ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 28, 379–394 (1981)

Causal and Diagnostic Inferences: A Comparison of Validity

MICHAEL BURNS AND JUDEA PEARL

University of California, Los Angeles

Decision-support technologies are founded on the paradigm that direct judgments are less reliable and less valid than synthetic inferences produced from more "fragmentary" judgments. Moreover, certain types of fragments are normally assumed to be more valid than others. In particular, judgments about the likelihood of a certain state of affairs given a particular set of data (diagnostic inferences) are routinely fabricated from judgments about the likelihood of that data given various states of affairs (causal inferences), and *not* vice versa. This study was designed to test the benefits of causal synthesis schemes by comparing the validity of causal and diagnostic judgments against "ground-truth" standards. The results demonstrate that the validity of causal and diagnostic inferences are strikingly similar; direct diagnostic estimates of conditional probabilities were found to be as accurate as their synthetic counterparts deduced from causal judgments. The reverse is equally true. Moreover, these accuracies were found to be roughly equal for each causal category tested. Thus, if the validity of judgments produced by a given mode of reasoning is a measure of whether it matches the format of human semantic memory, then neither one of the causal or diagnostic schema is a more universal or more natural format for encoding knowledge about common, everyday experiences. These findings imply that one should approach the "divide and conquer" ritual with caution; not every division leads to a conquest, even when the atoms are cast in causal phrasings. Dogmatic decompositions performed at the expense of conceptual simplicity may lead to inferences of lower quality than those of direct, unaided judgments.

Most decision-aiding technologies are based on the assumption that synthetic conclusions produced from "fragmentary" judgments are more valid than direct, unaided inferences. Quoting Slovic, Fischhoff, and Lichtenstein (1977):

> Most of these decision aids rely on the principle of divide and conquer. This "decomposition" approach is a constructive response to the problem of cognitive overload. The decision aid fractionates the total problem into a series of structurally related parts, and the decision maker is asked to make subjective assessments for only the small components. Such assessments are presumably simpler and more manageable than assessing more global entities. Research showing that decompo-

This work was supported in part by Office of Naval Research Contract N00014-78-C-0372 and, National Science Foundation Grant MCS 78-07468. Requests for reprints should be sent to Dr. Judea Pearl, Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California, Los Angeles, CA 90024. Michael Burns is now at Bell Laboratories, Piscataway, NJ.

0030-5073/81/060379-16$02.00/0

> sition improves judgment has been reported by Armstrong, Denniston, & Gordon (1975), Gettys, Michel, Steiger, Kelley, & Peterson (1973), and by Edwards, Phillips, Hayes, & Goodman (1968). (pp. 17–18)

A close look at decision-aiding techniques reveals that the structuring procedures that are used fall into three major categories: cascading, aggregation, and inversion. Cascading entails the chaining of a sequence of local relations $r_1(x_1, x_2), r_2(x_2, x_3), \ldots, r_{n-1}(x_{n-1}, x_n)$ to produce a global relation $R(x_1, x_n)$. Aggregation combines the relations $r(y, x_1), r(y, x_2), \ldots, r(y, x_n)$ to form $R[y, (x_1, x_2, \ldots, x_n)]$. Inversion entails converting the direction of certain relations to an easier format, e.g., $r_1(x, y)$ is converted to $r_2(y, x)$.

A typical example of *cascading* is involved when we wish to infer the consequence of a long sequence of actions. This inference is normally done by separately considering the effect of each individual action in the chain predicated upon the conditions created by the sequence that precedes it, then computing the overall effect by some formal rule. Another example is that of inferring the consequence of actions intertwined with a sequence of uncertain events. In the practice of decision analysis, the quality of actions is invariably inferred from judgments about the likelihood of the actions' consequences cascaded by judgments about the desirability of those consequences. Decision analysts seldom accept direct judgments about preferences on actions.

Aggregation is best exemplified by the task of assessing the cumulative impact of a large set of data on a given set of hypotheses. If the assessment task is complex, it may be helpful to assess separately the likelihood ratios for each individual datum, then aggregate these ratios by some normative rule (usually multiplication) to synthesize the overall joint impact. The synthetic conclusion is generally assumed to be more valid than that obtained by making a direct judgment of the joint impact of all the data (Edwards *et al.*, 1968).

