JOURNAL OF EXPERIMENTAL RESEARCH IN PERSONALITY 3, 168–171 (1968)

METHODOLOGICAL CRITIQUES

Seer over Sign: The First "Good" Example?

LEWIS R. GOLDBERG

University of Oregon and Oregon Research Institute

In two highly influential articles previously published in this journal, Lindzey (1965) presented some findings purportedly relevant to the clinical vs. statistical prediction issue, and Meehl (1965) praised Lindzey's research as having provided "the first good example" of seer over sign. Meehl's conclusions are here questioned on the grounds that (*a*) Lindzey's data do not support Meehl's assertion that any seer demonstrated superior "validity generalization" over any sign, (*b*) the validity of the seers and signs did not differ at a statistically significant level in either of Lindzey's two studies, and therefore (*c*) these studies cannot be cited as "the first good example" of anything at all. The implications of this argument for future studies based upon the "validity generalization" paradigm is discussed.

In 1965, Gardner Lindzey published a provocative paper entitled "Seer versus Sign," in which he attempted to demonstrate that in one clinical situation (the prediction of homosexuality from TAT protocols), clinical judgments ("seers") were finally found to be superior to actuarial predictions ("signs"). In a companion paper in the same issue of the journal, Paul Meehl—the official arbiter of the clinical vs. statistical establishment—cited Lindzey's study as "the first good example" of seer over sign—no mean feat in a continuing series of contests which Meehl (1965) summarized as follows:

> "Monitoring of the literature yields a current bibliography of some fifty empirical investigations in which the efficiency of a human judge in combining information is compared with that of a formalized ("mechanical," "statistical") procedure. The design and the range of these investigations permits much more confident generalization than was true on the basis of the eighteen studies available to me in 1954. They range over such diverse substantive domains as success in training or schooling, criminal recidivism and parole violation, psychotherapy (stayability and outcome), recovery from psychosis, response to shock treatment, formal psychiatric nosology, job success or satisfaction, medical (nonpsychiatric) diagnosis, and general trait ascription or personality description. The current "box score" shows a significantly superior predictive efficiency for the statistical method in about two-thirds of the investigations, and substantially equal efficiency in the rest. . . . It would be difficult to mention any other domain of psychological controversy in which such uniformity of research outcome as this would be evident in the literature. Since Professor Lindzey's paper is the first and only empirical comparison of the relative efficiency of the two methods showing clear superiority for the clinical judge, it is deserving of special attention" (Meehl, 1965, p. 27).

Since my initial reading of Professor Lindzey's paper had convinced me that his data did not support these conclusions, I was perplexed to find my own reactions so strongly at variance with those of Paul Meehl, and I eagerly awaited what I assumed would be an immediate dogfight between Lindzey and Meehl, on the one hand, and the actuaries, on the other. Two years have now gone by, Lindzey's paper has begun to be cited as evidence favoring clinical over statistical prediction, and still no one has leaped into the fray. Therefore, if only to test my own clinical judgment, I want to now put forth the view that Lindzey's study cannot be considered evidence of seer over sign.

In Lindzey's paper, he reports the results of two studies on the differentiation of homosexuals from nonhomosexuals on the basis of their TAT protocols. The first study compared "20 undergraduate male students who had acknowledged overt homosexuality and a group of 20 undergraduates comparable in sex, age, and educational level but with no known history of homosexuality." Their 40 TAT protocols were sorted into two diagnostic groups by *one* judge (Judge B), and his accuracy was compared with that achieved by various composites of 20 signs coded from the TAT protocols. In the second study, Lindzey compared 14 inmates of a state maximum security prison who "were known to have been overtly homosexual prior to imprisonment" with 16 inmates of the same prison who "provided no evidence of homosexuality prior to imprisonment or during incarceration," the two groups having been matched "in terms of age, education, intelligence, period of imprisonment, and place of residence." The same judge (Judge B) again sorted the 30 TAT protocols into two groups, and his diagnostic accuracy was compared with that of various actuarial indices derived from the first study. In addition, a second judge (Judge A) also sorted the protocols; this judge had not been involved in the first study. The major findings from both studies are summarized in Table 1.

The values in Table 1 indicate the number (and percentage) of diagnostic hits and misses for the "best" actuarial formula, and for each of the two judges. Note that *only* Judge B participated in both studies; in Study I, he delivered a 95% hit-rate (as compared to 90% for the "best" formula), while in Study II he achieved a 60% hit-rate (as compared to 57% for the formula). I conclude from these findings that Judge B and the formula "behaved" remarkedly alike: they both were quite accurate in diagnosing homosexual from non-homosexual students (90 vs. 95%) and they both were poor at diagnosing homosexual from non-homosexual prisoners (57 vs. 60%). Stated in terms of "validity generalization"—a criterion proposed by Meehl as one potential superiority of seer over sign—the cross-sample shrinkage in validity for Judge B was 35%, while that for the formula was 33%. Clearly, then, one would have to conclude from the comparison of this one judge with the formula that for this problem seer and sign yielded roughly equivalent validities.

