantic, sense) toward another? One need not be a member of the Roman church to agree with Thomas Aquinas that *caritas* consists of willing a person's good. To an empowered predictor, this means making what is more

likely to be the best decision. If a certain mode of data combination has been clearly shown to be more efficient (i.e., more likely to benefit the person), but the practitioner opts for a less efficient one because it fosters a pleasant, warm, cozy feeling of an I–thou relationship, this has the effect of treating the other person as a means to the practitioner's sentimental, emotional ends. It violates one of Kant's formulations of the categorical imperative, that we should treat persons as ends, not as means. Aquinas wrote in his *Summa Theologica*, "Accordingly, it is evident that [charity and justice] are not in the sensitive appetite, where passions are found, but in the rational appetite—the will—where there are no passions" (translation by Goodwin, 1965, p. 85 [Article 5]). This antistatistical argument is especially offensive because it commits an immorality behind a moral mask.

"*The Studies Included Naive Clinicians: If the Clinicians Were Properly Informed About Their Mistakes (Where the Actuarial Formula Was Correct), Then in the Future They Would Beat the Formula*"

A number of studies provided the clinicians with feedback, and the evidence is conflicting as to whether it helped and how much. For example, in Goldberg's (1965) study, judges were given immediate feedback on the accuracy of their judgments, for a total of 861 trials. However, this massive opportunity to learn better judgment practices did not result in clinicians' doing nearly as well as a four-variable equally weighted regression equation. The evidence to date is not encouraging and surely does not warrant the confident dismissal of actuarial prediction on the basis of hope. This is a quantitative matter that will probably vary over prediction domains and with the general level of clinician education and experience. The meta-analysis suggests that when feedback does produce improvement, it only moves the less accurate clinicians closer to the (naively) better ones, but it does not enable the latter to surpass the formula. Even if this alleged effect were stronger and more consistent than the studies show it to be, the pragmatic context makes this a finding that is not useful. Most practitioners in most settings are in fact naive in this sense, and the absence of adequate feedback (such as occurs for physicians in the clinicopathological conference) is part of the reason why clinicians do not do better than the formula, or as well as they assume they do. Finally, this complaint subtly applies a double standard to the clinician and the actuary. Suppose "naive" clinicians can sometimes be effectively transformed into "sophisticated" clinicians by feedback on their errors, including information as to how these errors are related to the correct predictions of the formula. Such analytic investigations of subsets of decisions will enable the actuary to improve the statistical formula as well. In comparing two procedures, one cannot rationally or fairly say that the procedure one prefers (clinical judgment) could be improved by educational measures not currently practiced but that the competitor (statistician) is required to remain naive with respect to the formula's mistakes. What epistemological or mathematical argument is offered to show that the clinician is improvable but the formula is not? We have seen none.

Given that both modes of prediction could benefit from feedback, an interesting asymmetry arises that favors the actuary. Qualitative insight gained from education and feedback never guarantees that the clinician will reliably apply what has been learned and assign optimal weights. If an item is, in reality, predictive—and it must be shown to be so by the analysis of disaggregated subsets of predictions by both parties—then the

statistician, no longer naive, can include it in the actuarial equation. The improved equation proceeds consistently and reliably (except for clerical error) and with weights that are certain to be closer to optimal than the subjective weights of the clinician.

"*The Assessment Process Is Not Predictive Because Predictions Are Often Inexplicit—The Goal Is to Understand, Not to Predict*"

This is an immoral argument. The practitioner is spending the taxpayer's or the patient's or the insurance company's money to make decisions about mentally ill people, law school applicants, or bank robbers and meanwhile is putting scholarly thirst for alleged understanding ahead of the institutionally defined pragmatic task. Whether intellectually satisfying comprehension facilitates making the best decisions is, of course, an empirical question, and that is what the comparative studies are about. It does not appear that some surplus understanding over and above those components of diagnosis that have actuarial predictive value accomplishes much of anything.

"*The Actuarial Method Uses Probabilities, Which Are Only Estimates, Not Exact*"

This is surely correct, but hardly relevant when the subjective (informal, impressionistic, in-the-clinician's-head) probabilities are equally inexact, usually more so. If a certain diagnostic sign predicts an event with probability of .40 on the actuarial data, the true probability for the whole reference class might be .43 or .37. Random sampling variations due to chance (as distinguished from bias or validity generalization to a different population) affect beta weights, proportions, and actuarial table tallies, but those random factors in a given clinical population exert precisely the same bad influence, in the purely statistical sense, on the clinician's cerebral memory bank. Thus, even if the clinician had optimal weights and used them consistently, this argument is two-edged and speaks with equal force against both methods. Again, the proof of the pudding is in the eating, and we must look to the studies to see how serious a deficiency this is.

