Bulletin of the Psychonomic Society
1991, 29 (6), 553-556

Accuracy of alternative methods for describing experts' knowledge of multiple influence domains

ROBERT M. HAMM
University of Colorado, Boulder, Colorado

Accuracy of different methods for modeling experts' knowledge of a domain with continuous additive relations among multiple variables was compared. Bootstrapped models of case judgments with premeasured cues were more accurate than models produced directly by the experts. Bootstrapped models of judgments were less accurate when experts perceived the cues than when the cues were measured for them, in two of three domains. A formal procedure for guiding the experts in producing the models did not improve accuracy, whereas correcting their slips did. More accurate use of a diagnostic (as opposed to predictive) cue-criterion relationship was observed when the experts judged cases perceptually than when they wrote abstract formulas.

Models of individuals' knowledge are useful for a number of reasons, including description for scientific purposes (Hammond, Hamm, Grassia, & Pearson, 1987) and the elicitation of knowledge for expert systems (Bamber, 1990). One form of knowledge concerns the functional influence upon one factor of a number of other factors. In some cases, particularly when the domain is known imperfectly, such knowledge can be modeled using additive linear models, such as

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3.$$

This model would summarize the judgments the individual would make about the y feature of situations in which information about the x_1, x_2, and x_3 features was available.

A variety of techniques can be used for producing models of this form. They vary both in the amount of effort required from the builder (e.g., knowledge engineer) and the informant (expert) and in the quality of the resulting model (accuracy, validity, generality). This paper reports a comparison of the results of using two techniques (two variants each) to produce models in three domains. It complements work by others (reviewed by von Winterfeldt & Edwards, 1986).

A fundamental distinction among methods for producing additive linear models of an individual's judgments of the dependency of one entity upon a set of entities may be drawn between methods in which the individual (1) *provides the structure* (variables and parameters) in the model, or (2) *judges individual cases* or instances, leaving the analyst to produce the model using analytic techniques, such as multiple regression. When the individual provides the structure, this can be done with or without guidance that assures that the model adheres to general mathematical principles. When the individual judges particular situations, the information about these cases can be either perceptual or abstract (i.e., already measured). This paper compares the accuracy of knowledge models produced using methods that differ in each of these respects.

The models were produced by Hammond, Hamm, Grassia, and Pearson (1983, 1987; see also Hammond, Hamm, & Grassia, 1986) in a study of highway engineers' judgments. Separate models were constructed for three characteristics: the safety, vehicle-bearing capacity, and aesthetic quality of highways. For each of these domains, each expert made models using three techniques. Because the earlier papers focused on different theoretical issues, the comparisons presented below have not previously been published.

METHOD

Two basic methods were used to produce models of experts' knowledge: (1) "bootstrapping" the expert—that is, making statistical descriptions of the relation of the expert's judgments about individual highways to the characteristics of those highways; and (2) having the expert build the model by writing its formula.

Twenty-one male highway engineers each provided three models about each of three highway characteristics. The knowledge models were made and evaluated with reference to a set of 40 rural two-lane Colorado highways (see Hammond et al., 1983, 1987). These provided "cases" for the experts to judge so that best-fit models could be produced by fitting their judgments to the highway characteristics, using multiple regression. The decisions about what variables to include in the model for each characteristic were made a priori, in consultation with other highway experts. Because objective measures of the modeled characteris-

The research project was supported by the Engineering Psychology Programs, Office of Naval Research, Contract N00014-81-C-0591, and the preparation of this paper was supported by a National Research Council Senior Associateship award through the Army Research Institute, Fort Leavenworth, KS. Kenneth Hammond and Janet Grassia collaborated on the design of the original study, and J. Grassia and Tamra Pearson on the analysis of the data. Mary Luhring provided access to the archives. Reprint requests may be sent to Robert M. Hamm, Institute of Cognitive Science, Box 345, University of Colorado, Boulder, CO 80309-0345.

tics were available, it was possible to assess the accuracy of models produced using the different methods.

Bootstrapping: Fitting Models to Experts' Judgments of Hypothetical Cases

In this method, the expert made a judgment (e.g., of capacity, in terms of the number of cars the road could carry in an hour) for each of the 40 highways. This was done twice for each task—once using pictures of the highways (filmstrips of 1- to 3-mile highway segments, one photograph each 50 feet) and once using abstract profiles (bar graphs measuring features of those segments). In doing this, the decision about model form was made a priori: a linear weighted average was used (product of the regression analysis method used for producing the best-fit model). Each engineer judged the 40 highways using both filmstrip and bar-graph displays for each of the three characteristics.

