ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 16,13 -22 (1976)

Man versus Model of Man: Just How Conflicting Is That Evidence?

LEWIS R. GOLDBERG

University of Oregon and Oregon Research Institute

The distributions of the values for four of the five cues in Libby's (1976) study were quite markedly skewed. When these values were rescaled by a simple normalizing transformation, the results changed remarkably: Instances of model outperforming expert jumped from 23 to 72%. As a consequence, Libby's data can hardly be invoked as "conflicting evidence."

Most psychological findings are not all that robust: Change the sample of subjects, the instructions, or the experimental apparatus a wee bit, and the original finding oft disappears. Not so, however, for one recent result from the literature on human judgment and decision making: the finding that predictions from multiple regression models of human judges are typically no less accurate than those provided by the judges themselves (Goldberg, 1970). Indeed, this finding has now been verified in at least five studies, which include quite substantial differences in the number and type of judges, the nature and content of the decisions, the number and type of cues employed, and the basic task structure (e.g., the intercorrelations among the cues, and between the cues and the criterion).[1]

As a consequence, any failure to replicate so hardy a phenomenon, such as has apparently been reported by Libby (1976), should be most instructive, since it could point to limiting conditions where the general finding may fail to hold. With this as our goal, let us review the details of Libby's study to see how his experimental and/or analytic procedures differed from those of his predecessors.

THE STUDY REPORTED BY LIBBY (1976)

Libby asked 43 bank loan officers to predict whether each of 60 large industrial corporations would or would not experience bankruptcy; their predictions were based on five financial ratios computed from each firm's

This project was supported by Grant No. MH-12972 from the National Institute of Mental Health, U. S. Public Health Service.

Requests for reprints should be sent to Lewis R. Goldberg, Oregon Research Institute, P. O. Box 3196, Eugene, Oregon 97403.

[1] For a description of the modeling procedures and an analysis of the conditions under which the performance of judges can possibly outperform their models, see Goldberg, 1970; for an extension of this argument and a review of more recent evidence, see Dawes and Corrigan, 1974.

annual statement. Half of the companies had indeed failed, while the other half were still in business. For each company, the judges made a dichotomous prediction and then indicated their confidence in this judgment on a three-place rating scale. The resulting predictions constitute a quasi-continuous variable:

(1) Very confident that the firm will *fail*.
(2) Confident that the firm will *fail*.
(3) Not very confident, but guess firm will *fail*.

(4) Not very confident; guess firm will *not* fail.
(5) Confident that the firm will *not* fail.
(6) Very confident that the firm will *not* fail.

In his analyses, Libby elected to treat these judgments in their dichotomous form and to use a discriminant function for modeling each of the 43 judges. Libby reported that 26 of the judges were more valid than their models; 10 judges were less valid; and seven judges were indistinguishable from their models. As an explanation for the discrepancy between these findings and those from previous studies, Libby has argued that his financial analysis task was better defined and measured than those in other studies, and his judges had much more task-related expertise.

There are, however, other distinctions between Libby's study and those that preceded it, distinctions that may turn out to be far more significant than the two he discussed. A few examples:

(a) Libby elected to use a discriminant function algorithm to construct his judgmental models, and he elected to report "hit-rates" as the index of predictive accuracy. All past studies have employed multiple regression models, and used correlations as a measure of validity.

(b) Libby elected to analyze only the dichotomous form of his judgmental rating scale, in spite of some evidence (presented in Table 2 of his study) that confidence and accuracy were at least weakly related. Previous studies have typically employed some sort of quasi-continuous rating scale.

(c) The cues or predictors that Libby employed were all *ratios*, constructed by dividing the firm's net income, cash, or sales by its total assets, current liabilities, or current assets. Given that half of these firms experienced bankruptcy soon after their financial statements were published, it is not unreasonable to suspect that a few of the denominators might be quite small, a few ratios quite large, and the general distribution of these ratios markedly skewed. In previous studies, the cue distributions were all reasonably symmetric.