"*The Studies Do Not Vary Sufficiently Over Predictive Domains to Make Any Generalization*"

This is simply false as a characterization of the research literature. The predictands include such widely varied ones as cancer recovery, parole violation, college grades, psychiatric nosology, nurses' fear of mental patients, kinds of jaundice, response to shock therapy, air crew training survival, business failures, and winning football games. If one argues that although the range of predictive tasks is wide, it is still not sufficient to make a generalization, a double standard of methodological morals is again imposed. This would make all summaries of research literature in all areas of social science uninterpretable. Given the massive, varied, and almost wholly consistent results, greater than for any other controversy in social science, one who advances this objection has an obligation to specify the predictive domain for which the informal method's superiority is claimed and then to do legitimate empirical comparisons.

"*Mathematics Assumes That the World Is Completely Orderly, Rigid, and Deterministic, Which It Is Not*"

The branch of mathematics that is relevant here—the statistical method—is explicitly probabilistic. If all events were certainly deterministic in the nomological sense assumed by classical mechanics and if we always had all that information available, the science of statistics would be a branch of pure mathematics and of no application to human medical or social problems. It is precisely when events are indeterministic, unpredictable, "chancy" that the probability calculus and its offspring, theoretical and applied statistics, acquire the great importance they have in the life sciences (and often even in chemistry and physics). If we had some way of knowing for sure that strict sociopsychological laws absolutely determined that Jones would rob another bank if released from jail, we would not be fooling around with discriminant functions or actuarial tables.

"*The World Changes All the Time, so Any Statistical Formula Will Quickly Become Out-Of-Date*"

The quantitative importance of this qualitative truism is an empirical question, not to be settled by armchair prejudices. A periodic recheck of a formula or table is of course welcome and if several years have passed, it would be strongly urged. We know of no empirical metageneralization on this subject that says how much time must elapse in a given kind of population of students, patients, offenders, or job applicants before regression weights become seriously altered, and reliance on the robustness of minimally or equally weighted predictors reduces the force of this argument to a very weak one. Here again, we have a double standard of morals, because it is assumed that the changes that take place in the world will not also begin to trip up the clinician, who is relying upon an informal computational system the inexplicit weights of which are a product of past training and experiences. Of course, if there is some major social change (e.g., in the laws regarding probation or in the availability of intensive psychotherapy) and there are good theoretical reasons for expecting that change to affect the formula's accuracy, it is necessary to perform appropriate empirical studies and update the actuarial procedure.

Explanation of Why Actuarial Prediction Works Better Than Clinical

What is the explanation for the statistical method being almost always equal or superior in accuracy to the informal, impressionistic, clinical method of data combination? Space does not permit more than a summary statement here; for more extensive treatment by clinical, social, and cognitive psychologists, see, for example, Arkes and Hammond (1986); Dawes (1988); Faust (1984); Hogarth (1987); Kahneman, et al. (1982); Meehl (1954/1996); Nisbett and Ross (1980); and Plous (1993); for a listing of sources of error in clinical judgment, see Meehl (1992a, pp. 353-354). Assume that the clinician does not usually (except, e.g., Freud) attempt to concoct an idiographic mini theory of an individual's psyche and the environmental forces that are likely to act upon that person but simply attempts to do a subjective, impressionistic, in-the-head approximating job of actuarial computation. Then the clinician's brain is functioning as merely a poor substitute for an explicit regression equation or actuarial table. Humans simply cannot assign optimal weights to variables, and they are not consistent in applying their own weights.