The most prevalent example of *inversion* is the decision analysts' preference for expressing the linkage between evidence and a hypothesis in causal phrasings. Judgments about the likelihood of a certain hypothesis given a set of data (diagnostic inferences) are routinely synthesized from judgments about the likelihood of that data given various states of affairs (causal inferences), and *not* vice versa. The following set of examples illustrates the prevalence of this practice.

In the context of medical diagnosis, we find:

> It is more expedient to ask the physician to estimate simpler probabilities and then use Bayes' theorem to evaluate the desired probabilities. The simpler probabilities can be classified into two types: the prior and the identification probabilities. The priors are the probabilities which the doctor assigns to each of the four diseases before any new tests are performed. The identification probabilities are the prob-

abilities that the pathologist says disease j, given that the boy has disease i. (Ginsberg & Offensend, 1968, p. 360)

In a business management environment, we find an example of a decision maker who attempts to infer the probability that a steel industry strike is about to break out from the fact that his competitor is not preparing for one (Morris, 1968). Morris automatically assumes that the desired probability ought not to be assessed directly but rather should be obtained from a Bayes inversion of beliefs such as, "if in fact there will be no strike, the probability of this competitor's preparing for no strike is .40; the probabilities that he will be preparing for a 30-, 45-, or 90-day strike are .20, respectively" (p. 59). In the area of political situation assessment, we are shown an intelligence analyst evaluating the intent of Country A to develop an independent nuclear weapons production capability within the next 5 years (Decisions and Designs, Inc., 1973). The analyst in this setting is instructed to derive the desired conclusion by first quantifying such "elementary" beliefs as: If Country A has the given intention, there will be an increase in the Nuclear R&D Program; if that program is increased, there will be an increase in the use of nuclear material; etc. (Chap. 14, pp. 5–10). Presumably these causal beliefs facilitate more valid quantification than their diagnostic equivalents, e.g., an increase in the use of nuclear material suggests an increase in the Nuclear R&D Program.

Causality is an elusive concept, one loaded with philosophical difficulties. While a time precedence relationship would invariably hold between a cause and its effect, time precedence is by no means a sufficient condition for causality. To our knowledge, no attempt to give a precise definition of causality based entirely on nonpsychological concepts has been completely satisfactory. At the same time, causal relations are very common and fairly intelligible in ordinary human discourse. We rarely find two persons sharing a common heritage who disagree on cause–effect directionality.

Understandably, the causal/evidential distinction has not been explicitly identified among the glossary of tools employed by decision analysts. However, the examples cited above clearly demonstrate that this distinction tacitly controls the actual practice of decision analysis over a wide variety of problem domains. The distinction also has a strong influence over the structuring of aggregation procedures; the assumption of conditional independence among variables (permitting multiplications of likelihood ratios) is usually upheld when the condition is perceived as the cause and the variables as its manifestations. Conditional independence is rarely assumed when the condition is perceived as a supporting piece of evidence for the remaining variables.

While the experiments of Armstrong *et al.* (1975), Gettys *et al.* (1973), and Edwards *et al.* (1968) were directed toward verifying the benefit of cascading and aggregation, this paper focuses on the issue of causal/diagnostic inversion. Human preference for causal relations may come from the fact that in many cases we can associate P(manifestation|cause) with a stable physical entity while no such association is feasible for P(cause|manifestation). For example, when considering the reliability of a weather indicator we regard the quantity P(indicator reads "foul"|weather is "fair") as a stable property, inherent with the indicator's mechanism, a quantity that can be measured in isolation regardless of other factors. The quantity P(weather is "fair" indicator reads "foul"), on the other hand, depends on both the indicator's mechanism and the general weather conditions in that neighborhood and, therefore, appears harder to assess. This asymmetry also underlies the celebrated urn model (Raiffa, 1970) and is typical to many statistical applications where P(data|hypothesis) is obtained by computation from a so-called statistical model like the assumption that a set of observations is normally distributed (Edwards *et al.*, 1968). However, it is not at all clear whether any bias in favor of causal schema exists in cases where a parametric statistical model is not obvious and where both P(data|hypothesis) and P(hypothesis|data) are inferred by recalling various segments of human memory about everyday experiences.

Tversky and Kahneman (1977) indeed detected what they called "causal biases" in decision making. They showed that subjects perceive causal information to have a greater impact than diagnostic information of equal informativeness. Further, if some information has both causal and diagnostic implications, then subjects' judgments are "dominated" by the causal rather than the diagnostic relationship. Granted that causal reasoning is more emphasized in ordinary inference tasks, the question of the conditions under which the causal mode of reasoning will lead to *more valid* inferences still remains.