It is possible that the apparent "conflict" between Libby's findings and those of others may stem solely from one or more of these factors. By reanalyzing Libby's data, it should be possible to discover the separate effects of each of them, and thus to judge the robustness of the purportedly conflicting evidence.

A REANALYSIS OF LIBBY'S DATA[2]

First, let us see whether Libby's findings can be replicated using the more traditional multiple regression methodology employed in past studies. Table 1 presents a summary of the results of these analyses for the 43 individual judges and for the "composite judge" (the average ratings for each firm from all of the judges). Values were computed separately for the six-step (continuous) rating scale (C) and for the dichotomous prediction (D). The first two columns of Table 1, which present the multiple correlations between the five cue values and each expert's judgments, indicate the linear predictability of these judges on this task. As would be expected, predictability is higher for the continuous than for the dichotomous rating scale, and higher for the composite judge than for the average of the individual judges. The next four columns in the table provide a breakdown of judgmental accuracy into both linear and nonlinear components, using the model suggested by Tucker (1964). In contrast to all previous studies, there would appear to be a slight nonlinear component of judgmental accuracy for the average judge on this task, and a quite substantial component for the composite judge. The six right-hand columns of Table 1 present the validity coefficients for the models, the judges, and the difference between the two (Δ); negative differences favor the expert, positive ones the model. Note that the models fare quite poorly. Using the dichotomous response, only 10 (of 43) models outperformed the expert; using the continuous scale, only 6 did so.

What has been demonstrated so far is that the results presented by Libby did not stem simply from his use of discriminant versus regression analysis, nor from his use of the dichotomous rather than the continuous rating scale. Might something else be going on? Inspection of the univariate frequency distributions for each of the five cue values reveals the problem: Four of the five distributions were markedly skewed. While the cue "current assets/total assets" was distributed in a reasonably symmetric manner, the cue "net income/total assets" had an extreme negative skew (a few of the firms having large *negative* net incomes), and the other three cues ("cash/total assets," "current assets/current liabilities," and "sales/current assets") all had enormous positive skews. For these four cues, most of the values were quite close to each other, with but one or two values spread far away. What would happen, then, if the cues were rescaled to eliminate that outlandish skewness?

Each of the five cue values was rescaled by a simple normalizing trans-

[2] The author is indebted to Robert Libby for providing these data.

TABLE 1

JUDGMENTAL INDICES FOR 43 JUDGES AND THEIR MODELS: ORIGINAL CUE VALUES[a]

	Linear predictability		Linear accuracy		Nonlinear accuracy		Validity coefficients					
							Continuous (C)			Dichotomous (D)		
	C	D	C	D	C	D	Model	Man	Δ	Model	Man	Δ
43 individual judges												
Highest value	.91	.81	.92	.92	.60	.57	.53	.66	.07	.53	.67	.12
Lowest value	.58	.56	.23	.03	−.08	−.22	.13	.07	−.19	.02	−.10	−.19
Mean	.79	.73	.76	.76	.35	.30	.44	.51	−.08	.44	.49	−.05
Composite judge	.90	.78	.80	.83	.53	.52	.46	.61	−.15	.48	.64	−.16

[a] All values are correlation coefficients ($N = 60$). Values are presented separately for judgments based on the six-step (continuous) scale (C), and those based on the dichotomous response (D).

TABLE 2

INTERCORRELATIONS AMONG THE FIVE CUES, AND THEIR CORRELATIONS WITH THE CRITERION AND WITH THE JUDGMENTAL COMPOSITE[a]

	Cues						Composite	
	1	2	3	4	5	Criterion	Continuous[b]	Dichotomous[c]
1	—	.13	.24	.48	−.38	.43	.80	.68
2	.14	—	.42	.26	−.28	−.06	.39	.26
3	.34	.48	—	.21	−.02	.08	.36	.29
4	.66	.31	.42	—	−.33	.48	.66	.53
5	−.31	−.26	−.08	−.35	—	−.08	−.28	−.08
Criterion	.49	−.06	.15	.57	.00			
Composite								
Continuous[b]	.84	.37	.51	.84	−.19			
Dichotomous[c]	.72	.25	.39	.68	−.03			

[a] Correlations based on the original values of the cues are presented above the main diagonal, while those based on the transformed cue values are presented below the diagonal. The linear predictability of the criterion (R_e) is .58 with the original cue values, and .67 with the transformed values. $N = 60$.

