COGNITIVE PSYCHOLOGY 18, 253–292 (1986)

The Effects of Statistical Training on Thinking about Everyday Problems

GEOFFREY T. FONG

Northwestern University

DAVID H. KRANTZ

Columbia University

AND

RICHARD E. NISBETT

The University of Michigan

People possess an abstract inferential rule system that is an intuitive version of the law of large numbers. Because the rule system is not tied to any particular content domain, it is possible to improve it by formal teaching techniques. We present four experiments that support this view. In Experiments 1 and 2, we taught subjects about the formal properties of the law of large numbers in brief training sessions in the laboratory and found that this increased both the frequency and the quality of statistical reasoning for a wide variety of problems of an everyday nature. In addition, we taught subjects about the rule by a "guided induction" technique, showing them how to use the rule to solve problems in particular domains. Learning from the examples was abstracted to such an extent that subjects showed just as much improvement on domains where the rule was not taught as on domains where it was. In Experiment 3, the ability to analyze an everyday problem with reference to the law of large numbers was shown to be much greater for those who had several years of training in statistics than for those who had less. Experiment 4 demonstrated that the beneficial effects of formal training in statistics may hold even when subjects are tested completely outside of the context of training. In general, these four experiments support a rather "formalist" theory of reasoning: people reason using very abstract rules,

This research was supported by NSF Grants SES82-18846 and SES85-07342, and ONR Grant 442PT/85-2281 to Richard Nisbett. Geoffrey T. Fong was supported by a CIC-Lilly Foundation Minority Graduate Fellowship and a Rackham Fellowship at the University of Michigan. Experiment 1 was conducted while David Krantz was at AT&T Bell Laboratories, whose support is gratefully acknowledged. We thank Michael Smith, Bonnie Thompson, Alison Frank, Roseann O'Rourke, and Kathy Wandersee for their assistance and Edward Rothman for his help on Experiment 4. We thank Reid Hastie, Darrin Lehman, Ziva Kunda, and Saul Sternberg for their comments on previous drafts of this article. And finally, we thank Drake's Candy Store in Ann Arbor, Michigan, for supplying the gumballs used in Experiments 1 and 2. Requests for reprints should be sent to Geoffrey T. Fong, Department of Psychology, Northwestern University, Evanston, IL 60201.

0010-0285/86 $7.50

and their reasoning about a wide variety of content domains can be affected by direct manipulation of these abstract rules.

Do people solve inferential problems in everyday life by using abstract inferential rules or do they use only rules specific to the problem domain? The view that people possess abstract inferential rules and use them to solve even the most mundane problems can be traced back to Aristotle. In modern psychology, this view is associated with the theories of Piaget and Simon. They hold that, over the course of cognitive development, people acquire general and abstract rules and schemas for solving problems. For example, people acquire rules that correspond to the laws of formal logic and the formal rules of probability theory. Problems are solved by decomposing their features and relations into elements that are coded in such a way that they can make contact with these abstract rules.

This formalist view has been buffeted by findings showing that people violate the laws of formal logic and the rules of statistics. People make serious logical errors when reasoning about arbitrary symbols and relations (for a review, see Evans, 1982). The best known line of research is that initiated by Wason (1966) on his selection task. In that task, subjects are told that they will be shown cards having a letter on the front and a number on the back. They are then presented with cards having an A, a B, a 4, and a 7 and asked which they would have to turn over in order to verify the rule, "If a card has an A on one side, then it has a 4 on the other." This research showed that people do not reason in accordance with the simple laws of conditional logic, which would require turning over the A and the 7. Subsequent work showed that people do reason in accordance with the conditional for certain concrete and familiar problems. For example, when people are given envelopes and asked to verify the rule, "If the letter is sealed, then it has a 50-lire stamp on it," they have no trouble with the problem (Johnson-Laird, Legrenzi, & Sonino-Legrenzi, 1972). Many investigators have concluded from results of the latter sort that people do not use abstract rules of logic when solving concrete problems. Instead, people use only domain-specific rules (e.g., D'Andrade, 1982; Golding, 1981; Griggs & Cox, 1982; Johnson-Laird et al., 1972; Manktelow & Evans, 1979; Reich & Ruth, 1982). If people solve a problem correctly, it is because they are sufficiently familiar with the content domain to have induced a rule that allows them to solve problems in that domain.

Research on inductive reasoning has followed a similar history. Kahneman and Tversky (e.g., 1971, 1973; Tversky & Kahneman, 1974) demonstrated that people fall prey to a multitude of failures to employ statistical rules when reasoning about everyday life problems. In particular, people often fail to reason in accordance with the law of large numbers, the regression principle, or the base rate principle. (For reviews see Einhorn & Hogarth, 1981; Hogarth, 1980; Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980).

We and our colleagues, however, have shown that people do use statistical concepts in solving particular kinds of problems in particular domains (Jepson, Krantz, & Nisbett, 1983; Nisbett, Krantz, Jepson, & Fong, 1982; Nisbett, Krantz, Jepson, & Kunda, 1983). For example, Jepson et al. (1983) presented subjects with a variety of problems drawn from three very broad domains. All of the problems dealt with events that are variable and, as such, can be analyzed in terms of statistical concepts such as sample size. One domain examined by Jepson et al. consisted of problems for which the random nature of the sample is obvious. In one problem, for example, the protagonist has to judge characteristics of a lottery. As expected, the great majority of the answers for these "probabilistic" problems were statistical answers, that is, they incorporated intuitive notions of the law of large numbers or the regression principle in their answer. At the other extreme, a different group of problems dealt with subjective judgments about the properties of some object or person. In one of these problems, for example, the protagonist has to decide which of two college courses he should take, either on the basis of one visit to each class or on the basis of the evaluations of students who took the courses the previous term. Statistical responses were relatively rare for these "subjective" problems, constituting only about a quarter of the total. In between these extremes, there were a number of problems that, while not containing broad hints as to the random nature of the events in question, dealt with events that are of a sufficiently objective nature that it is relatively easy to recognize that they are characterized by a degree of random variation. These problems dealt primarily with athletic events and academic achievements. For these "objective" problems, slightly more than half of the answers were statistical in nature.

