Behavior and Philosophy, *33*, 91-99 (2005).

CLINICAL, STATISTICAL, AND BROKEN-LEG PREDICTIONS

Kurt Salzinger
Hofstra University

ABSTRACT: Accurate prediction of behavior is a critical task for the psychologist, particularly for the practitioner. Outstanding among those who have successfully wrestled with this complicated task is Paul Meehl. Yet, small has been the influence of his work on the everyday practice of prediction. The object of this paper is to review Meehl's work in this area and, using behavior analysis, to seek an understanding of practitioners' continued opposition to his findings.
Key words: prediction of behavior, behavior analysis, clinical judgment, statistical prediction, actuarial prediction, clinical prediction

More than 50 years after Paul Meehl's (1954) publication of, as he later called it, "my disturbing little book" (Meehl, 1986), relatively few clinicians in the field have accepted his findings. Instead, they have continued to construct and repeat reasons for avoidance of his assertions. It is, therefore, wholly appropriate that we consider the reasons given for denying his, and over the years, many other investigators' assertions that statistical or mechanical devices are superior (or at the very least, are more accurate, cheaper, take less time, and require less expert knowledge) to clinical ways of predicting the behavior of individuals.

Let us begin by first reviewing what Meehl's assertions are, including where he believes there is a place for clinical judgment. Basically, Meehl, who was both a clinician and a statistician, both a psychoanalyst and an experimental psychologist, tried to make the point that although there are areas of functioning in which clinicians do best, there are other areas best left to what Meehl called clerks who tallied data and allowed equations, however simple, to combine those tallies of data in a mechanical way. Meehl's argument is based on the assumption that there are universal behavioral laws. He speaks of Skinner's approach, which respects the uniqueness of each laboratory rat while demonstrating the universal laws of operant conditioning. Meehl also cites the concept of response class to allow us to speak about the probability of behavior. Thus, if one describes a person as aggressive, one means that he or she emits responses all of which are members of the response class of aggressive responses, e.g., pounds on the table, kicks a chair, curses, hits another person, kicks a dog. The clinician then makes use of his or her estimate (consciously or unconsciously) of the frequency with which the person in question emits aggressive responses. When measures of these frequencies of

AUTHOR'S NOTE: Please address all correspondence to Kurt Salzinger, Ph.D., Senior Scholar in Residence, Department of Psychology, Hofstra University, Hempstead, NY 11549; Email: Kurt.Salzinger@hofstra.edu

response classes exist, the actuary can then combine them in a mechanical manner, whereas the clinician combines them and other observations, as well as interpretations of the observations "in his head." Although actual counts are superior, numerical estimates of frequency (i.e., ratings) could also be used in an equation. It is the manner of combination of observations that is at the core of the argument between clinical and actuarial approaches.

Meehl (1986) provides a wonderful example that argues for actuarial prediction. Imagine yourself checking out at the supermarket, he suggests; would you survey what you have in your basket and estimate the cost? Would the clerk do that? Obviously there is a less "creative" solution to the problem. The clerk adds it all up; in fact, with current electronic devices, he or she simply passes the items over a sensing device that lights up the small screen with the cost of each item and eventually adds up the entire set of item-costs. Can we do that with behavior? That is the crux of the argument. Should we combine our observations in our heads or should we resort to mechanical means of combination? Meehl freely admits that there are other kinds of clinician activities, such as those of hypothesis generation and theory creation, that cannot be given over to a clerk in the same way as can the observations stemming from psychometric tests, tallies of occurrence of various classes of behaviors, or ratings generated by clinicians. The former (theory creation and hypothesis generation) have to do with the context of discovery, for which tallying is of no value; it is only in the context of justification that we can talk about frequencies and reliability, of calculation and validity.

Before we go any further, let us review some of the arguments against the use of the actuarial method and the answers given to those objections when they are not obvious (Grove & Meehl, 1996; Meehl, 1986):

1. Most clinical psychologists are not acquainted with the relevant statistics or mathematics. Neither needs to be fully understood, however, to make use of actuarial tables to take but one example of statistical prediction.

2. The threat of being displaced by a clerk is unwelcome. The possibility of losing income is obviously frightening. In fact, however, having a clerk take over some of the clinician's duties would release the expert's time to do what he or she is best equipped to do, namely clinical observation and creation of hypotheses and theories to explain the behavior that is statistically combined.