[b] Judgments based on a six-step (continuous) scale.

[c] Judgments based on a dichotomous response.

formation (e.g., Leverett, 1947), and all analyses were repeated. Table 2 presents the intercorrelations among the cues, both before and after the scale transformations. Also included in the table are the correlations for both sets of cues with the judgmental composite, and the point-biserial correlations between the cues and the criterion. Four of the five cues were positively correlated; for those four cues, all of their intercorrelations increased after rescaling. More importantly, the multiple correlation between the five cues and criterion also increased, from .58 to .67.

Table 3 is identical in format to Table 1, but now based on the transformed cue values. The major differences between the two tables are summarized in Table 4, which shows that Libby's conflicting evidence has disappeared after the cues were transformed. For the dichotomous prediction, the format analyzed in his original paper, instances of model outperforming expert jumps from 23 to 72% as a result of the cue transformations![3]

[3] Table 4 must not be misinterpreted as implying that the models were more valid when based on the dichotomous, as compared to the continuous, scales. As Table 3 indicates, the accuracy of the models was virtually identical in the two cases. However, the validity coefficients for the judges were somewhat higher for the continuous than for the dichotomous scales, and consequently the difference in accuracy between judge and model (Δ) was somewhat less in the dichotomous than in the continuous case.

TABLE 3

JUDGMENTAL INDICES FOR 43 JUDGES AND THEIR MODELS: TRANSFORMED CUE VALUES[a]

	Linear predictability		Linear accuracy		Nonlinear accuracy		Validity coefficients					
							Continuous (C)			Dichotomous (D)		
	C	D	C	D	C	D	Model	Man	Δ	Model	Man	Δ
43 individual judges												
Highest value	.96	.86	.92	.93	.46	.42	.61	.66	.17	.62	.67	.18
Lowest value	.62	.60	.20	−.04	−.07	−.17	.13	.07	−.09	−.03	−.10	−.07
Mean	.84	.77	.80	.80	.18	.16	.53	.51	.02	.53	.49	.04
Composite judge	.95	.82	.83	.88	.32	.37	.56	.61	−.05	.58	.64	−.06

[a] All values are correlation coefficients ($N = 60$). Values are presented separately for judgments based on the six-step (continuous) scale (C) and those based n the dichotomous response (D).

TABLE 4
THE COMPARATIVE VALIDITY OF MODEL AND MAN, BEFORE AND AFTER THE CUE TRANSFORMATION

	Original cue values	Transformed cue values
Continuous rating scale		
Number (of 43) models > man	6	25
Mean Δ	−.08	.02
Range Δ	−.19 to .07	−.09 to .17
Dichotomous response		
Number (of 43) models > man	10	31
Mean Δ	−.05	.04
Range Δ	−.19 to .12	−.07 to .18

Badly skewed cue distributions can make models perform poorly for one or both of two reasons: (i) The presence of one or two extreme values can so change the regression weights that the model itself is rendered impotent; and/or (ii) the presence of those extreme values even when processed through an optimal model will lead to such a skewed distribution of predicted values that the resulting correlations may be severely attenuated.[4] Which of these two factors had the most pernicious effect in Libby's study?