Aside from its psychological interest, this question has also acquired technological import. One application has already been alluded to, that of guiding the procedures used by decision analysts in eliciting likelihood estimates. The second application concerns organization of knowledge-based computer expert systems (Feigenbaum, 1977). In this latter application judgments from experts are encoded in the form of heuristic rules which are later combined to yield expert-like conclusions, explanations, and interpretations. The appropriate format for these fragmentary judgments is still subject to debate. Some knowledge-based systems (e.g., Shortliffe's MYCIN, 1976) insist on diagnostic inputs. Others (e.g., Ben-Bassat's MEDAS, 1980) require the more traditional causal judgments. The issue is whether experts, such as physicians, find it more comfortable

to estimate the likelihood of a disease given a set of symptoms or to evaluate the likelihood that a given disease be accompanied by a certain set of symptoms. Comfort aside, which form of input yields more valid therapeutical recommendations?

The experiment reported in this paper was designed to shed light on some of these issues. The problem of testing judgment validity, which has long been exacerbated by the lack of suitable criteria for measuring the quality of judgments about real-life experiences, was circumvented by "creating" our own ground-truth data.

METHOD

Subjects

One hundred seven undergraduate engineering students and 58 graduate students from various departments at UCLA participated in this study. The undergraduates, who were enrolled in one of two upper-level undergraduate engineering classes, served in the experiment as part of an in-class lecture. The graduate students were recruited via advertisements posted around the campus and in the campus newspaper, and they were paid according to the accuracy of their judgments.

Materials

The undergraduates participated in the first phase of the study. Their task was to answer 24 yes/no questions concerning their activities and beliefs; the answers provided the data base (ground truth) for the estimation phase which followed. The questions were of two types: "X questions" and "Y questions," equal in number and randomly ordered on the questionnaire. Each X query questioned a condition, activity, or belief considered by the experimenters to be a causal agent for a condition, activity, or belief specified in one Y query. For example, since the color of a person's eyes is perceived to be influenced by that person's parents, not vice versa, X may represent the event of a mother having blue eyes and Y may denote the condition of her daughter having blue eyes. Four categories of causal relations were employed: (1) genetic causality, where a genetic condition specified by the X question serves as a cause of the condition designated by the Y question; (2) training causality, where the X condition provides training for the Y activity; (3) habit-forming causality, where the X condition serves as a habit-forming agent for the behavior specified by the Y condition; and finally, (4) self-interest causality, where the X question defines a particular self-interest that leads to the belief unveiled by the Y question. Table 1 shows the four causal categories and the corresponding X and Y questions for each category.

The questions regarding the definition of causality (see Introduction) have prompted a separate pilot study aimed at verifying whether subjects

TABLE 1
CAUSAL AND DIAGNOSTIC ASSERTIONS USED FOR COMPILING THE DATA BASE

X Questions	*Y* Questions
Genetic causality	
1. Mother has blue eyes.	Student has blue eyes.
2. At least one of student's parents is left-handed.	Student is left-handed.
3. Student is a male over 5 ft 9 in.	Student played on high school basketball team.
Training causality	
4. Student took musical lessons as a child.	Student currently plays a musical instrument.
5. Student ran or jogged regularly in high school.	Student currently runs or jogs regularly.
6. Student took typing in high school.	Student types ≥40 words/min now.
Habit-forming causality	
7. Student attended church regularly in high school.	Student attends church regularly now.
8. Student is currently married.	Student is wearing a wedding ring.
9. Student's father was "handy" around home.	Student changes his own oil in his car.
Self-interest causality	
10. Student finds it financially difficult to complete his college studies	Student favors UCLA increasing financial aid at expense of larger classes.
11. Student's family finds medical expenses constitute a substantial burden.	Student favors nationalized medical-care plan.
12. Student closely follows UCLA football.	Student favors UCLA building on-campus football stadium.

perceived the X conditions to be causal agents for the Y conditions. One hundred five students were asked to identify the direction of causality in the 12 $X-Y$ conditions shown in Table 1. For each of the 12 pairs, Appendix 1 lists the percentage of subjects who identified X as the causal agent for Y. Clearly, there was almost complete agreement regarding the cause–effect directionality.