3. The view that "I do not engage in that kind of dehumanizing activity; I am a Freudian" is hard to answer because it arbitrarily limits one's capacity to be helpful to others. Not trying out a method that might turn out to be useful is truly self-defeating.

4. Some insist that they use both the clinical and actuarial method of prediction. However, this cannot always be true since at least some of the time a committee might vote (on a "clinical" basis) for admission of a student, or release of a prisoner, or application of an intervention, whereas the equation might predict a different course for each of these recommendations. One must, in the case of such contradictions, come down on one side or another.

5. Some critics mistakenly believe that the actuarial approach requires psychometric scores and denies the usefulness of other behavioral descriptions.

Any behavioral observation can, however, be quantified, even if only by designation of presence or absence of a characteristic such as "sociable" or "motivated." What the actuarial technique contributes is a way of combining disparate data sets. It is apparently such clinical combination that we fail to do effectively.

6. Another objection states that actuarial tables may not be available, but more often than not some form of table is available at another clinic or school, or one can be constructed from data collected by one's own institution.

7. Actuarial prediction is falsely interpreted to refer to groups only. Here Meehl's wonderful example of being forced to decide which of two guns to use when placed against one's forehead is relevant. Suppose a madman forced you make a choice: He will aim a gun at your head containing five bullets out of six chambers or a gun containing but one bullet in the six chambers. Even though only one person will yank the trigger one time, wouldn't you base your decision of the smaller chance of encountering a bullet from the one-bullet gun?

8. Related to number 7 above is the notion that a given individual may have a particular characteristic that makes the equation or actuarial table inappropriate. The so-called "broken leg" phenomenon is invoked when you are trying to predict whether a particular man is going to see a particular movie. All the information that you have about his predilections may still lead you to the wrong prediction because he has a broken leg and therefore simply cannot go to the movies at all. (You have to also assume that there is no wheelchair around and that movie houses do not provide special places for wheelchairs, as they do nowadays.) This exception must be answered by pointing out that psychology does not have many "broken legs" and that one should invoke such an exception sparingly in view of the data contradicting such phenomena. In other words, there are very few conditions that trump universal generalizations. Nevertheless, such special conditions ought to be investigated to determine how often they do occur and what might be a way of defining a "broken leg" in psychology.

9. Some critics of actuarial methods have suggested that the factors that help to make predictions change over time, requiring that formulas be modified to accommodate those changes. That, of course, is an empirical question and one would think would apply equally to clinical predictions.

The debate of clinical versus actuarial prediction, which had been confined purely to academic journals, eventually broke out into the open, so to speak. Faust and Ziskin (1988a) wrote an article for *Science* magazine that constituted a kind of prolegomenon to the article one year later by Dawes, Faust, and Meehl (1989). The former article cited Meehl's (1954) book and discussed the difficulty that human beings have of combining information in a reliable and accurate way, but it concentrated on the general problem of clinical judgment and its usefulness when expressed by so-called experts in the courts. The authors concluded on the basis of many studies that actuarial procedures are usually better than the clinical judgment, and thus (when applicable in a particular situation) make "the expert's involvement in the interpretive process. . .unnecessary" (p. 33). Giving no quarter, they went on to maintain not only that clinical judgment is usually inferior to that of actuarial

prediction, and they insisted that "attempts to 'refine' or modify actuarial conclusions [by clinical judgment] produce inferior overall results" (p. 34).

It did not take long for Faust and Ziskin's (1988a) paper to elicit official responses from both the American Psychological Association (Fowler & Matarazzo, 1988) and from the American Psychiatric Association (Spitzer, Williams, & Pincus, 1988). Fowler and Matarazzo argued that Faust and Ziskin's (1988a) fault-finding should result in a cautious approach to the use of expert witnesses, not in the elimination of the practice. Faust and Ziskin (1988b) responded that although as representatives of the APA the writers wish to continue to rely on expert witnesses, in their scientific writings they appeared to share the views held by Faust and Ziskin. Spitzer and colleagues (1988) responded to Faust and Ziskin by defending psychiatric diagnosis. They pointed out that it had been significantly improved; besides, they stated, medicine in general does not provide perfect reliability either. As you might imagine, Faust and Ziskin (1988c) rejoined by arguing that low reliability in other parts of medicine does not excuse or improve low reliability of psychiatric diagnosis. Furthermore, many other flaws in the diagnostic process, such as its lack of validity, do not exactly provide reason for having psychiatrists serve as expert witnesses in the courts.