Nisbett et al. (1983) interpreted these and similar results as reflecting the fact that people possess intuitive but abstract versions of statistical rules. They called these intuitive rules "statistical heuristics," and argued that people call on such heuristics to the degree that (a) problem features are readily coded in terms of statistical rules, that is, when the sample space and sampling process are clear, and when the events can be coded in common units (as is the case for athletic events and academic achievements, for example); (b) the presence of chance factors or random variation is signaled by the nature of the events or by other cues in the problem; and (c) the culture recognizes the events in question as being associated with random variation (for example, gambling games) and thus prescribes that an adequate explanation of such events should make reference to statistical principles.

This account presumes that statistical heuristics are abstract. It explains people's frequent failures to use abstract rules as being the result of difficulty in coding problem elements in terms that trigger the rules or as the result of the presence of competing heuristics. But the evidence to date does not rule out the view that statistical heuristics are not abstract at all, but rather are local, domain-bound rules that happen to overlap with formal statistical rules. These rules are better developed in some domains than in others, and it is for this reason that people are much more likely to give statistical answers for some problems than others.

If statistical heuristics are abstract, then it should be possible to improve people's statistical reasoning about everyday events by formal instruction in the rule system, without reference to any domain of everyday events. Such abstract instructional methods should help people apply the rules over a broad range of problem content. On the other hand, if such formal instruction fails to help people to solve concrete problems, despite the fact that people can be shown to have learned a substantive amount about the formal properties of the rules, this would be discouraging to the formal view. It would also be discouraging to the formal view if it were to turn out that abstract instruction affects only people's solution of probabilistic problems, where the relevance of statistical rules is obvious, and where competing rules have relatively little strength.

In order to test the view that formal training per se results in an increase in people's use of statistical principles across a variety of domains, we trained subjects, in brief but intensive laboratory sessions, on the concepts associated with the law of large numbers. We then presented them with a number of problems in each of three broad domains, dealing, respectively, with events generally construed as probabilistic, with objectively measurable events, and with events that are measurable only by subjective judgments.

We also tested the formal view in another way. Some subjects were not given formal instruction, but instead were shown how to apply the law of large numbers for three concrete example problems, all of which dealt with objectively measurable events. If subjects are capable of inducing generalized rules of some degree of abstraction from such training, then they might be expected to reason more statistically about problems in the other domains as well, even though they have not been presented with examples in those domains. Whereas the empirical view suggests that statistical training will be domain specific, with training in one domain failing to generalize to other domains, the formalist view predicts that statistical training in one domain should generalize readily to other domains.

All of the problems presented to subjects concerned everyday life events and were of a type that, in previous work, we have found at least some subjects answer in a statistical fashion. All questions were open ended, and we coded the written answers according to a system that distinguished among varying degrees of statistical thinking. This procedure provided us with a great deal of information about how people reason about events in everyday life and allowed us to determine whether training can enhance not only the likelihood of employing statistical concepts, but also the likelihood that those concepts will be employed properly.

EXPERIMENT 1

Testing Method

Subjects' intuitive use of statistical reasoning was tested by examining their answers to 15 problems to which the law of large numbers could be applied and 3 for which the law of large numbers was not relevant. In this section we describe the instructions that introduced the test problems, the design of the 18 problems, and the system of coding the open-ended answers. The actual text of the problems is given in Appendix A.

Instructions

The instructions for the control subjects read as follows:

> We are interested in studying how people go about explaining and predicting events under conditions of very limited information about the events. It seems to us to be important to study how people explain and predict under these conditions because they occur very frequently in the real world. Indeed, we often have to make important decisions based on such explanations and predictions, either because there is too little time to get additional information or because it is simply unavailable.
>
> On the pages that follow, there are a number of problems that we would like you to consider. As you will see, they represent a wide range of real-life situations. We would like you to think carefully about each problem, and then write down answers that are sensible to you.

For groups that received training, the first paragraph of the above instructions was presented as part of the introduction to the training materials. After the training, the test booklet was introduced by the second paragraph, which ended with the sentence, "In many of the problems, you may find that the Law of Large Numbers is helpful."

Problem Types and Problem Structure

The 18 problems were divided into three major types as follows:

Type 1. Probabilistic. In these six problems, subjects had to draw conclusions about the characteristics of a population from sample data generated in a way that clearly incorporated random variation. Randomness was made clear in various ways: by the explicitly stated variation in sample outcomes (for example, the number of perfect welds out of 900 made by a welding machine ranged from 680 to 740), by including in the problem a random generating device (for example, shaking a jar of pennies before drawing out a sample), or by simply stating that a sample was "random."

Type 2. Objective. In these six problems, subjects had to draw conclusions about characteristics of a population on the basis of "objective" sample data but with no explicit cue about randomness of the data. One problem, for example, asked subjects to decide which of two makes of car was more likely to be free of troublesome repairs, on the basis of various facts about the repair records. Other problems dealt with the outcomes of athletic events and with academic accomplishments.

Type 3. Subjective. In these six problems, subjects had to draw conclusions about subjective characteristics of a population from "subjective" sample data. In one problem, for example, a high school senior had to choose between two colleges. The underlying subjective characteristic in this problem was liking for the two schools and the data consisted of his own and his friends' reactions to the schools.

In order to systematize the kinds of problems we presented to subjects across the three domains, we selected six different underlying problem structures and for each structure we wrote one problem of each of the above three types. The structures varied in types of samples drawn, type of decision required, and type of competing information.

Structure 1 problems required subjects to draw conclusions about a population from a single small sample. Structure 2 problems pitted a small sample against a large sample. Structure 3 problems required subjects to explain why an outcome selected because of its extreme deviation was not maintained in a subsequent sample (i.e., regression). Structure 4 problems were similar to those in Structure 2, except that the large sample was drawn from a population that was related to, although not identical to, the target population. Structure 5 problems pitted a large sample against a plausible theory that was not founded on data. Structure 6 (false alarm) problems involved conclusions drawn from a sample that was large, but also highly biased. As such, criticism or arguments in these problems should be based on the sample *bias,* but not on sample size. We included these problems to determine whether subjects who received training on the law of large numbers would then proceed to invoke it indiscriminately, or if they would apply it only to the problems of Structures 1–5, for which it was genuinely relevant.

In short, the 18 test problems followed a 3 × 6 design, with problem type crossed with problem structure. The order of the 18 test problems was randomized for each subject, with the constraint that no 2 problems with the same structure appeared successively.