The influence of unreliable data combination by informal judgment is dramatically illustrated by the Goldberg paradox. Goldberg (1970) used 29 clinicians' ratings of profiles on the MMPI for psychosis versus neurosis. First using each clinician's ratings as

the predictand (rather than the external criterion of psychiatric diagnosis), Goldberg then found that these strangely derived multiple regression equations predicted the external criterion more accurately than the clinicians [did]; this was true for each clinician. The explanation of this counterintuitive result lies in rater unreliability; the clinicians' subjective regression weights, though nonoptimal, do better than the clinicians themselves, because they do not apply their own weights consistently. The Goldberg paradox, though numerically small, is robust, having been replicated in 15 studies on a variety of predictive tasks (Camerer, 1981). The paradox is less interesting than it seems if one accepts the generalization of Dawes and Corrigan (1974) that randomly chosen weights perform as well as those modeling the clinician's judgments. We do not know whether Goldberg's clinician-based weights would out-perform an unweighted composite.

The human brain is a relatively inefficient device for noticing, selecting, categorizing, recording, retaining, retrieving, and manipulating information for inferential purposes. Why should we be surprised at this? From a historical viewpoint the superiority of formal, actuarially-based procedures seems obvious, almost trivial. The dazzling achievements of Western post-Galilean science are attributable not to our having any better brains than Aristotle or Aquinas, but to the scientific method of accumulating objective knowledge. A very few strict rules (e.g., don't fake data, avoid parallax in reading a dial) but mostly rough "guidelines" about observing, sampling, recording, calculating, and so forth sufficed to create this amazing social machine for producing valid knowledge. Scientists record observations at the time rather than rely on unaided memory. Precise instruments are substituted for the human eye, ear, nose, and fingertips whenever these latter are unreliable. Powerful formalisms (trigonometry, calculus, probability theory, matrix algebra) are used to move from one set of numerical values to another. Even simple theories can now be derived by search algorithms (e.g., Langley, Simon, Bradshaw, & Zytkow, 1987; Shrager & Langley, 1990), although inventing rich theories postulating theoretical entities interacting in complex ways are as yet a uniquely human mind task. However theories are concocted, whether appraisal of their empirical merits is best conducted informally, as presently (except in meta-analysis, cf. Glass, McGaw, & Smith, 1981; Hunter, Schmidt, & Jackson, 1982), is not known and has been forcefully challenged (Faust, 1984; Faust & Meehl, 1992; Meehl, 1990a, 1990e, 1992a, 1992c). However, we need not look to science for the basic point to be made, as it holds—and is universally accepted, taken for granted—in most areas of daily life.

> Surely we all know that the human brain is poor at weighting and computing. When you check out at a supermarket, you don't eyeball the heap of purchases and say to the clerk, "Well it looks to me as if it's about $17.00 worth; what do you think?" The clerk adds it up. (Meehl, 1986a, p. 372)

This everyday example also casts commonsensical doubt on the antiactuarial claim that formal procedures will only work better for linear prediction functions but that the clinician's informal mode of data combination is needed when the true function is nonlinear and, especially, configural (cf. Meehl, 1954/1996, pp. 131-135, for a definition of patterning). Suppose the supermarket made use of a nonlinear and configural combination rule for the commodity basket, such as "add the logarithm of the vegetable price to half the product of hamburger and kitty litter prices"; would this complication lead us to prefer subjective eyeballing? Of course not.

While acknowledging that they do not function as well as even a second-rate desk calculator, clinicians may believe that they can usually formulate a correct idiographic (structural and dynamic) mini theory of the patient and can employ the laws of the mind to reach predictions on this mini theory. This has a certain plausibility for the advanced sciences such as astronomy or chemistry, but the analogy in the social sciences is grossly incorrect. In order to mediate predictions accurately by causal theories (that is, via attribution of particular states of affairs within a framework of causal laws), one must have (a) a fairly complete and well-supported theory, (b) access to the relevant variables that enter the equations of that theory, and (c) instruments that provide accurate measures of those variables. No social science meets any of these three conditions. Of course, the actuarial method also lacks adequate knowledge of the events and social pressures to which the person may be exposed during the time span for which prediction is made. However, the actuarial method has the distinct advantage that the statistics have already discounted the collective influence of all of these unknown factors (which is why a multiple correlation may be .75 instead of .95). These unknown and unpredictable events and forces, called "contingency factors" by Horst (1941), must be assigned values when we try to mediate predictions by a causal theory, whereas they are all part of the error variance in the actuarial method and their collective influence is given the weight that it deserves, as shown by the actuarial data.

Why Do Practitioners Continue to Resist Actuarial Prediction?