All of this was followed by Dawes et. al. (1989). Using examples from three different areas of prediction studies, the authors found actuarial prediction to prevail over clinical prediction every time. Goldberg's (1968) MMPI study in which he used a simple formula to combine scale results produced greater validity than clinicians' attempts to combine the scales of that test intuitively. A second study reviewing progressive brain dysfunction measured by intellectual testing scales showed that the actuarial combination achieved greater diagnostic accuracy than the intuitive combination created by a group of clinicians. A third study compared prediction of survival time based on combining nine histological dimensions of biopsy slides of patients who had Hodgkin's disease. When the same clinicians who had produced the ratings did the combination, it was inferior to the combination achieved by actuarial means. The article went on to provide other studies that demonstrated the superiority of actuarial over clinical combination of various measurements. Unlike other papers written almost exclusively about psychological measurements in combination, Dawes and his colleagues succeeded in showing the superiority of actuarial over clinical combination in other fields as well.

Kleinmuntz (1990) argued, like others before him, that it is not a case of clinical versus actuarial prediction but rather a combination of both. Faust, Meehl, and Dawes (1990) responded again that the evidence is that the combination is generally worse than the actuarial approach, allowing only psychoanalytic combination during the therapeutic hour as an exception. Interestingly enough, Zubin (1955) also argued for combination of clinical and actuarial approaches shortly after Meehl's book came out:

> The two types of prediction supplement each other and the discrepancies between the two should be studied for improving each other reciprocally. Meehl

> has pointed out that behind the clinician looms the shadow of the actuary and that the latter like the undertaker will have the last word. I doubt this. For behind this actuary is another clinician looking over his shoulder to see just where the formula fails and behind him is a new actuary to see whether corrections introduced by the clinician hold, etc. (p. 124)

The combination of approaches he suggested, in other words, was serial and successively corrective, a combination nobody could disagree with.

Still, a recent meta-analysis of the largest sample of studies so far (Grove, Zald, Lebow, Snitz, & Nelson, 2000) concluded as other studies with less sophisticated methods have: "Superiority for mechanical-prediction techniques was consistent, regardless of the judgment task, type of judges, judges' amounts of experience, or the types of data being combined" (p. 19).

Finally, let us examine, using behavior analysis, why psychological practitioners view equations as anathema and how we could modify that clinical behavior of our practitioners. It might be useful to examine the variables that control the behavior of practitioners, especially those who work by themselves. What are the consequences of the practitioner's behavior? Everyone in everyday life must predict the outcome of his or her behavior. Thus, if you find yourself at a train station and you need to find out what platform your train leaves from, you would predict that individuals wearing a conductor's uniform would be more likely to know where to go, that fellow travelers would know less, and, of those, the child travelers would be less helpful than the adults. Over the course of one's lifetime one learns how to behave to obtain information under various conditions because the consequence of getting information is immediate with respect to whom to ask for help. In the same way one learns to walk carefully over a wet floor and to go to a bakery rather than a tailor to obtain bread.

It is different when we talk of the clinical field where the consequences of our behavior are more significant although less definite, less immediate, and sometimes never revealed to us. Yet, in more complicated social relations that we engage in with relatives, friends, fellow workers, bosses, and those who work for us, we constantly make predictions about how our behavior will impinge on those other people and how successful we will be in assessing their behavior. We do not get these consequences immediately when it comes to social relations because the consequences are manifold, delayed, and often not recognized until it is too late for us to do anything about our behavior. Still, we do all get along to varying degrees, making our intuitive predictions of how the behavior of those around us will change or stay the same. The point is that on balance, we are successful in our social relations and if we find that not to be case, we might ourselves hunt up a psychotherapist. And we must always add that sometimes we are not aware, or for that matter, need be aware of the consequences of our behavior because of all the other variables impinging on the people that we respond to. The behavior of those whose behavior we try to predict might well be a function not of the variables that we invoke but a consequence of other people's behavior, other histories of those people whose behavior we are trying to predict, etc. It turns out that we receive a great deal of reinforcement for our behavior even though it is not our behavior that

is responsible for evoking it. All of this is also true, to varying degrees, for practitioners who work with patients or who try to predict the behavior of criminals, wayward children, or troubled "normals." For practitioners who consider themselves to be successful, as indicated to them by the things their patients say or do not say, or in the fact that they do not return with the same problems that brought them in, a good deal of reinforcement has been forthcoming inculcating the belief that they need no statistical information, thank you, to improve their own functioning. But their behavior is a function of both contingent and happenstance reinforcement (as it is for all of us, except that is more important when you are a psychotherapist or other kind of practitioner).