Coding System

To study the use of statistical reasoning, a simple 3-point coding system was developed for the 15 problems for which the law of large numbers was applicable (Structures 1–5). To illustrate this coding system, we present examples of responses to the "slot machine problem," the probabilistic version of Structure 2 (small sample vs large sample). The protagonist of the story, Keith, was in a Nevada gas station where he played two slot machines for a couple of minutes each day. He lost money on the left slot machine and won money on the right slot machine. Keith's result, however, ran counter to the judgment of an old man sitting in the gas station, who said to Keith, "The one on the left gives you about an even chance of winning, but the one on the right is fixed so that you'll lose much more often than you'll win. Take it from me—I've played them for years." Keith's conclusion after playing the slot machines was that the old man was wrong about the chances of winning on the two slot machines. Subjects were asked to comment on Keith's conclusion. Every response to the test problems was classified into one of three categories:

1 = an entirely deterministic response, that is, one in which the subject made no use of statistical concepts. In responses of this type, there was no mention of sample size, randomness, or variance. The following was coded as a deterministic response to the slot machine problem: "Keith's reasoning was poor, provided the information given by the man was accurate. The man, however, may have been deceiving Keith."

2 = a poor statistical response. Responses given this score contained some mention of statistical concepts, but were incomplete or incorrect. These responses contained one or more of the following characteristics: (1) the subject used both deterministic and statistical reasoning, but the deterministic reasoning was judged by the coder to have been preferred by the subject; (2) the subject used incorrect statistical reasoning, such as the gambler's fallacy; (3) the subject mentioned luck or chance or the law of large numbers but was not explicit about how the statistical concept was relevant. The following is an example of a poor statistical response to the slot machine problem:

> I think that Keith's conclusion is wrong because the old man had better luck on the left one, so he thought it was better. Keith had better luck on the right one so he thought it was better. I don't think you could have a better chance on either one.

3 = a good statistical response. Responses given this score made correct use of a statistical concept. Some form of the law of large numbers was used, and the sampling elements were correctly identified. If the subject used both deterministic and statistical reasoning, the statistical reasoning was judged by the coder to have been preferred by the subject. In general, the subject was judged to have clearly demonstrated how the law of large numbers could be applied to the problem. The following was coded as a good statistical response to the slot machine problem:

> Keith's conclusion is weak. He is wrong in making the assumptions against the old man. Keith is judging the machines on only a handful of trials and not with the sample number the old man has developed over the years. Therefore, Keith's margin of error is much more great than the old man's.

The coding system thus distinguished each response on the basis of whether or not a statistical concept had been used and, within the class of statistical responses, whether or not it was a "good" statistical response, that is, one that showed a correct use of the law of large numbers.

Such coding obviously runs into borderline cases. A coding guidebook was created which documented the principal types of borderline cases and the recommended treatment of them, for each problem. Reliability was tested by having four coders code a sample of 20 test booklets (300 law of large numbers problems). There was exact agreement among all four coders on 86% of these responses. Having achieved a high level of reliability, the primary coder (who had been one of the four coders), coded all of the responses, blind to conditions. His coding comprised the data we present here and in Experiment 2.

The coding of the three Structure 6 (false alarm) problems is described in a separate section below.

Training Procedures

All training procedures began with an introductory paragraph about decisions with limited information (quoted in full above as the first paragraph in the testing instructions for the control subjects).[1] Next followed a paragraph introducing the law of large numbers. This always began as follows:

> Experts who study human inference have found that principles of probability are helpful in explaining and predicting a great many events, especially under conditions of limited information. One such principle of probability that is particularly helpful is called the *Law of Large Numbers.*

[1] All training materials can be obtained from the authors.

Rule Training Condition

Subjects read a four-page description of the concept of sampling and the law of large numbers. This description introduced the important concepts associated with the law of large numbers and illustrated them by using the classic problem of estimating the true proportion of blue and red gumballs in an urn from a sample of the urn. Thus, the gumballs in the urn constituted the *population,* the proportion of blue and red gumballs in the urn formed the *population distribution* (in the example, the population distribution of gumballs was set at 70% blue and 30% red), and a selection of gumballs from the urn constituted a *sample.*

The concept of sampling was then presented by explaining that since it is often impractical or impossible to examine the entire population to determine the population distribution ("Imagine counting a million gumballs!"), it is necessary to rely instead on samples to *estimate* the population distribution. *Sample distributions,* subjects were told, vary in their closeness to the population distribution, and that the only factor determining the closeness of a *random* sample to the population is *sample size.* Finally, the law of large numbers was presented in the following way:

As the size of a random sample increases, the sample distribution is more likely to get closer and closer to the population distribution. In other words, the larger the sample, the better it is as an estimate of the population.

When subjects had finished reading this description, the experimenter performed a live demonstration of the law of large numbers, using a large glass urn filled with blue and red gumballs. In order to maximize subjects' understanding of the concepts they had just read, the demonstration was designed to adhere closely to the description. Each of the concepts introduced in the description was illustrated in the demonstration. For example, the population distribution of the urn was 70% blue and 30% red, just as it had been in the description.

After reintroducing all of the concepts, the experimenter drew four samples of size 1, then four of size 4, and finally, four of size 25. (The gumballs were returned to the urn after each sample.) The experimenter summarized each sample on a blackboard, keeping track of the deviation between each sample and the population. Subjects were told that the average deviation of a sample from the population would decrease as the sample size increased, in accordance with the law of large numbers. Thus, for example, samples of size 25 would, on the average, deviate less from the population than would samples of size 4 or 1. (By good luck, these expected results were obtained in all the training sessions.)

Examples Training Condition

Subjects in the examples training condition read a packet of three example problems with an answer following each problem that provided an analysis of it in terms of the law of large numbers. The three example problems were drawn from Structure 1 (generalizing from a small sample), Structure 3 (regression), and Structure 5 (large sample vs theory without supporting data), and were presented in that order. The three examples were all drawn from the domain of objective problems. After the paragraph that introduced the law of large numbers, there followed a single sentence describing one example of the principle (a public opinion poll based on a large sample is more likely to be accurate than one based on a small sample). The example problems were then introduced in the following way:

> The basic principles involved in the law of large numbers apply whenever you make a generalization or an inference from observing a sample of object, actions, or behaviors. To give you an idea of how broad the law of large numbers is, we have, in this packet, presented three situations in which the law of large numbers applies. Each situation is analyzed in terms of the law of large numbers.