Readers unfamiliar with this controversy may be puzzled that, despite the theoretical arguments from epistemology and mathematics and the empirical results, the proactuarial position is apparently held by only a minority of practitioners. How is it possible that thousands of MDs, PhDs, and MSWs, licensed to practice in their jurisdictions, and even academics teaching in clinical training programs, could be so wrong, as we allege? Having answered their objections on the merits, we think it not arguing ad hominem or committing the genetic fallacy to suggest some sociopsychological factors that may help to explain this remarkable resistance to argument and evidence.

Fear of technological unemployment. If one of 20 social workers engaged in writing presentence investigation reports is told that 18 could be dispensed with and that the other two, supervised by a PhD-level psychologist or statistician, could do as well or better in advising the court judges, then that is cause for concern.

Self concept. Income aside, most professionals have a self-image and a personal security system that are intimately tied in with the value that they and society place on their scholarly and technical functions. As an analogy, consider how unhappy senior partners in a law firm would be, even if assured of their jobs, to learn that paralegals with a few years of experience could predict the opinions of an appellate court as accurately as a partner can.

Attachment to theory. Most researchers and clinicians have a fondness for certain concepts and theories, and the idea that our theory-mediated predictions do not contribute anything predictively beyond what an atheoretical actuarial table could or that the theory may even make matters worse produces cognitive dissonance. Most intellectuals, whether practitioners or not, take concepts and theories seriously.

Misperception of the actuarial method as dehumanizing to clients or patients. The objection of the actuarial method as being dehumanizing has been dealt with above.

General dislike of computers' successfully competing with human minds. Personal ego involvement and employment aside, many persons seem to have some diffuse resentment toward the very idea that a computer can duplicate human cognitive performance. Thus, for instance, that computer chess programs are now able to defeat a few grand masters sometimes bothers people who are not themselves chess masters. For some reason, people just do not like the idea that a mere machine can do better than a person at any cognitive task.

Poor education. Poor education is probably the biggest single factor responsible for resistance to actuarial prediction; it does not involve imputation of any special emotional bias or feeling of personal threat. In the majority of training programs in clinical psychology, and it is surely as bad or worse in psychiatry and social work, no great value is placed upon the cultivation of skeptical, scientific habits of thought; the role models—even in the academy, more so in the clinical settings—are often people who do not put a high value upon scientific thinking, are not themselves engaged in scientific research, and take it for granted that clinical experience is sufficient to prove whatever they want to believe. There are probably not more than two dozen American psychology departments whose clinical training programs strongly emphasize the necessity for scientific proof, either as experiments or statistical study of file data, as the only firm foundation for knowledge. As a sheer matter of information, many psychologists, psychiatrists, and social workers are literally unaware that any controversy about the merits of prediction procedure exist or that any empirical comparisons of the two methods have been performed. The common position is, "Well, of course, a deep psychological understanding will enable a clinician to predict an individual's future behavior better than any generic mathematical equation possibly could." Even if motivational forces were absent (and they are hardly likely to be totally absent in any of us who engage in clinical work), inadequate scientific education would be more than sufficient to account for the compact majority being in error.

If this is a shocking deprecation of typical doctoral education, we invite sophisticated readers to reflect on the intellectual quality of the 17 antistatistical arguments rebutted above. A few are plausible, involving interesting epistemological, mathematical, or unresearched factual questions (e.g., "broken leg" cases, generalizing weights, defining a reference class, Windelband's dichotomy), but a large majority are confused, uninformed, or tendentious (double standard).

Conclusions and Policy Implications

We know of no social science controversy for which the empirical studies are so numerous, varied, and consistent as this one. Antistatistical clinicians persist in making what Dawes (1994, pp. 25, 30, 96) calls the "vacuum argument," in which (imagined, hoped-for) supportive evidence is simply hypothesized, whereas negative evidence that has actually been collected is ignored. For example, "But clinicians differ; some are better than others." Reply: "True, but even the best ones don't excel the equation." "But, even the best ones were naive; they should have feedback so as to improve their performance." Reply: "The effectiveness of feedback is not a robust finding and is small." "But, they were not given the right kind of feedback," and so forth. One observes a series of tactical retreats, reformulations, ad hoc explanations, coupled with a complacent assurance that if the "right sort" of study were done, things would turn out differently.