Now, some have suggested that comparisons between actuarial and clinical prediction showing the former to be better than the latter were made with inexperienced clinicians (those presumably who received no reinforcement about the accuracy of their predictions). If one taught them how to do it properly, they would be much better than the actuaries. In addition, we know that clinicians do not, as a rule, get sufficient feedback on their decisions. As a behavior analyst I am moved by the fact of insufficient reinforcement. A study by Goldberg (1968), however, questions the premise that practitioners can learn to make their predictions as well as the equations can. He asked clinicians to make a differential diagnosis of psychosis vs. neurosis. He compared clinical experts with graduate students and with non-psychologists. All judges alternated weeks of training with weeks of testing. The training sessions consisted of providing the judges with immediate information concerning the accuracy of their judgments. All three groups showed some learning; the non-psychologists showed the most but did not reach the level of the other two groups which, in turn, did not improve much at all. Only when the judges were given the optimal formula (used by the actuarial prediction) did the latter two groups improve materially, although they failed to reach the level of the formula itself.

This result is very interesting and relates to a study done by Verplanck (1962). He showed, using game cards, that one can reinforce card sorting of one kind (e.g., red border cards on the right, black border cards on the left) while at the same time reinforcing partially-related or unrelated hypotheses supposedly underlying the sorts independently. With so many possibilities, overlapping hypotheses could still provide the subjects with reinforcement while not being entirely right on the hypotheses or on the card sort. Thus, we have here not only evidence for the difficulty of learning to make predictions with the same accuracy as statistical prediction, we also have an explanation in operant conditioning terms for how the practitioner might have great difficulty learning to make the relevant discriminations underlying the predictions. In one of my own studies (Salzinger, Portnoy, Zlotogura, & Keisner, 1963) we found that the reinforcement of a response class of plural nouns, while subjects were emitting continuous speech, actually conditioned a subclass of plural nouns (i.e., those ending in the sound z as in dogs rather than all plural nouns). Whether or not clinicians can learn to improve their predictions through comparing their behavior with that of statistical prediction apparently depends on the precision of the conditioning process, both

the response class that gets conditioned and the discrimination they are learning to make.

We have already listed a large number of reasons for clinicians' opposition to using statistical prediction. Yet with a large literature telling us why we should make use of it, why have we not learned to live with it? What does it take to modify the behavior of practitioners who have for a long time done things in a particular way? Given a situation in which we have formulas, what should we do when the formula tells us with a 75% certainty that following it will provide us with the best answer, but a clinician tells us that he or she has a feeling that we ought to go with an alternative? Meehl (1957) discussed this problem in a paper titled "When shall we use our heads instead of the formula?" The clinician uses his or her head based on some special intuition, as described above, and he or she is using it because of a special circumstance, what we have already described as the "broken leg" condition.

Basically, we are left with the problem of having to determine whether we do have something as definite as a broken leg or merely something as vague as a hunch in psychopathology. Moreover, when hunch contradicts statistical equation, combination is not possible.

Meehl suggests that we (surreptitiously?) keep a record of the number of times that the clinician contradicts the statistical prediction, the number of times that he or she is right, and the number of times that he or she is wrong. The point is that by keeping such a record we can all profit because, after all, correct prediction is what we are all seeking. The next step is to persuade some large organization to adopt the formula approach, to allow the broken leg exception, and to keep a record of each time the broken leg is allowed to trump the statistical prediction. Also, we must insist that the broken leg exception be specified in large detail so that the different kinds of "legs" can be recorded and evaluated with regard to their validity. If enough of them occur, if they are found to be valid, and if they are well enough specified, they can then be added to our formula. That would allow us to eventually incorporate a better way of making predictions. This does not supply us with a solution to the problem of how to convince our clinical siblings to accept this kind of procedure, but it is interesting to note that statistical prediction does not stand alone in being slow to evoke acceptance.