For each example in turn, subjects read the problem and were asked to consider it for a few moments before turning the page to read the law of large numbers answer. The answers to the example problems were constructed so that subjects could learn how the law of large numbers might be applied to a variety of real-life situations. The format of the answers was constant across training domain and structure and included the following characteristics:

1. A statement about the goal of the problem;
2. Identification of the sample or samples and their distributions in the problem;
3. Explanation of how the law of large numbers could be applied to the problem. This identified the population distribution(s) and explained the relationship between the sample(s) and the population(s).
4. The conclusion that could be drawn from the application of the law of large numbers.

The three example problems are presented in Appendix B.

Full Training Condition

Subjects received rule training, followed by examples training, except that the first sentence of the passage introducing the examples was replaced by the following sentence: "One reason that the law of large numbers is important to learn is that it applies *not only* to urns and gumballs."

Demand Condition

Subjects received only the one-sentence definition of the law of large numbers that introduced the examples training, along with the brief example. We included this condition in order to assess whether training effects might be due to experimenter demand or to simply making statistical rules salient to subjects. If performance of the demand group turned out not to be higher than that of the control group, these alternative explanations would be ruled out.

In addition, there was a *control* condition, which received no training before answering the test problems.

In summary, there were five conditions in Experiment 1, as shown in Fig. 1. They were defined by crossing the presence or absence of rule training with presence or absence of examples training. Note that the bottom-left cell of Fig. 1, where neither type of training was given, contains both the control and demand conditions.

Subjects and Procedure

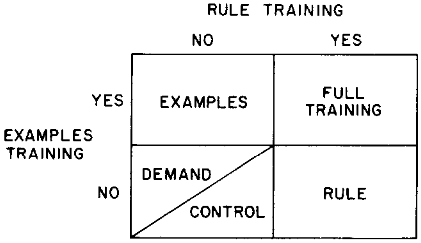
The 347 subjects were adults (229) and high school students (118) from various New Jersey suburban communities. They were paid to participate in the experiment. The adult subjects varied widely in age and education, but almost all were females who were not employed fulltime outside the home. Most of them had participated previously in psychology experiments at Bell Laboratories. Because adults and high school students showed the same pattern of results, their responses were combined in the analyses we present.

Subjects were scheduled in groups of 4–6, with the same training condition presented to the entire group. Training condition was randomly determined. Subjects were told the general nature of the experiment, given the appropriate training, and then given the 18-problem test booklet. They were given 80 min to complete the problems.

Results

Overview of Data Analysis

Recall that subjects' responses were coded using a 3-point system: A code of "1" was given for responses that contained no mention of statis-



		RULE TRAINING	
		NO	YES
EXAMPLES TRAINING	YES	EXAMPLES	FULL TRAINING
	NO	DEMAND CONTROL	RULE

FIG. 1. Design of Experiment 1.

tical concepts such as variability or sample size, whereas a "2" or "3" was given for responses that incorporated statistical notions. Within the class of statistical responses, a "2" was given for "poor" statistical responses, and a "3" was given for "good" statistical responses.

We analyzed the data in terms of two dichotomies. The first one asks whether the response was deterministic (code = 1) or statistical, regardless of quality (code = 2 or 3). We refer to analyses based on this dichotomy as analyses of *frequency* of statistical responses. The second dichotomy asks, for statistical responses only, whether the response was poor (code = 2) or good (code = 3). We refer to analyses based on this dichotomy as analyses of *quality*. The quality dichotomy is conditional: it is defined only for statistical responses and is undefined (missing) for deterministic responses.

These two analyses allowed us to separate the questions of whether training increased the incidence of any kind of statistical reasoning from whether it increased the *proper* use of statistical principles. If we found that training led to an increase in frequency but a decrease in quality, this would lead to the pessimistic conclusion that training merely serves to make statistical concepts salient to subjects without conveying any real sense about how such concepts should be used properly. On the other hand, if training was found to increase *both* frequency and quality, then this would support the optimistic notion that training not only makes salient the usefulness of statistical principles in analyzing inferential problems, but also improves the ability to use those principles correctly.

Because our basic variables were dichotomous, we used a log-linear modeling approach (e.g., Bishop, Fienberg, & Holland, 1975), in which we modeled frequency and quality as a function of (1) training differences, (2) individual differences within training groups, (3) problem differences, and (4) problem × training interaction. This approach closely parallels a three-factor ANOVA model, in which training is a between-subjects variable and problems are crossed with subjects (i.e., problems are treated as repeated measures).

TABLE 1
Frequency and Quality of Statistical Answers in Experiment 1

		Frequency		Quality	
Condition	n	Overall proportion	Log-linear effect	Overall proportion	Log-linear effect
Control	68	.421	−0.515	.542	−0.501
Demand	73	.440	−0.420	.577	−0.316
Rule	69	.557	0.188	.666	0.165
Examples	69	.535	0.074	.659	0.181
Full training	68	.643	0.673	.708	0.471

Effect of Training on Frequency of Statistical Reasoning

Column 3 of Table 1 shows the overall frequency of statistical responses for each of the five experimental groups.[2] It is clear that training increased the frequency of statistical responses, as predicted. Specifically, there resulted a three-level ordering of the conditions. At the lowest level, subjects who received no training (the control and demand conditions) were least likely to employ statistical principles in their answers (42 and 44%, respectively, across all 15 problems). At the middle level, subjects who received only rule training or only examples training were more likely to reason statistically (56 and 54%, respectively). And at the highest level, subjects in the full training condition (those who received both rule and examples training) were most likely to use statistical reasoning in their answers (64%).

The statistical reliability of these proportions cannot be directly assessed from the binomial, since they involve repeated measures over subjects. An alternative strategy would be to employ an analysis of variance on subject means. Such an approach, although quite feasible, would ignore problems as a source of variance, and thus would be inappropriate for our purposes.

Instead, we assessed the reliability of group differences by log-linear analysis. The log-linear effects of training groups, subjects within groups, and problems were all large and highly reliable; the training group × problem interaction was small and only marginally significant.