Input variables perceived by the expert: The filmstrip display. In this condition, each highway segment was presented for the expert's judgment as a moving filmstrip (see Hammond et al., 1983). The expert judged its safety (or capacity or aesthetic value) using an appropriate numerical scale. An objective numerical measure was then taken of the pertinent variables, using information in Highway Department records and frame-by-frame analysis of the filmstrips. A model of the expert's knowledge was produced by regressing the judgments onto the measured predictors. The accuracy of the model was assessed by correlating its predictions for the 40 highways with the objective measure of the variable in question.

Input variables measured by the researchers: The bar-graph display. In this condition, each highway segment was described in terms of its measures on the pertinent variables (the same measures used for modeling the filmstrip judgments), which were presented as bar-graph profiles with numerical labels for the lowest and highest values on the dimension and for the actual value of the dimension for this highway (the length of the bar; see Hammond et al., 1983).

Experts' Production of Models

In this method, the expert built a model, embodied in a formula, to express his knowledge of the effect of the offered input variables (highway features) on the characteristic of highways (safety, capacity, or aesthetics). These models were built in two ways by different experts: guided (n = 3) or not guided (n = 18). The engineers were given the variables that were available for use in the model, with a definition of the variable and a metric for measuring it.

Nonguided production of formulas. Eighteen of the engineers produced the models of their knowledge without guidance. They were told to produce a mathematical formula and to limit the input variables to a given set of 8 to 10 variables.

Guided production of formulas. Three of the engineers were guided when producing their formulas (three formulas each). The procedure consisted of a sequence of steps, managed through the use of a set of forms, that ensured the formulas provided sensible measurements. The steps are summarized in Hamm (1991), and the forms are available in Hamm (1990).

The guided or nonguided formula was written by the expert at the conclusion of the session. It was then applied to a highway, as a check. The expert could make changes in the formula at this stage.

For data analysis, the formula was programmed as a FORTRAN-like SPSS statement. Graphs were reduced to a formula using a best-fit program on several points. Tables were represented by nested loops that covered every cell of the table. Applied to the vector of measures describing each of the 40 Colorado highways, the formula predicted a value for each highway. Formula accuracy was assessed by correlating these predictions with the actual measures of the highway features.

RESULTS

Comparison of Guided and Nonguided Expert Models

The accuracies of the experts' own models when they were guided and nonguided are shown in Figure 1. Data are displayed as correlations, but the statistics were calculated using Fisher Z-transformed correlations, which are more normally distributed. The formulas predicting highway aesthetics were more accurate (mean = .842) than were those predicting capacity (.573; t = 12.7, p = .000) or safety (.558; t = 12.6, p = .000). The high accuracy of the aesthetics models depends on three facts (see Hammond et al., 1987). First, factor analysis of the predictor variables showed loadings on only one factor. Thus, all input variables are intercorrelated. Second, the impact of any variable on the criterion (the mean aesthetic value judgment of a number of citizens) is fairly obvious. Third, the engineers generally used simple weighted-average formulas. With correlated predictors, the weights are not important as long as one has the correct direction (Dawes, 1979). With safety and capacity, on the other hand, the criterion was very specific. The capacity measure was produced using a standard procedure involving several tables and graphs. The safety measure was the total accident rate on the particular stretch of highway. In these two tasks, the predictors were not highly intercorrelated, and the experts tried complicated formulas.

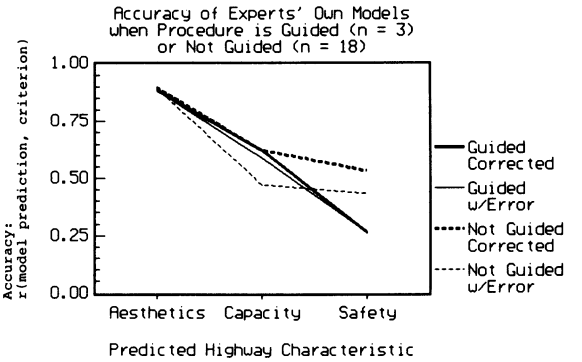

Figure 1. Accuracy of experts' own models when procedure was guided (n = 3) or nonguided (n = 18).

Use of complicated formulas introduces the possibility of errors due to slips between intention and action (Norman, 1981). Such errors were discovered when the engineers' formulas were being coded. Separate models were made for the expert's uncorrected formula and for a formula corrected to conform to the expert's intent. The mean correlations of each type of model with the criterion are shown in Figure 1.