We can unconfound these two factors by using the original (skewed) cue values in the regression equations derived from the transformed cues, and analogously using the transformed cue values in the regression equations derived from the skewed cues. Table 5 provides a summary of such analyses, both on criterion linear predictability (the multiple correlation between the five cues and the criterion) and on the accuracy of the models for the composite and the average individual judge. Apparently, the effect

[4] There are two major reasons why cues with highly skewed distributions should not be directly employed in linear models. First of all, the maximum possible value of any product–moment correlation is attenuated by differences between the distributions of the two variables being correlated; only when the two distributions are identical in form can the maximum correlation reach unity. There is, however, an even more serious problem with skewed distributions in this context: The few very extreme values at the tail of the distribution are given far more weight than each of the more typical (modal) values in affecting the direction and magnitude of the correlation coefficient. Specifically, chance perturbations in those extreme values can markedly affect the resulting correlations and, consequently, the regression weights (which are a function of the intercorrelations among the cues, and the correlations between the cues and the criterion). In addition, since the predicted values from such a regression model may themselves be highly skewed, the few extreme predicted values will again be unduly weighted in their relative effect on the resulting validity coefficient.

TABLE 5

Unconfounding the Effects of Skewed Cue Values and Nonoptimal Regression Weights[a]

		Original cue values		Transformed cue values	
		Original weights	New weights	Original weights	New weights
Criterion	Linear predictability	.58	.47	.54	.67
Accuracy of model	Average judge	.44	.37	.46	.53
	Composite judge	.48	.34	.48	.58

[a] Analyses based on the dichotomous response; $N = 60$.

of those skewed cue distributions was twofold, serving both to perturb the regression weights in the original equations and to attenuate the predictions based on other weights. While both types of effect were of substantial size, the latter was a bit more detrimental than the former.

DISCUSSION

How conflicting, then, *is* Libby's evidence? Table 6 provides a detailed comparison of the findings from the reanalysis of Libby's data and the findings from another study of professional experts: clinical psychologists diagnosing 861 psychiatric patients as either neurotic or psychotic on the basis of their MMPI profiles (Goldberg, 1970). The column headings in Table 6 include the major summary statistics suggested by Goldberg (1970) for comparing judges and their models. In both studies, the average individual judge was less valid than his model (and by virtually the same small margin), and in both studies the composite judge bested its model. The major difference between the two studies is that the bank loan officers (and their models) were more accurate than the clinical psychologists (and their models), and this increase in accuracy appears to have stemmed both from linear and from nonlinear components. As a consequence, it becomes all the more important that here in Libby's study, where there was some nonlinear component to judgmental accuracy, the major finding from all past studies still remains: "Linear regression models of clinical judges can be more accurate diagnostic predictors than are the humans who are modeled" (Goldberg, 1970; p. 430).

TABLE 6

THE COMPARATIVE VALIDITY OF MAN VERSUS MODEL: SUMMARY TABLE[a]

		N	Δ	$=$	r_m	$-$	r_a	$=$	G	$\cdot$	R_e	$\cdot$	$1-R_s$	$-$	C	$\cdot$	$(1-R_e^2)^{1/2}$	$(1-R_s^2)^{1/2}$
	Goldberg	861	.03		.31		.28		.68		.46		.23		.08		.89	.64
Typical judge	Libby (C)[b]	60	.02		.53		.51		.80		.67		.16		.18		.74	.54
	Libby (D)[c]	60	.04		.53		.49		.80		.67		.23		.16		.74	.64
	Goldberg	861	−.02		.33		.35		.72		.46		.11		.13		.89	.45
Composite judge	Libby (C)[b]	60	−.05		.56		.61		.83		.67		.05		.32		.74	.31
	Libby (D)[c]	60	−.06		.58		.64		.88		.67		.18		.37		.74	.57

[a] The values from Libby's study are based on the transformed values of the cues (Table 3).

[b] Judgments based on a six-step (continuous) scale.

[c] Judgments based on a dichotomous response.

REFERENCES

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

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

Leverett, H. M. Table of mean deviates for various portions of the unit normal distribution. *Psychometrika,* 1947, **12,** 141–152.

Libby, R. Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance,* 1969, **16,** 1–12.

Tucker, L. R. A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd. *Psychological Review,* 1964, **71,** 528–530.

RECEIVED: February 26, 1976.