The simplest way to assess the effects of training is given by the effect sizes for an additive log-linear model based only on training group and

[2] Each of the frequency means represents the proportion of problems for which subjects in that condition utilized some kind of statistical concept. Thus, the frequency mean of .42 for the control condition is based on 1007 responses (68 subjects × 15 problem each, minus 13 unanswered problems).

problems as factors.[3] These effects are shown in Table 1, Column 4. The standard error of each pairwise difference was 0.19, which we obtained from jackknifing.[4] Hence, the difference between the control and the demand conditions and between the rule and examples conditions were not statistically reliable, whereas all of the other pairwise differences were highly reliable ($p < .01$). Thus *both* formal training and training by "guided induction" over examples were effective in increasing the use of statistical heuristics. In addition, training effects were not due to mere experimenter demand or mere salience of statistical rules, since the demand condition was significantly lower than any of the training conditions. In fact, there was no evidence that the demand instructions had any effect whatsoever, compared to controls.

Effect of Training on Quality of Statistical Reasoning

But does training have a beneficial effect on people's ability to use statistical principles *appropriately?* The right-most columns in Table 1 show the overall quality proportions and corresponding effects.[5] The jackknifed estimate of the standard error of the differences in quality between any two conditions was 0.18.

[3] The additive log-linear model can be expressed as: $\log p_{ijk} - \log(1\text{-}p_{ijk}) = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$, where p_{ijk} is the probability of a statistical response by the kth subject in the jth training group for the ith problem. The parameters were estimated by maximizing the likelihood of the 15 × 347 (problem × subject) matrix of zeroes and ones, subject to the identifying constraints that the sum of the problem effects, $\Sigma\alpha_i$, and the sum of the training group effects, $\Sigma\beta_j$, are zero. The estimation was accomplished by the Loglin function of the statistical package S (a product of AT&T Bell Laboratories). The Loglin function uses an algorithm developed by Haberman (1972). The entries in Table 1, column 4, are the estimated values of β_j. The fit was barely improved by including the problem × training interaction parameters, γ_{ij}, to the model. The fit was considerably improved by including subject parameters, δ_{jk}, to the model, but this created difficulties in identifying β_j, because a few of the subject parameter estimates were $+\infty$ or $-\infty$, corresponding to 15 out of 15 or 0 out of 15 statistical answers. Therefore, we stuck with the simple additive model when we tested for differences among training conditions. The $\beta_j s$ from the above model are good descriptive statistics for assessing the effects of training condition, and their sampling properties can be estimated by jackknifing (see Footnote 4).

[4] Jackknifing was performed with 10 subsamples, each formed by randomly dropping 10% of the subjects. The estimated standard error of the pairwise differences (that is, differences between any two $\beta_j s$) varied only slightly from one pair of groups to another.

[5] The quality data were analyzed using the same models as for the frequency data (see Footnote 3). The corresponding parameter estimates, β_j, are shown in the right-most column of Table 1. The 15 × 347 data matrix of zeros and ones for quality had nearly half missing data, since quality was defined only for statistical answers. The nonlinearity of the log-linear model leads to some minor differences between the quality proportions and their corresponding log-linear effects. For example, note that although the rule proportion is greater than the examples proportion, the rule log-linear effect is actually less than the examples log-linear effect.

The effect of training on the quality of statistical responses was strikingly similar to the effects of training on frequency, though somewhat smaller in magnitude. As degree of training increased, the ability to utilize statistical concepts properly increased. This resulted in a similar three-level ordering of the conditions. However, the log-linear analysis indicated that the differences between the full training condition and the rule and examples conditions were significant only at the .10 level.

The effects of training on frequency and quality can be seen clearly in Fig. 2, where the five conditions in Experiment 1 are represented by the filled points. (The open points are from Experiment 2, which are added to demonstrate the stability of training effects across experiments and across different subject populations.) Each training group is represented by one point, with the log-linear frequency effect on the abscissa, and log-linear quality effect on the ordinate. The standard errors of differences for frequency and for quality are shown by a horizontal and vertical bar, respectively.

The diagonal line in Fig. 2 is the least-squares regression line for the five conditions in Experiment 1. It is clear that there is a very stable relationship between the training effect on frequency and on quality, $r(3) = .98$, $p < .005$. The slope of the line is 0.80, which corresponds to the finding that the effect of training on quality was slightly less than the effect on frequency. (Equal effects would be indicated by a slope of 1.00.) This slope is an interesting way to characterize the nature of training procedures. One can imagine procedures that would lead to a much lower

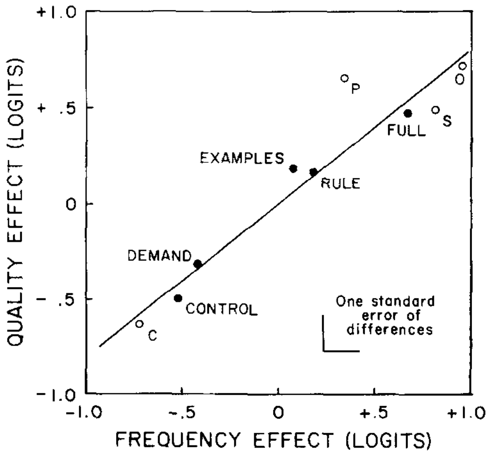


FIG. 2. Effects of training on frequency and quality of statistical answers in Experiment 1 and Experiment 2. Closed points (●) = Experiment 1; open points (○) = Experiment 2: P = probabilistic examples training; O = objective examples training; S = subjective examples training; C = control.

slope (for example, emphasizing the identification of chance processes without much concern for explaining the principles underlying them), or a much higher slope (for example, emphasizing the principles of mathematical statistics, with advice to use great caution in applying such principles broadly).

To summarize, training on the law of large numbers increased the likelihood that people will employ statistical concepts in analyzing everyday inferential problems. Moreover, there appears to be a three-level ordering such that either rule or examples training alone improves performance and that training on both has an additional effect. Training also serves to increase the proper application of statistical concepts in the same way, although this effect is somewhat weaker.

The Effect of Problem Type on the Use of Statistical Principles

Collapsing across training condition, subjects were most likely to employ statistical reasoning for probabilistic problems (75%), less likely to do so for objective problems (48%), and least likely for subjective problems (33%).[6] This result is consistent with the findings of Nisbett et al. (1983) that the use of statistical reasoning is associated with features of the inferential problem that relate to the clarity of the sampling elements and sample space, the salience of the presence of chance factors, and the cultural prescriptions concerning whether causal explanations should include statistical concepts.