Now, how does Judge A fit into the picture? Since he did not participate in the first study, we do not know how accurately he would have performed in it. Unlike Judge B and the formula, his abilities were *not* tested across two diverse samples, and thus the talents of this judge in Study II are in an important sense irrelevant to the crucial validity generalization issue raised by Meehl. Just as there may be some other formula which would work well in Study

TABLE 1
SUMMARY OF THE FINDINGS FROM TWO STUDIES BY LINDZEY (1965)[a]

	Best actuarial formula		*Judge B*		(Judge A)	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Study I	36 (90%)	4 (10%)	38 (95%)	2 (5%)	— —	— —
Study II	17 (57%)	13 (43%)	18 (60%)	12 (40%)	24 (80%)	6 (20%)

[a] The main values in the table are the number of correct and incorrect diagnoses. Figures in parentheses are the percentages of correct and incorrect diagnoses. The values for the "best actuarial formula" in Study I are *not* cross-validated from some previous study and thus are spuriously high (see Lindzey, 1965; p. 20).

II (though not in Study I), so Judge A might perform much more poorly in Study I than he did in Study II. This, of course, is only a supposition; since Judge A is deceased, the data for the missing cell must remain forever missing. However, it is my contention that it is solely this estimate of how well Judge A *would have performed* in Study I (had he participated in it) which led Meehl to conclude that the seer's validity generalization had afforded him some superiority over the actuary. For, clearly if Judge A had predicted with, say, a 20% hit-rate in Study I, then all one could conclude is that whatever processes were involved, there was virtually no validity generalization demonstrated by *either* sign or seer.

What Meehl appears to be assuming is that Judge A would have performed accurately in Study I, had he participated in it. Perhaps he reasoned that: (*a*) since the diagnostic task in Study I seems to have been easier than that in Study II (for both one judge and several signs), and since (*b*) Judge A performed more accurately than Judge B in Study II, *therefore* (*c*) Judge A would have performed at least reasonably well in Study I. While this chain of reasoning sounds compelling, the available evidence on the *generality* of judgmental accuracy (e.g., Goldberg, 1965, p. 13; Crow and Hammond, 1957), as well as recent work on the clinical judgment process (e.g., Goldberg, 1968), would lead me to the opposite conclusion. I would reason that (*a*) since judges appear to be rather insensitive to sample differences (i.e., they do not generally seem to change their diagnostic model as they move to new samples), and (*b*) since it seems clear that the TAT "signs" associated with homosexuality among prison inmates appear to differ from those associated with homosexuality among college students, therefore, (*c*) the probabilities favor the hypothesis that Judge A was using a judgmental model which is more appropriate for prisoners than for undergraduates and that he would have suffered some considerable loss in accuracy had he moved to the college sample. Obviously, this is all sheer speculation; the evidence is simply unavailable. However, until this study is replicated so as to include some data for the crucial missing cell, this study can *not* be considered "the first good example" of seer over sign.

Moreover, even in Lindzey's study, the validity of Judge A in Study II did *not* differ at a statistically significant level from that achieved by the formula ($\chi^2 = 3.77$; $df = 1$; $p > .05$; my calculations), nor—perhaps of more relevance—did the *average* validity of the two judges in Study II (70%) differ significantly from that of the formula (57%). In addition, the study clearly violates Canon Number One of Meehl's (1954) basic racing rules: both clinician and actuary must have access to the *same* data. In Lindzey's two studies, the clinicians had access to the entire TAT protocols, the actuary only to a set of 20 precoded signs. If these 20 signs did not include all of the cues relevant for this diagnostic problem, then the clinicians had data unavailable to the actuary. While I do not see this as a serious problem, a stickler for the rules undoubtedly would. More importantly, however, no actuary worth his IBM cards would want to (*a*) generalize from a sample of only 40 cases (*b*) to a new sample which is obviously different in crucial ways (e.g., prisoners vs. college students) from the derivation sample. He would simply say that this was hardly a fair test of actuarial procedures. In which case, how can it be considered "the first good example" of anything?

Paradoxically, this entire argument stems in part from the very completeness of Lindzey's presentation of his findings. If Lindzey had reported solely the findings from Study II and omitted any mention of Study I (save as a derivation sample for his actuarial index), the study might have appeared all the more as a victory for seer over sign (especially if significance tests are ignored). But, this would have been misleading, and therefore Lindzey's study serves as an excellent example for any future researcher who might wish to infer validity generalization on the basis of data from a single sample. For, the actuary does not ordinarily seek validity generalization, but rather assumes *only* that his signs will generalize from an initial random sample of a population to another random sample from the *same* population; if the signs are to be used with some different population (e.g., students versus prisoners), the actuary will certainly demand an empirical examination of their validity in the new population. Now, if the thrust of Meehl's argument is solely that clinicians validly change their diagnostic models from one sample to another quite different one (while the actuary must begin anew), then it is immediately apparent that any contest between the two on this (new) issue must include data on the validities of the same signs and seers in *more than one* sample. In Lindzey's study, only Judge B meets this criterion. And, in Lindzey's study, Judge B performed at approximately the same level of accuracy as the formula. Therefore, shouldn't this study be tallied in the "box score" along with those others showing "substantially equal efficiency" of seer and sign?

References

Crow, W. J., and Hammond, K. R. The generality of accuracy and response sets in interpersonal perception. *Journal of Abnormal and Social Psychology,* 1957, **54,** 384-390.

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

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

Lindzey, G. Seer versus sign. *Journal of Experimental Research in Personality,* 1965, **1,** 17-26.

Meehl, P. E. Seer over sign: The first good example. *Journal of Experimental Research in Personality,* 1965, **1,** 27-32.

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