The data compiled from the undergraduates' responses served as the estimation targets in the second phase of the experiment. In this phase, the graduate students' task was to estimate the proportion of undergraduates responding in particular ways on the questionnaire. For a given $X-Y$ relation, each graduate student was instructed to estimate either a *causal triplet* or a *diagnostic triplet*. When estimating the *causal triplet*, the estimator first considered $P(X)$ (e.g., "What percentage of people said their mother had blue eyes?"), then $P(Y|X)$ (e.g., "What percentage of the people who said their mother had blue eyes also said they themselves

have blue eyes?"), and then $P(Y \mid \bar{X})$ (e.g., "What percentage of the people who said their mother did *not* have blue eyes said they themselves have blue eyes?"). In assessing a *diagnostic triplet*, the student first estimated $P(Y)$ (e.g., "What percentage of people said they have blue eyes?"), then $P(X|Y)$ (e.g., "What percentage of the people who said they have blue eyes said their mother also had blue eyes?"), and then $P(X|\bar{Y})$ (e.g., "What percentage of the people who said they do *not* have blue eyes said their mother had blue eyes?"). Note that the three components of each triplet represent statistically independent quantities and, moreover, that every component can be deduced from the three members of the opposing triplet via Bayes' theorem.

Procedure

The undergraduates answered the questionnaire during a regularly scheduled class meeting. The graduate students were assembled in groups ranging in size from 4 to 15 persons. Before they began the task, the graduate students were told about the nature of the estimations they would be making, and about the "pay scale" which was dependent on the proximity of their estimates to the actual proportions computed from the undergraduates' responses. Since people tend to be easily confused between conditional probabilities and joint probabilities, special care was exercised to ensure that subjects thoroughly understood the meaning of the statement, "the percentage of people who said X who also said Y." Subjects were explicitly told to think only of the population of those persons who said X and to estimate the proportion of those who said Y among this population. These instructions were accompanied by Venn diagrams depicting these proportions and were followed by an actual numerical example.

Half of the graduate students estimated causal triplets for odd-numbered $X-Y$ relations and diagnostic triplets for even-numbered relations. For the other half of the subjects, this pattern was reversed.[1] Each graduate student estimated one triplet for each of the 12 relations, thus making a total of 36 probability estimates. The estimators were given as much time as needed to contemplate the estimates required. Most of the students took between 20 and 30 min to complete the task.

RESULTS AND DISCUSSION

The task of evaluating judgment validity requires a choice of a validity criterion. A variety of criteria has been proposed and utilized for measur-

[1] The responses to the football stadium question of 33 graduate students were deleted from the data analysis because of an inadvertent discrepancy between the wording of the questions posed to those subjects and the wording of the corresponding question given to the undergraduates in the first phase of the study.

ing the degree of disparity between a given actual proportion P_a and an estimate P_e of that proportion (Pearl, 1978). We have examined both the quadratic error,

$$Q = (P_e - P_a)^2 \tag{1}$$

and the logarithmic error,

$$L = P_a \log P_a/P_e + (1 - P_a) \log (1 - P_a)/(1 - P_e).$$

Both gave rise to practically identical patterns, so this paper will present data based on the quadratic error only.

For each query we took $\bar{Q}$, the arithmetic mean of the quadratic errors across subjects, as a measure of the inaccuracy of the corresponding estimate. These mean quadratic errors, along with the actual proportions and mean estimates, are shown in Table 2. These estimates are called *direct estimates* to distinguish them from *synthetic estimates*, which will be discussed later.

Table 2 reflects a slight tendency for the mean estimates to regress toward the .50 probability level in relation to the actual probability. That is, in 65% of the cases, the proportions were actually "more extreme" (closer to .00 or 1.00) than their associated estimates. This effect is more apparent in Fig. 1, which displays the relationship between the actual proportions (along the horizontal axis) and their associated estimates (along the vertical axis).