Analysis of the quality proportions for the three problem types showed a quite different pattern. There was no significant differences. (The overall proportions for probabilistic, objective, and subjective problems were .63, .53, and .55, respectively.) This suggests that the source of the differences among problem types in statistical reasoning is in the likelihood that a person will notice the relevance of statistical principles to begin with. Given that a person has done so, the three problem types do not differ significantly in whether the person will be able to generate a *good* statistical response.

Thus, frequency of statistical answers was strongly associated with problem type while quality was only weakly associated with problem type. This result is consistent with the notion that people solve problems by use of abstract rules rather than by use of domain-dependent rules: different domains differ with respect to the likelihood that people will recognize the relevance of statistical rules, but once the relevance is rec-

[6] The predicted ordering of the three problem types with respect to frequency of statistical answers (probabilistic > objective > subjective) resulted for each of the five problem structures for which the law of large numbers was relevant (Structures 1–5). The probability of this occurring by chance is extremely low, $p = (1/6)^5 < .001$.

ognized, the same abstract rules are applied across domains with approximately the same degree of success.

Relationship between Training and Problem Type

Are the effects of statistical training limited to the more obvious probabilistic problems, or do they extend to the objective and subjective problems? Figure 3 presents the frequency of probabilistic answers by training condition and problem type. The profiles are nearly parallel, which suggests that there is no interaction between training and problem type.

The log-linear analysis verifies this: Although the interaction between training condition and the 15 problems was significant ($\chi^2(56) = 80, p < .05$), the pattern of residuals from the additive model indicates very clearly that the source of the interaction was due to variation of problems *within* problem type and not at all to systematic differences *between* problem types. Thus, training increased statistical reasoning for subjective events just as much as it did for objective and probabilistic events.

Figure 4 presents the quality of probabilistic answers by training condition and problem type. Note that the three profiles are much closer to each other than are the profiles in Fig. 3: this reflects the fact that frequency was strongly related to problem type, whereas quality was unrelated. We used the same analytic approach to test whether the effect of training on quality of statistical reasoning interacted with problem type. The training × problem interaction was not significant, $\chi^2(54) = 60, p >$

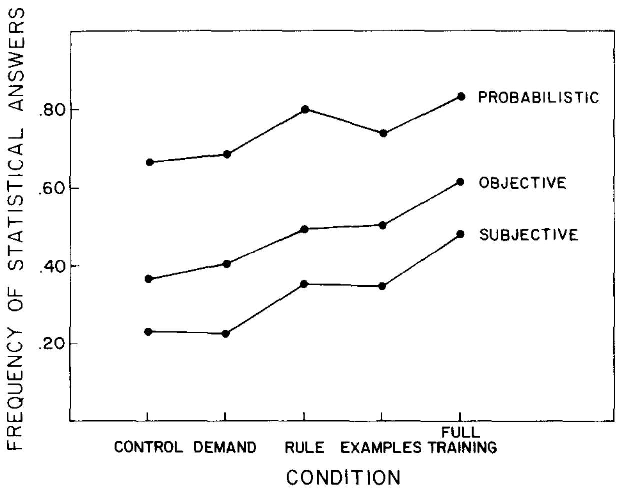


FIG. 3. Frequency of statistical answers as a function of condition and problem type in Experiment 1.

.20. Thus, as with frequency, training effects on quality did not interact with problem type.

These results are consistent with a strong version of the formalist view. Formal rule training improves statistical reasoning and enhances the quality of such reasoning for all kinds of events, not just for probabilistic problems for which there are few plausible alternative kinds of solutions. This finding suggests that operations directly on the abstract rules themselves may be sufficient to produce change in subjects' analysis of essentially the full range of problems they might confront.

These results support the formalist view in a second way. The examples training consisted of example problems only in the domain of objective events. The empirical view predicts domain specificity of training: examples training should lead to greater use of the law of large numbers for the objective test problems but should have less effect for probabilistic and subjective problems. The formalist view, in contrast, predicts domain independence of training. In this view, examples training, insofar as it makes contact with people's relatively abstract rule system of statistical principles, should generalize to other domains as well.

As shown in Figs. 3 and 4, the results are much more consistent with the formalist view. Training on objective example problems improved performance on both probabilistic and subjective problems essentially as much as it improved performance on the objective problems. There was no residual advantage for problems in the domain on which training took place.

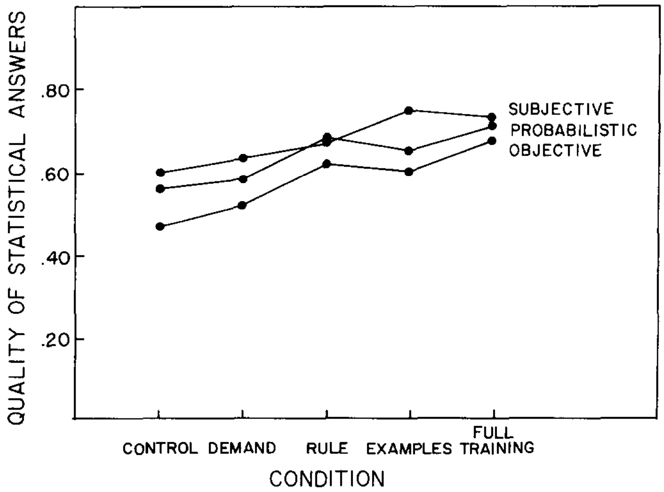


FIG. 4. Quality of statistical answers as a function of condition and problem type in Experiment 1.

False Alarms

Since subjects can only learn so much in a 25-min training session, and since a little learning is a dangerous thing, we should be concerned that our training session may be dangerous in some way. One danger is false alarms, that is, the use of the law of large numbers in situations where it is inappropriate. For example, subjects might claim that the sample size is too small even for problems in which the sample size is quite large. It should be clear that the overuse of the law of large numbers as well as the failure to use it can lead to erroneous conclusions. We explored the possibility that our training may have promoted the indiscriminate use of the law of large numbers by including false-alarm problems in our test package.

There were seven problems for which false-alarm data could be examined. In the three Structure 5 (large sample vs theory without supporting data) problems, the conclusion based on a large sample was contradicted by an opposing argument that was plausible but which was unsupported by data. An answer was given a false-alarm code if it stated that the sample was too small to combat the argument. The three Structure 6 (false alarm) problems involved conclusions drawn from large but biased samples. A false-alarm code was given if a subject accepted the criticism that the sample size was too small. And the objective version of Structure 1 (which we will refer to as O1) asked subjects to comment on two conclusions—one based on a large sample (part a), and one based on a very small sample (part b). Part a was used to assess subjects' tendency to false alarm; part b was used to assess the subjects' ability to use the law of large numbers correctly.