Sobell (1996) described the problem of getting practitioners to accept the use of evidence-based health care. She began her discussion by reminding us of the great difficulty that knowledgeable people had to convince the British navy to accept the use of lemon juice by its sailors to avoid scurvy and ultimately death for 62% of them. Apparently some 400 years intervened between the first discovery of the usefulness of lemon juice and its final acceptance. With that as background, she mentioned a series of maneuvers for bridging the gap, as she put it, between scientists and practitioners—involvement of the practitioners in the planning, development, and implementation of the clinical trials; tailoring the implementation to the needs of practitioners; providing ongoing clinical support by the research team; workshops; and making the clinical materials readily available to the practitioners and the community agencies. Sobell made no explicit mention

of reinforcement, but clearly any change in behavior by practitioners must be accompanied by positive consequences. If practitioners are to assume use of an actuarial approach in their daily work, then their behavior must be followed by immediate consequences for their new behavior and those consequences must be obvious and positively reinforcing. If the consequence were simply that their work begins to resemble that of a clerk and they get to do nothing more interesting than that, we should expect little by way of change. It also means that the person in authority would have to be willing (better yet, eager) to institute the new way of judging the behavior of patients and of the clinicians who work with them, buttressing all with positive reinforcement.

When I embarked on writing this paper I decided to determine how much this issue was still enjoying the attention of our field. I went to three different search engines: Google, Scholar Google, and PsycINFO. It is not clear how many of the 36,100 items displayed by Google give reasons for denying the usefulness of actuarial prediction and how many are actually using it; whatever it is, the subject is clearly not hiding from sight. When I proceeded to Scholar Google, which is still in formation, I came across 2,210 items. Finally, examination of PsycINFO yielded 3,340 references. Clearly, the issue is still being examined, indeed more so than many others.

This controversy has managed to bring to the surface several findings that we should definitely make use of. The first is the need to empirically validate the effectiveness of our predictions, no matter how we make them. The second is the fact that simpler is often better. The statistical predictions that have been so powerful made use of just a few variables. Indeed, it is quite possible that the reason for clinical prediction being less effective than the statistical one is that the latter is not distracted by variables that bear no relation to what is being predicted. Finally, the superiority of simplicity has also manifested itself in the usefulness of simple forms of combination, such as addition and subtraction rather than the sophisticated equations that appear at first to be needed. All of this will make using statistical prediction easier. Let us, therefore, get started in doing it, and in doing so let us honor Paul Meehl.

References

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

Faust, D., Meehl, P. E., & Dawes, R. M. (1990). Response. *Science, 247*, 146-147.

Faust, D., & Ziskin, J. (1988a). The expert witness in psychology and psychiatry. *Science, 241*, 31-35.

Faust, D., & Ziskin, J. (1988b). Response. *Science, 241*, 1143-1144.

Faust, D., & Ziskin, J. (1988c). Response. *Science, 241*, 652.

Fowler, R. D., & Matarazzo, J. D. (1988). Psychologists and psychiatrists as expert witnesses. *Science, 241*, 1143.

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

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

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

Kleinmuntz, B. (1990). Clinical and actuarial judgment. *Science*, *247*, 146.

Meehl, P. E. (1954). *Clinical versus statistical prediction*. Minneapolis: University of Minnesota Press.

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

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

Salzinger, K., Portnoy, S., Zlotogura, P., & Keisner, R. (1963). The effect of reinforcement on continuous speech and on plural nouns in grammatical context. *Journal of Verbal Learning and Verbal Behavior*, *1*, 477-485.

Sobell, L. C. (1996). Bridging the gap between scientists and practitioners: The challenge before us. *Behavior Therapy*, *27*, 297-320.

Spitzer, R. L., Williams, J. B. W., & Pincus, H. A. (1988). Psychiatric diagnosis. *Science*, *242*, 651-652.

Verplanck, W. S. (1962). Unaware of where's awareness: Some verbal operants—notates, monents, and notants. *Journal of Personality*, *30*, 130-158.

Zubin, J. (1955). Clinical vs. actuarial prediction: a pseudoproblem. *Proceedings of the 1955 Invitational Conference on testing Problems*. Princeton: Educational Testing Service, 107-128.