This sublime confidence in the yet-to-be-done super study persists despite the social fact that many of the published investigators (including Meehl, 1959a, trying to implement Meehl, 1957b) were motivated to come up with a good antiactuarial result. When we have 136 interpretable studies with only 5% deviant, ranging over a wide diversity of predictands (e.g., winning football games, business failures, response to shock therapy, parole violation, success in military training), it is time to draw a conclusion "until further notice," the more so as the facts are in accord with strong theoretical expectations. One must classify continued rejection (or disregard) of the proactuarial generalization as clear instances of resistance to scientific discovery (Barber, 1961), or, more generally, as exemplifying H. L. Mencken's dictum that most people believe what they want to believe. This seems a harsh but warranted judgment. Given that harsh judgment, scholarly justice requires us to note that the distinguished clinical psychologist Robert Holt, Meehl's friendly critic for 40 years, has, in his latest publication on this topic, explicitly conceded the point originally at issue. He writes,

> My main quarrel with Paul Meehl is that he did not see that I was trying to mediate, or did not agree at all about the ways I wanted to change the focus, and persisted in charging through what looks to me like an open door. Maybe there are still lots of clinicians who believe that they can predict anything better than a suitably programmed computer; if so, I agree that it is not only foolish but at times unethical of them to do so…. If I ever accused him or Ted Sarbin of "fomenting the controversy," I am glad to withdraw any implication that either deliberately stirred up trouble, which I surely did not intend. (Holt, 1986a, p. 378)

From a theoretical viewpoint the issue may be rather uninteresting, because it is trivial. Given an encodable set of data—including such first-order inferences as skilled clinicians' ratings on single traits from a diagnostic interview—there exists an optimal formal procedure (actuarial table, regression equation, linear, nonlinear, configural, etc.) for inferring any prespecified predictand. This formula, fallible but best (for a specific clinical population), is known to Omniscient Jones but not to the statistician or clinician. However, the statistician is sure to approximate it better, if the job is done properly. If the empirical comparisons had consistently favored informal judgment, we would have considerable explaining to do. Yet the empirical comparisons were necessary, as we see from the widespread inability to accept them despite their metapredictability from mathematics and cognitive science.

The policy implications of the research findings are obvious. Two main theses emerge from the empirical conclusion. First, policy makers should not accept a practitioner's unsupported allegation that something works when the only warrant for this claim is purported clinical experience. Clinical experience is an invaluable source of ideas. It is also the only way that a practitioner can acquire certain behavioral skills, such as how to ask questions of the client. It is not an adequate method for settling disputes between practitioners, because they each appeal to their own clinical experience. Histories of medicine teach us that until around 1890, most of the things physicians did to patients were either useless or actively harmful. Bleeding, purging, and blistering were standard procedures, as well as prescribing various drugs which did nothing. In 1487, two Dominican monks, Kraemer and Sprenger (1487/1970), published a huge treatise, *Malleus Maleficarum*, that gave details on how to reach a valid diagnosis of a witch. It is estimated that more than 100,000 persons were hanged, burned alive, drowned, or

crushed with stones as witches; the basis for the detailed technical indications in that book was the clinical experience of inquisitors. All policy makers should know that a practitioner who claims not to need any statistical or experimental studies but relies solely on clinical experience as adequate justification, by that very claim is shown to be a nonscientifically minded person whose professional judgments are not to be trusted (cf. Meehl, 1997a). Further, when large amounts of taxpayer money are expended on personnel who employ unvalidated procedures (e.g., the millions of dollars spent on useless presentence investigation reports), even a united front presented by the profession involved should be given no weight in the absence of adequate scientific research to show that they can do what they claim to do.

Regardless of whether one views the issue as theoretically interesting, it cannot be dismissed as pragmatically unimportant. Every single day many thousands of predictions are made by parole boards, deans' admission committees, psychiatric teams, and juries hearing civil and criminal cases. Students' and soldiers' career aspirations, job applicants' hopes, freedom of convicted felons or risk to future victims, millions of taxpayer dollars expended by court services, hundreds of millions involved in individual and class action lawsuits for alleged brain impairment (Faust, Ziskin, & Hiers, 1991; Guilmette, Faust, Hart, & Arkes, 1990), and so forth—these are high stakes indeed. To use the less efficient of two prediction procedures in dealing with such matters is not only unscientific and irrational, it is unethical. To say that the clinical–statistical issue is of little importance is preposterous.