By and large, one cannot detect a marked difference in accuracy between causal estimates (i.e., $P_e(Y|X)$ and $P_e(Y|\bar{X})$) and their diagnostic counterparts (i.e., $P_e(X|Y)$ and $P_e(X|\bar{Y})$). In Fig. 1, for example, where accuracy is reflected by proximity to the diagonal line, the two families of estimates appear equally dispersed. However, such a comparison is not entirely reliable. Since the values of the actual proportions $P_a(Y|X)$ are generally smaller than those of $P_a(X|Y)$, a direct comparison between their estimates may not reflect true differences in validity. An estimation error in the neighborhood of $P = .50$ is far less severe than an error of equal magnitude near the extremes (.00 and 1.00). On four of the $X-Y$ relations the actual proportions $P_a(Y|X)$ and $P_a(X|Y)$ are fairly close to one another (within .15). In all four cases $P_e(Y|X)$ is at least slightly more accurate than $P_e(X|Y)$, lending some support to the hypothesis that causal reasoning leads to better inference making than diagnostic reasoning. However, if the same procedure is employed with the $P_e(Y|\bar{X})$ estimate (invoking causal reasoning) and the $P_e(X|\bar{Y})$ estimate (based on diagnostic reasoning), only 3 of the 7 comparisons show an advantage for $P_e(Y|\bar{X})$. Since the difference between causal and diagnostic reasoning in these 11 comparisons is generally of small magnitude, there is not a noticeable advantage for the former, as had been anticipated.

TABLE 2

ACTUAL PROPORTIONS, MEAN ESTIMATES, AND MEAN QUADRATIC ERRORS FOR DIRECT ESTIMATES

	$P(X)$			$P(Y)$			$P(Y\|X)$			$P(X\|Y)$			$P(Y\|\bar{X})$			$P(X\|\bar{Y})$		
$X-Y$ relation	Prop	Mean	$\bar{Q}$	Prop	Mean	$\bar{Q}$	Prop	Mean	$\bar{Q}$	Prop	Mean	$\bar{Q}$	Prop	Mean	$\bar{Q}$	Prop	Mean	$\bar{Q}$
1. Blue eyes	.20	.34	.053	.22	.25	.024	.76	.60	.067	.67	.49	.072	.09	.23	.048	.06	.26	.072
2. Left-handed	.15	.24	.037	.05	.17	.027	.19	.43	.115	.60	.46	.098	.02	.20	.064	.13	.17	.017
3. Basketball	.39	.38	.037	.08	.15	.041	.12	.26	.082	.56	.70	.091	.06	.11	.032	.38	.37	.042
4. Musical instrument	.56	.49	.054	.24	.34	.069	.33	.41	.062	.77	.66	.072	.13	.13	.009	.49	.34	.108
5. Running	.34	.30	.036	.20	.38	.094	.36	.74	.183	.62	.55	.070	.11	.29	.056	.27	.23	.034
6. Typing	.56	.39	.081	.23	.24	.031	.25	.48	.132	.60	.79	.066	.21	.22	.037	.55	.28	.115
7. Church	.36	.30	.041	.22	.28	.032	.49	.54	.056	.79	.78	.047	.07	.10	.013	.24	.34	.045
8. Wedding ring	.06	.16	.034	.04	.15	.040	.50	.68	.123	.75	.82	.079	.01	.02	.000	.03	.16	.048
9. Repairs	.74	.52	.097	.64	.37	.142	.66	.55	.086	.76	.62	.093	.57	.34	.107	.69	.47	.119
10. Financial aid	.34	.48	.077	.38	.39	.067	.44	.62	.144	.39	.69	.148	.35	.28	.075	.30	.32	.070
11. Medical care	.26	.40	.106	.50	.57	.056	.57	.66	.072	.30	.68	.198	.48	.33	.085	.23	.24	.056
12. Football[a]	.46	.60	.046	.42	.59	.121	.57	.73	.057	.62	.81	.063	.29	.34	.063	.34	.18	.042

[a] Total of 25 subjects answered this question.