Of the seven false-alarm problems, three of them (O5, S5, and S6) elicited virtually no false alarms (less than 2%). For a fourth problem (P6), the false-alarm rate was about 10%, with the false alarms distributed approximately equally among the five conditions. The results for these four problems suggest that trained subjects do indeed increase their use of the law of large numbers in a discriminating fashion.

For the other three problems, the false-alarm rates were only somewhat higher for the three trained groups (about 16%) than for the two untrained groups (about 10%). And it is interesting that the specific pattern of false alarms across the three trained groups varied depending on whether subjects had received examples training. In P5 (the probabilistic version of Structure 5), for instance, subjects exposed to examples training (the examples and full training conditions) were less likely to false alarm than those exposed to rule training only. This is probably because the examples training package included a Structure 5 problem. These subjects had thus been alerted to the possibility that large samples were indeed "large enough" to make confident conclusions and were therefore less likely to false alarm on P5. In contrast, subjects receiving only rule training were not given any information about when a sample was large enough. It is not surprising, then, that these subjects were more likely to false alarm to this problem.

There is also evidence from problem O1 that the tendency to false alarm was negatively related to the proper use of the law of large numbers. For this problem, there was a strong negative relationship between false alarms to part a and the quality of statistical responses to part b. Of the subjects who false alarmed on part a, none gave a good statistical answer to part b, that is, quality was equal to .00. In contrast, for those subjects who had not false alarmed, quality was equal to .16. This analysis suggests that a little learning can be somewhat dangerous, but that subjects who absorb the training more thoroughly are able to use it in a discriminating fashion.

In summary, our 25 min training session did *not* lead to widespread overuse of the law of large numbers.[7] Instead, subjects were surprisingly sophisticated in avoiding the improper use of the law of large numbers, sometimes citing intuitive versions of statistical concepts such as power and confidence intervals in their answers. Moreover, subjects who did false alarm were also less likely to use the law of large numbers correctly when it was appropriate.

EXPERIMENT 2

The results of Experiment 1 indicate very clearly that people can be taught to reason more statistically about everyday inferential problems. They can be taught through example problems showing how statistical principles can be applied, and they can also be taught through illustrating the formal aspects of the law of large numbers. These results are consistent with the formalist view that people possess abstract inferential rules and that these can be improved both by guided induction through examples and by direct manipulation.

One of the important results in Experiment 1 was the absence of an interaction between training and problem type. Examples training had an equal effect in enhancing statistical reasoning across all three problem types. Thus, training on objective problems increased the use of statistical thinking no more for objective events than for subjective events, such as choosing a college or explaining a person's compassionateness, or for probabilistic events, such as those involving lotteries or slot machines. That training effects were entirely domain independent is quite

[7] Complete details of the false-alarm analyses for Experiments 1 and 2 can be obtained from the authors.

remarkable when contrasted with the strong domain specificity of subjects' spontaneous *use* of statistical reasoning. Subjects were much more likely to use statistical principles for probabilistic problems than for objective problems and much more likely to use them for objective problems than for subjective problems.

Experiment 2 was designed to explore more fully whether training effects might vary as a function of the training domain. In Experiment 1, all subjects who received examples training were given example problems only in the objective domain. In Experiment 2, subjects were taught how to apply the law of large numbers in one of the three problem domains: probabilistic, objective, or subjective. All subjects were then tested on all three problem domains. This design makes it possible to see whether there are domain-specific effects of training. The empirical view suggests that subjects would be expected to show more improvement for problems in the domain in which they were trained than for other problems. The formal view, on the other hand, predicts that there will be no such interaction between training domain and testing domain.

Method

Subjects

The subjects were 166 undergraduates at the University of Michigan who were enrolled in introductory psychology classes. They participated in the 2-h experiment in small groups.

Design and Procedure

Subjects were randomly assigned to one of four conditions. The *control* condition was identical to that in Experiment 1. In the other three conditions, subjects were given training identical to the full training condition in Experiment 1, except that the type of example problems varied. Subjects in the *probabilistic training* condition read three probabilistic example problems and were shown how each could be analyzed by the application of the law of large numbers. Subjects in the *objective training* condition were given the same three objective example problems that were used in Experiment 1. And subjects in the *subjective training* condition were given three subjective example problems. The probabilistic and subjective examples matched the objective examples in structure: they were drawn from Structures 1, 3, and 5.

All subjects then answered the same set of 18 test problems (15 law of large numbers problems and 3 false-alarm problems) used in Experiment 1.

The subjects' responses to the open-ended questions were coded by two raters under the same coding system used in Experiment 1. The reliability of the coding was high—there were exact matches by the two coders on 88% of the responses.

Results

The data analytic procedures we used in Experiment 1 were employed here. From the 3-point coding system, we derived frequency and quality dichotomies and then used log-linear models to estimate the effects of training, test problem, and training × test problem interaction. The jackknifed estimate of the standard error of the difference between any two conditions for frequency and quality were 0.20 and 0.18 on the log-linear scale, respectively. These standard errors correspond very closely to those found in Experiment 1.

Effects of Training

As in Experiment 1, training significantly enhanced the frequency of statistical responses. Subjects in the control conditions were least likely to use statistical concepts for the 15 test problems (53% of responses were statistical). The three training groups were significantly more likely than controls to give statistical answers (72, 81, and 79% for the probabilistic, objective, and subjective training groups, respectively. All comparisons with the control condition were significant at the .001 level). In addition, subjects trained on probabilistic examples were less likely than subjects trained on objective or subjective examples to reason statistically ($p <$ .01 and .05, respectively); the objective and subjective example conditions did not differ from each other.

Training also increased the quality of statistical answers. The quality proportions were .47 for the control group and .70, .70, and .66 for the probabilistic, objective, and subjective groups, respectively. Once again, training significantly enhanced the quality of statistical responses (all comparisons with the control condition were significant at the .001 level). But, in contrast to the frequency data, no training domain was more effective than any other in enhancing the quality of statistical answers.

The relationship between the training effects on frequency and on quality was very consistent with Experiment 1, as can be seen by looking back to Fig. 2, where the open points represent the frequency and quality effects of the three training conditions and the control condition for Experiment 2.