References

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt. 167

Arkes, H. R., & Hammond, K. R. (1986). *Judgment and decision making: An interdisciplinary reader*. New York: Cambridge University Press.

Barber, B. (1961, Septermber 1). Resistance by scientists to scientific discovery. *Science*, *134*, 596-602.

Bloch, D. A., & Moses, L. E. (1988). Nonoptimally weighted least squares. *American Statistician*, *42*, 50-53.

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* (pp. 205-249). Springfield, IL: Illinois Committee on Indeterminate-Sentence Law and Parole.

Burt, C. (1950). The influence of differential weighting. *British Journal of Psychology, Statistical Section*, *3*, 105-123.

Butcher, J. N. (1986). *Users guide for the Minnesota Clinical Report*. Minneapolis: National Computer Systems.

Butler, J. (1736). *The analogy of religion*. London: Printed by J. Jones for George Ewing.

Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance*, *27*, 411-422.

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

Dawes, R. M. (1988). *Rational choice in an uncertain world*. Chicago, IL: Harcourt Brace Jovanovich.

Dawes, R. M. (1994). *House of cards*. New York: Free Press.

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

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668-1674.

Dawes, R. M., Faust, D., & Meehl, P. E. (1993). Statistical prediction versus clinical prediction: Improving what works. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 351-367). Hillsdale, NJ: Erlbaum.

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment*, *50*, 387-395.

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

Faust, D. (1984). *The limits of scientific reasoning*. Minneapolis, MN: University of Minnesota Press.

Faust, D. (1991). What if we had listened? Present reflections on altered pasts. In W. M. Grove and D. Cicchetti (Eds.), *Thinking clearly about psychology. Vol. 1: Matters of public interest* (pp. 185-216). Minneapolis, MN: University of Minnesota Press.

Faust, D., & Meehl, P. E. (1992). Using scientific methods to resolve enduring questions within the history and philosophy of science: Some illustrations. *Behavior Therapy*, *23*, 195-211.

Faust, D., Ziskin, J., & Hiers, J. B. (1991). *Brain damage claims: Coping with neuropsychological evidence* (2 vols.). Los Angeles, CA: Law & Psychology Press.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, *105*, 387-396.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

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

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

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

Goldberg, L. R. (1970). Human mind versus regression equation: Five contrasts. In W. M. Grove and D. Cicchetti (Eds.), *Thinking clearly about psychology. Vol. 1: Matters of public interest* (pp. 173-184). Minneapolis, MN: University of Minnesota Press.

Goodwin, R. P. (1965). *Selected writings of St. Thomas Aquinas*. New York: Macmillan.

Gough, H. G. (1962). Clinical versus statistical prediction in psychology. In L. Postman (Ed.), *Psychology in the making* (pp. 526-584). New York: Knopf.

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

Guilmette, T. J., Faust, D., Hart, K., & Arkes, H. R. (1990). A national survey of psychologists who offer neuropsychological services. *Archives of Clinical Neuropsychology*, *5*, 373-392.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley & Sons. 167

Hogarth, R. M. (1987). *Judgement and choice: The psychology of decision*. New York: Wiley.

Holt, R. R. (1978). *Methods in clinical psychology. Vol. 2: Prediction and Research*. New York: Plenum Press.

Holt, R. R. (1986). Clinical and statistical prediction: A retrospective and would-be integrative perspective. *Journal of Personality Assessment*, *50*, 376-386.

Horst, P. (Ed.). (1941). *Prediction of personal adjustment*. (Bulletin No. 48). New York: Social Sciences Research Council.

Hovey, H. B., & Stauffacher, J. C. (1953). Intuitive versus objective prediction from a test. *Journal of Clinical Psychology*, *9*, 349-351.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Jevons, W. S. (1958). *The principles of science*. New York: Dover. (Original publication 1874)

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgments under uncertainty: euristics and biases*. Cambridge, Eng.: Cambridge University Press.

Kendall, M. G. (1949). On the reconciliation of theories of probability. *Biometrika*, *36*, 101-116.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin*, *107*, 296-310.

Kraemer, H., & Sprenger, J. (1970). *Malleus Malaeficarum*. New York: Blom. (Original work published 1487)

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1990). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Laughlin, J. E. (1978). Comment on "Estimating coefficients in linear models: It don't make no nevermind." *Psychological Bulletin*, *85*, 247-253.