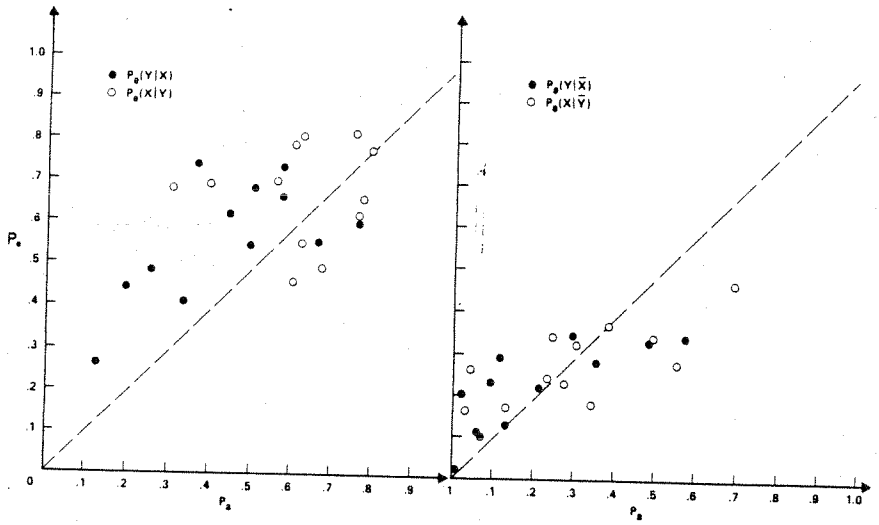

FIG. 1. Mean estimates versus actual proportions.

Another way to circumvent the "apples versus oranges" difficulty is to synthesize causal and diagnostic estimates that can be compared on equal ground. To do this we aggregated subjects' estimates by Bayes' theorem to calculate synthetic estimates according to the following equations:

$$P_s(X) = P_e(X|Y)\,P_e(Y) + P_e(X|\bar{Y})\,[1 - P_e(Y)], \tag{2}$$

$$P_s(Y) = P_e(Y|X)\,P_e(X) + P_e(Y|\bar{X})\,[1 - P_e(X)], \tag{3}$$

$$P_s(Y|X) = \frac{P_e(X|Y)\,P_e(Y)}{P_s(X)}, \tag{4}$$

$$P_s(X|Y) = \frac{P_e(Y|X)\,P_e(X)}{P_s(Y)}, \tag{5}$$

$$P_s(Y|\bar{X}) = \frac{[1 - P_e(X|Y)]\,P_e(Y)}{[1 - P_e(X|Y)]\,P_e(Y) + [1 - P_e(X|\bar{Y})]\,[1 - P_e(Y)]}, \tag{6}$$

$$P_s(X|\bar{Y}) = \frac{[1 - P_e(Y|X)]\,P_e(X)}{[1 - P_e(Y|X)]\,P_e(X) + [1 - P_e(Y|\bar{X})]\,[1 - P_e(X)]}. \tag{7}$$

Note that the synthetic estimates in (2), (4), and (6) should be regarded as diagnostic since they are deduced from diagnostic inputs. Similarly, the estimates constructed by formulas (3), (5), and (7) are causal.

Furthermore, the synthetic estimates are more reflective of the transformations employed by common decision analysis procedures. For example, formula (5) represents the celebrated transformation from prior to posterior which was pioneered (posthumously) by Reverend Bayes in

1761 as a means to infer the "probability of causes." It has since become almost a ritual to assume that this transformation automatically produces more valid judgments than the direct estimate $P_e(X|Y)$.

Table 3 shows the mean quadratic error, $\bar{Q}$, for both the direct estimate and the synthetic estimate for each of the four conditional probabilities. The direct estimates for $P(Y|X)$ and $P(Y|\bar{X})$ involve causal reasoning and the direct estimates for $P(X|Y)$ and $P(X|\bar{Y})$ involve diagnostic reasoning, while this relationship reverses when the synthetic estimates are considered.

Also shown is an indicator called the *normalized error difference* which gives a measure of the significance of the difference between the direct estimate and the synthetic estimate. It was computed by the formula

$$\text{normalized error difference} = \frac{(n)^{1/2}(\bar{Q}_{\text{diagnostic}} - \bar{Q}_{\text{causal}})}{(\sigma^2_{\text{causal}} + \sigma^2_{\text{diagnostic}})^{1/2}}, \qquad (8)$$

where $\bar{Q}_{\text{diagnostic}}$ and $\bar{Q}_{\text{causal}}$ stand for the mean quadratic error across subjects for either the direct or synthetic estimates, as appropriate, and σ^2 represents the variance of those quadratic errors. One property of this normalized error difference is rather obvious: Its value is made increasingly positive when the validity of the causal estimate becomes significantly greater than that of the diagnostic estimate, and negative when the reverse is true. Clearly, since the same actual proportion applies to both the direct estimate and the synthetic estimate for a particular probability, the "apples versus oranges" problem is eliminated.

Across all estimates, there are nine instances where the normalized error difference is significant at the .05 level according to a standard two-tailed t distribution. In six of these cases, it is the causal estimate that is better than the diagnostic, which leaves three cases in which the diagnostic is better. Thus, there is little evidence in these data for the superiority of causal reasoning over diagnostic reasoning. In fact, only one of the problems (musical instrument) shows a positive normalized error difference for all four conditional probabilities, while one other problem (typing) has a negative normalized error difference for all four conditionals.