With the aesthetics models, correction and guidance had no effect on the accuracy of the models. In fact, seven models had correctable errors, but these made little difference in formula accuracy. With the capacity models, the corrected models of the nonguided experts were as accurate as the models of the guided experts. There were 10 engineers whose models required a correction, and the corrections increased the mean accuracy of the expert's formulas from .468 to .620 [not statistically significant; $t(16)$ = 1.5, p = .15].

With the safety models, the guided engineers were worse than the nonguided. Correcting the nonguided engineers' formulas (seven required correction) increased

their mean accuracy from .430 to .531. The guided experts were worse [mean accuracy = .264; $t(19) = 2.0$, $p = .06$] because of the terrible performance of one expert's model (−.226). The other two guided experts had model accuracies of .581 and .436, in the same range as the nonguided experts' formulas. The bad guided model was inaccurate not because of a slip, but rather because of the expert's understanding: it expressed the relations he thought were true. (He was a retired specialist in landscape and had not studied safety quantitatively. His filmstrip judgments for aesthetics and for capacity were the best of the 21 engineers; his filmstrip judgments for safety were the worst.) As such, this model would have been inaccurate even if it had been nonguided, and, hence, it reflects on the random assignment of engineers to the guided or nonguided condition and the small number of engineers in the guided condition rather than on the efficacy of the guided condition.

In conclusion, use of a formal procedure for guiding engineers when they produced continuous models of their knowledge of relations in a domain did not improve their accuracy. It may be that the quantitative training and practice of most engineers makes the guidance unnecessary. There were, however, differences that are not reflected in the correlation that measures accuracy. The structure of the guided models was more complicated; the guided models were also better calibrated.

Comparison of Bootstrapped Models and Expert Models

Figure 2 presents the mean accuracy of the models the experts wrote in comparison with the models produced by fitting their judgments with a regression equation. Both the uncorrected and corrected versions of the experts' formulas are displayed. The nonguided and guided experts' models are collapsed together here. All bootstrapped models were weighted averages, whereas the experts' formulas could use other organizing principles. Generally, the models fit to the experts' judgments of highways displayed as abstract bar graphs were at least as accurate as the models the experts wrote, even when corrected. The models fit to the experts' judgments of filmstrips of the highways were less accurate for aesthetics and capacity, but more accurate for safety.

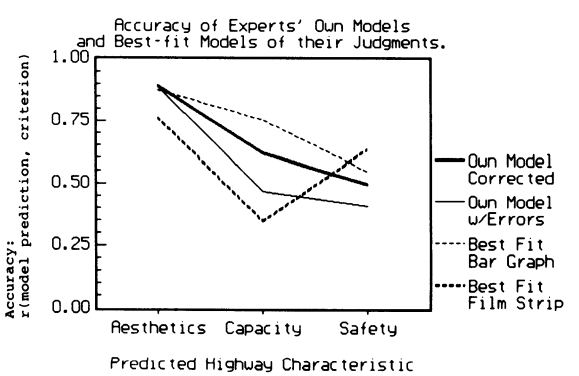

Figure 2. Accuracy of experts' own models and best-fit models of their judgments.

In predicting aesthetics, the experts' formulas (.893) and the bar-graph models (.873) were equally accurate, while the models fit to judgments of filmstrips were significantly less accurate [.759; compared with bar-graph models, $t(20) = 3.3$, $p = .004$; compared with experts' formulas, $t = 3.6$, $p = .002$], perhaps reflecting individual taste.

In predicting safety, the bootstrapped models were more accurate than the experts' formulas (corrected: .493). The filmstrip models (.638) were significantly more accurate [$t(20) = 4.3$, $p = .000$], while the bar-graph models (.543) were not ($t = 1.0$, $p = .34$). Only in this safety task were the filmstrip models more accurate than the bar-graph models ($t = 2.4$, $p = .028$).

The capacity results are interesting in that the experts' formulas (corrected: .620) were significantly more accurate than the filmstrip judgment models [347; $t(19) = 4.3$, $p = .000$], but significantly less accurate than the bar-graph judgment models (.752; $t = 2.4$, $p = .025$). The difference in the bootstrapped models' accuracy in safety and capacity is understandable if we take into account the nature of the cues. In safety, many of the cues are obvious in the filmstrips: curves, obstacles, lane width, and shoulder width (see Hammond et al., 1987). In contrast, some important cues to capacity are difficult to measure from the filmstrips (including, for example, the percent of automobiles vs. trucks in the traffic); therefore, providing measurements (in the bar-graph display) may increase the accuracy.