Linstone, H. A., & Turoff, M. (Eds.). (1975). *The Delphi method: Techniques and applications*. Reading, MA: Addison-Wesley.

Marchese, M. C. (1992). Clinical versus actuarial prediction: A review of the literature. *Perceptual and Motor Skills*, *75*, 583-594.

McCormack, R. L. (1956). A criticism of studies comparing item-weighting methods. *Journal of Applied Psychology*, *40*, 343-344.

McNemar, Q. (1955). [Review of the book *Clinical versus actuarial prediction*]. *American Journal of Psychology*, *68*, 510.

Meehl, P. E. (1956a). Clinical versus actuarial prediction. In *Proceedings of the 1955 Invitational Conference on Testing Problems* (pp. 136-141). Princeton: Educational Testing Service.
Meehl, P. E. (1956b). Symposium on clinical and statistical prediction (with C. C. McArthur & D. V. Tiedeman). *Journal of Counseling Psychology*, *3*, 163-173.
Meehl, P. E. (1956c). Wanted—a good cookbook. *American Psychologist*, *11*, 263-272.
Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology*, *4*, 268-273.
Meehl, P. E. (1959). A comparison of clinicians with five statistical methods of identifying MMPI profiles. *Journal of Counseling Psychology*, *6*, 102-109.
Meehl, P. E. (1965). Seer over sign: The first good example. *Journal of Experimental Research in Personality*, *1*, 27-32.
Meehl, P. E. (1967). What can the clinician do well? In D. N. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 594-599). New York: McGraw-Hill.
Meehl, P. E. (1970). Nuisance variables and the ex post facto design. In M. Radner & S. Winokur (Eds.), *Minnesota studies in the philosophy of science: Vol. IV. Analyses of theories and methods of physics and psychology* (pp. 373–402). Minneapolis, MN: University of Minnesota Press.
Meehl, P. E. (1973). *Psychodiagnosis: Selected papers*. Minneapolis, MN: University of Minnesota Press.
Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806-834.
Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment*, *50*, 370-375.
Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant using it. *Psychological Inquiry*, *1*, 108-141, 173-180.
Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195-244. In R. E. Snow & D. Wiley (Eds.), *Improving Inquiry in social science* (pp. 13-59). Hillsdale, NJ: Erlbaum, 1991.
Meehl, P. E. (1992a). Cliometric metatheory: The actuarial approach to empirical, history-based philosophy of science. *Psychological Reports*, *71*, 339-467.
Meehl, P. E. (1992b). The Miracle Argument for realism: An important lesson to be learned by generalizing from Carrier's counter-examples. *Studies in History and Philosophy of Science*, *23*, 267-282.
Meehl, P. E. (1996). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Northvale, NJ: Jason Aronson. (Original work published 1954)
Meehl, P. E. (1997). Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice*, *4*, 91-98.
Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of human judgment*. Englewood Cliffs, NJ: Prentice-Hall.
Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.
Reichenbach, H. (1938). *Experience and prediction*. Chicago, IL: University of Chicago Press.
Richardson, M. W. (1941). The combination of measures. In P. Horst, *Prediction of personal adjustment* (Bulletin No. 48; pp. 377-401). New York : Social Sciences Research Council.
Sarbin, T. R. (1943). A contribution to the study of actuarial and individual methods of prediction. *American Journal of Sociology*, *48*, 593-602.
Sarbin, T. R. (1986). Prediction and clinical inference: Forty years later. *Journal of Personality Assessment*, *50*, 362-369.
Sawyer, J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin*, *66*, 178-200.
Shrager, J., & Langley, P. (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufman.
Sines, J. O. (1970). Actuarial versus clinical prediction in psychopathology. *British Journal of Psychiatry*, *116*, 129-144.
Thompson, R. E. (1952). A validation of the Glueck Social Prediction Scale for proness to delinquency. *Journal of Criminal Law and Police Science*, *43*, 451-470.
Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago, IL: University of Chicago Press.
Tukey, J. W. (1948). Approximate weights. *Annals of Mathematical Statistics*, *19*, 91-92.
Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 213-217.
Wainer, H. (1978). On the sensitivity of regression and regressors. *Psychological Bulletin*, *85*, 267-273.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, *3*, 23-40.
Wittman, M. P. (1941). A scale for measuring prognosis in schizophrenic patients. *Elgin Papers*, *4*, 20-33.