Aside from comparing causal and diagnostic estimates, Table 3 also enables us to compare the validities of direct versus synthetic estimates. A suspicion that the latter may be more valid than the former could be based on the argument that each synthetic estimate combines the output of three knowledge sources. If these were independent mental processes in the sense that the estimator providing them would consult different data or invoke different procedures for their production, then one would be justified in hypothesizing superiority for synthetic estimates over their

TABLE 3
MEAN QUADRATIC ERRORS FOR DIRECT VERSUS SYNTHETIC ESTIMATES

$X-Y$ Relation	$P(Y\|X)$			$P(Y\|\bar{X})$			$P(X\|Y)$			$P(X\|\bar{Y})$		
	Dir.	Syn.	NED	Dir.	Syn.	NED	Dir.	Syn.	NED	Dir.	Syn.	NED
1. Blue eyes	.067	.156	**2.842**	.048	.044	−.137	.072	.058	.553	.072	.027	1.537
2. Left-handed	.115	.060	**−2.460**	.064	.015	**−3.954**	.098	.093	.216	.017	.035	−1.453
3. Basketball	.082	.067	−.503	.032	.015	−.625	.091	.062	.968	.042	.042	0
4. Musical instrument	.062	.130	**2.504**	.009	.059	**2.608**	.072	.046	1.179	.108	.069	2.034
5. Running	.183	.140	−1.530	.056	.064	.403	.070	.053	.542	.034	.044	−.504
6. Typing	.132	.117	−.454	.037	.024	−.461	.066	.067	−.031	.115	.120	−.177
7. Church	.056	.060	.174	.013	.037	1.004	.047	.068	−.910	.045	.029	.669
8. Wedding ring	.123	.077	**−2.227**	.000	.014	1.538	.079	.036	**2.059**	.048	.023	**2.747**
9. Repairs	.086	.135	1.538	.107	.143	1.567	.093	.070	.782	.119	.131	−.522
10. Financial aid	.144	.114	−1.071	.075	.044	−1.256	.148	.151	−.104	.070	.067	.122
11. Medical care	.072	.082	.255	.085	.075	−.317	.198	.159	1.140	.056	.061	−.159
12. Football[a]	.057	.125	**2.393**	.063	.092	.703	.063	.047	−.847	.042	.057	−.695

Note. Dir. = direct estimates; Syn. = synthetic estimates; NED = normalized error difference. The boldface entries indicate cases where the normalized error difference is significant at the .05 level. NEDs in boldface type indicate significance at the .05 level according to two-tailed t distribution.

[a] Total of 25 subjects answered this question.

direct counterparts. Comparing the data, one finds that in five of the nine significant cases, synthetic estimates are better than their direct counterparts.

Table 4 shows the mean quadratic errors for direct estimates and synthetic estimates with questions grouped according to the type of causality implied in the $X-Y$ relations. These were obtained by averaging the quadratic errors over the $X-Y$ relations within each causal category. In general, genetic relations induce slightly more accurate estimates than do training and habit-forming relations, while self-interest relations induce the worst estimates of all. This pattern is true for both direct estimates and synthetic estimates. For each causality category the synthetic estimates are more valid than the direct estimates of $P(X|Y)$, and less valid for $P(Y|X)$. In each case, the more valid estimates are those based on causal reasoning, a fact which lends support to the conjecture that causal reasoning is more naturally invoked in interpreting common observations. However, when considering the other four columns ($P(X)$, $P(Y)$, $P(Y|\bar{X})$, $P(X|\bar{Y})$), the pattern of results no longer reflects causal superiority.

CONCLUSIONS

Admittedly, having ourselves adhered to the belief that causal reasoning is a more natural mode of inference making, we were somewhat surprised that the results do not show a stronger validity differential in this direction. Taking Table 3, for example, the overall mean of the normalized error difference is equal to .25, which clearly does not support the hypothesis of general causal superiority. In the few $X-Y$ relations where significant validity differentials were detected, there was not a sizable bias favoring the causal mode. Thus, if the validity of judgments produced by a given mode of reasoning is a measure of whether that mode matches the format of human semantic memory, then neither the causal nor diagnostic schema is a more universal or more natural format for encoding knowledge about common, everyday experiences. It appears that semantic memory contains both causal schema and diagnostic schema. The choice of which schema to invoke for a particular observational relation may depend on the nature of the relation, the anticipated mode of usage, and the level of training or familiarity of the observer.