Diagnostic Versus Predictive Use of Cues

In the course of this study, a reason for inaccuracy in the formulas that experts write was discovered: When writing formulas, the experts neglected a "diagnostic" relation between an input variable and the output variable in favor of a "predictive" relation. They did not do so in judging cases from filmstrips.

The specific factor was *lane width* in the safety forecasting problem. A wide lane "predicts" a *safer* highway: cars have more room to maneuver. At the same time, a wide lane "diagnoses" a *more dangerous* highway: the Highway Department has widened the lanes more than usual in order to compensate for other factors that make the road particularly dangerous; even after widening, it is still dangerous. Because many engineers neglected the diagnostic relation in their formulas, the lane-width factor was used in the wrong direction, which contributed to the formulas' being less accurate than the models fit to the experts' judgments.

DISCUSSION

Linear models of the experts' judgments of individual highways, from premeasured cues, predicted the highway characteristics equally or more accurately than did the formulas the experts produced. This goes against popular conceptions about the power of analytical reason and the complicated nature of the world that we know. However, it is consistent with previous research that showed that linear models of experts' judg-

ments kept the "knowledge" and discarded the "noise" in their judgments (see Hammond, McClelland, & Mumpower, 1980).

A subset of the engineers were guided when writing formulas expressing their highway knowledge, using an elaborate procedure (analogous to a multiattribute utility assessment) to ensure that the model was coherent and adhered to general measurement principles. The resulting models were no more accurate than models produced without guidance. Although only three engineers were guided, they each did three tasks with similar results, which implies that the lack of difference would be general for this kind of task. It is interesting that correcting the experts' slips in writing formulas produced a greater gain in accuracy than did the elaborate formal guidance.

The linear models fit to two forms of judgment had different accuracies. When judging filmstrips, the expert both perceptually "measured" the input variables and integrated them into a judgment. When judging bar-graph profiles, the input variables were already measured and the expert just read them and integrated them into a judgment. The linear models of the bar-graph judgments predicted highway aesthetics and capacity at least as well as did the formulas the experts wrote, and more accurately than did the models of the filmstrip judgments (consistent with the findings of Lusk & Hammond, 1991). However, in the safety realm, the models of the filmstrip judgments, when the experts themselves measured the input variables, were more accurate. This may be due to differences between these realms in the visual obviousness of the input variables, but research is needed to discover what distinguishes those realms where more accurate models are produced using experts' judgments of expert-measured versus objectively measured cues.

REFERENCES

BAMBER, D. (1990). *Knowledge acquisition for the development of expert systems: An analysis* (Tech. Rep. No. 1322). San Diego, CA: Naval Ocean Systems Center.

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

HAMM, R. M. (1990). *Materials for guiding elicitation of self-reported formulas describing knowledge in a multi-object domain* (Tech. Rep. No. 90-14). Boulder, CO: University of Colorado, Institute of Cognitive Science.

HAMM, R. M. (1991). *Modeling expert forecasting knowledge for incorporation into expert systems* (Tech. Rep. No. 91-12). Boulder, CO: University of Colorado, Institute of Cognitive Science.

HAMMOND, K. R., HAMM, R. M., & GRASSIA, J. (1986). Generalizing over conditions by combining the multitrait multimethod matrix and the representative design of experiments. *Psychological Bulletin*, **100**, 257-269.

HAMMOND, K. R., HAMM, R. M., GRASSIA, J. L., & PEARSON, T. (1983). *Direct comparison of intuitive, quasi-rational, and analytic cognition* (Report No. 248). Boulder, CO: University of Colorado, Center for Research on Judgment and Policy.

HAMMOND, K. R., HAMM, R. M., GRASSIA, J., & PEARSON, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, & Cybernetics*, **SMC-17**, 753-770.

HAMMOND, K. R., MCCLELLAND, G. H., & MUMPOWER, J. (1980). *Human judgment and decision making*. New York: Praeger.

LUSK, C. M., & HAMMOND, K. R. (1991). Judgment in a dynamic task: Microburst forecasting. *Journal of Behavioral Decision Making*, **4**, 55-73.

NORMAN, D. A. (1981). Categorization of action slips. *Psychological Review*, **88**, 1-15.

VON WINTERFELDT, D., & EDWARDS, W. (1986). *Decision analysis and behavioral research*. New York: Cambridge University Press.

(Manuscript received May 20, 1991.)