These findings imply that one should approach the "divide and conquer" ritual with caution; not every division leads to a conquest, even when the resultant atoms are cast in causal phrasings. Forced transformations from diagnostic to causal judgments performed at the expense of conceptual simplicity may lead to inferences of lower quality than direct, "holistic" judgments.

TABLE 4
MEAN QUADRATIC ERRORS FOR DIFFERENT CAUSAL CATEGORIES

	$P(X)$		$P(Y)$		$P(Y\|X)$		$P(X\|Y)$		$P(Y\|\bar{X})$		$P(X\|\bar{Y})$		Mean	
	Dir.	Syn.	Dir.	Syn.	Dir.	Syn.	Dir.	Syn.	Dir.	Syn.	Dir.	Syn.	Dir.	Syn.
Genetic	.042	.038	.031	.056	.088	.094	.087	.071	.048	.025	.044	.035	.057	.053
Training	.057	.057	.065	.050	.126	.129	.069	.055	.034	.037	.086	.078	.073	.068
Habit-forming	.057	.073	.071	.043	.088	.091	.073	.058	.040	.060	.071	.061	.067	.064
Self-interest	.076	.077	.081	.062	.091	.107	.136	.119	.074	.070	.051	.062	.086	.083
Mean	.058	.061	.062	.053	.098	.105	.091	.076	.049	.048	.065	.059		

Note. Dir. = direct estimates; Syn. = synthetic estimates.

APPENDIX 1

One hundred and five subjects answered the following question:

> For each of the following pairs of actions or conditions, please put a check by the one which you judge to be the cause of the other. Please make one check per pair.

The percentage of subjects judging the directions of causality to match those of Table 1 is listed below:

Blue eyes	97%
Left-handed	90%
Basketball	91%
Musical instrument	95%
Running	94%
Typing	99%
Church	94%
Wedding ring	91%
Repairs	97%
Financial aid	94%
Medical care	95%
Football	95%
Overall	95%

REFERENCES

Armstrong, J. S., Denniston, W. B., Jr., & Gordon, M. W. The use of the decomposition principle in making judgments. *Organizational Behavior and Human Performance*, 1975, **14**, 257-263.

Ben-Bassat, M., *et al.* Pattern-based interactive diagnosis of multiple disorders: The MEDAS system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1980, **PAMI-2**, 148-160.

Decisions and Designs, Inc. Hierarchical inference. In *Handbook for decision analysis*. McLean, Va.: Decisions and Designs, Inc., 1973. Chap. 14.

Edwards, W., Phillips, L. D., Hayes, W. L., & Goodman, B. G. Probabilistic information processing systems: Design and evaluation. *IEEE Transactions on Systems Science and Cybernetics*, 1968, **SSC-4**, 248-265.

Feigenbaum, E. A. The art of artificial intelligence. I. Themes and case studies of knowledge engineering. *Proceedings of the 5th International Joint Conference on Artificial Intelligence*, 1977, **2**, 1014-1029.

Gettys, C., Michel, C., Steiger, J. H., Kelley, C. W., & Peterson, C. R. Multiple-stage probabilistic information processing. *Organizational Behaviour and Human Performance*, 1973, **10**, 374-387.

Ginsberg, A. S., & Offensend, F. L., An application of decision theory to a medical diagnosis-treatment problem. *IEEE Transactions on Systems Science and Cybernetics*, 1968, **SSC-4**, 360.

Morris, W. T. *Management science*. Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Pearl, J. An economic basis for certain methods of evaluating probabilistic forecasts. *International Journal of Man-Machine Studies*, 1978, **10**, 1975–183.

Raiffa, H. *Decision analysis*. Redding, Mass.: Addison–Wesley, 1970.

Shortliffe, E. H. *Computer-based medical consultations: MYCIN*. New York: Elsevier, 1976.

Slovic, P., Fischhoff, B., & Lichtenstein, S. Behavioral decision theory. *Annual Review of Psychology*, 1977, **28**, 1–39.

Tversky, A., & Kahneman, D. Causal schemata in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology*. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.

RECEIVED: June 26, 1980