Effect of Problem Type

The strong effect of problem type found in Experiment 1 was replicated here. Collapsing across conditions, subjects were most likely to reason statistically for probabilistic problems (91%), less likely to do so for objective problems (68%), and least likely for subjective problems (56%).[8]

As in Experiment 1, the quality of statistical answers varied only

[8] Although the pattern of these proportions are similar to those in Experiment 1, their magnitude is substantially greater. One reason is that whereas the five conditions in Experiment 1 varied considerably in the degree of training, three of the four conditions in Experiment 2 were essentially full training conditions (all were given rule training). When averaging across conditions, the proportions for Experiment 2 will reflect this more extensive training.

slightly across the three problem types. The quality proportions were .69, .65, and .60 for the probabilistic, objective, and subjective problems, respectively. These differences were not statistically significant.

Relationship between Training Domain and Test Domain

The primary goal of this experiment was to examine the relationship between training domain and test domain. Figures 5 and 6 present the frequency and quality of statistical answers as a function of training domain and test domain. If training effects were domain specific, we should find that frequency and quality for problems in a given domain will be highest for those subjects who were trained on that domain. These domain-specificity data points are represented as larger data points in the two figures. Figures 5 and 6 make it clear that this was not the case: the domain-specific data points are not consistently higher than the other data points. For example, subjects who were trained on problems in the probabilistic domain were actually *less* likely to think statistically on the probabilistic test problems than were subjects trained on objective or subjective problems. In short, training significantly increased statistical reasoning; the domain of training had no differential effect.

The log-linear analysis confirms the absence of domain specificity of training. There was no significant interaction between training domain and test domain, either for frequency, $\chi^2(42) = 55, p = .10$, or for quality, $\chi^2(42) = 49, p > .15$.

Finally, the false-alarm rates for Experiment 2 were generally higher than they were for Experiment 1, for the control group as well as for the

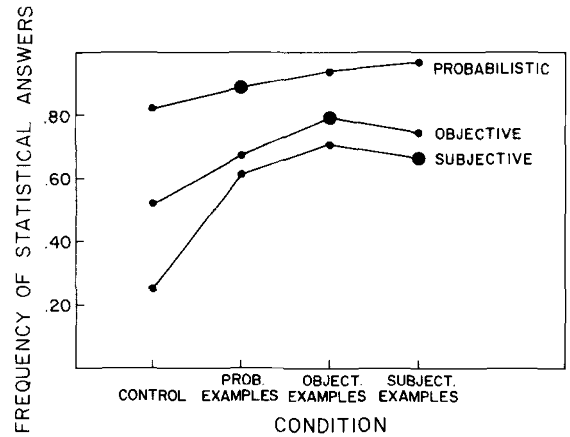


FIG. 5. Frequency of statistical answers as a function of condition and problem type in Experiment 2.

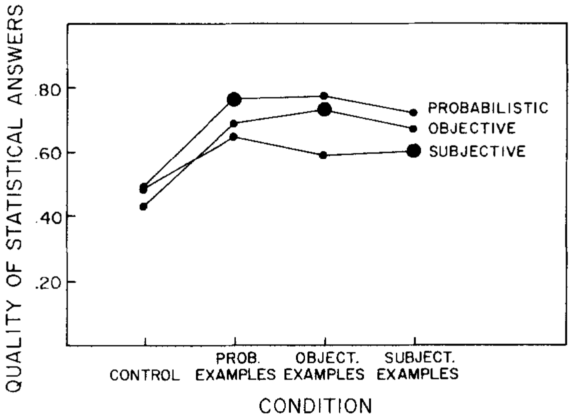


FIG. 6. Quality of statistical answers as a function of condition and problem type in Experiment 2.

trained groups. The difference may be due to the fact that the subjects in Experiment 2 were college students, but this is only speculation.

Discussion

The results of Experiments 1 and 2 show that instruction in statistics can have a marked effect on the way people reason about a broad range of everyday problems. Such training affects not only their reasoning about transparently probabilistic events such as lotteries, but also their reasoning about events that most people analyze using only deterministic rules.

Both formal training, restricted to descriptions of the formal aspects of the law of large numbers, and "guided induction," that is, teaching the rule by means of examples, were effective in improving both the frequency and the quality of statistical reasoning. The former finding suggests that the more abstract aspects of academic training in statistics may, by themselves, be sufficient to produce significant improvement in the way people reason. We test this hypothesis in Experiments 3 and 4. The latter finding indicates that the use of examples adds greatly to people's ability to use their abstract rule systems.

The two types of training were approximately additive on the log-linear scale, that is, examples training plus rule training added as much improvement, both in frequency and quality, as would be expected from the sum of the effects of each type of training in isolation. It is important to note that, in the present experiments at least, the effect of examples training does not appear to be in the form of rules about how to "map" the law of large numbers onto the content of particular domains. This is because there was no domain specificity of training effects. In general, subjects taught examples in one domain learned no more about how to solve problems in that domain than they did about how to solve problems in other domains. There are two hypotheses that may account for this domain independence of examples training. What subjects learn from examples training may be an abstracted version of the law of large numbers. Alternatively, or perhaps in addition, they may learn an abstracted version of how to apply the principle to problems in general.

The domain independence of training effects we found should not be presumed to be highly general, however. Every teacher knows that students sometimes apply a rule beautifully in a domain in which they have been taught the rule and yet fail to apply it in another domain in which it is just as applicable. Two aspects of the present work probably contributed to the domain independence of statistical training that we found. First, the domains we used were very broad, constituting three haphazard samples of problems, one sample united only by the fact that some obvious randomizing device was present, another consisting of problems where a protagonist had to make a judgment about some objectively measurable aspect of a person or object, and another consisting of problems where a protagonist had to make a judgment about some subjective aspect of a person or object. Had we studied substantially narrower domains—the domain of sports, for example, or the domain of judgments about personality traits—and had we taught subjects specific tools for coding events in those domains and for thinking about their variability, we might well have found some domain specificity of training effects.

A second factor that almost surely contributed to the lack of domain specificity of training effects was the fact that testing immediately followed training. Thus subjects could be expected to have their newly improved statistical rules in "active memory" at the time they were asked to solve the new problems. This fact could be expected to reduce domain-specificity effects to a minimum.

It may have occurred to the reader to suspect that the temporal relation between testing and training might not only reduce domain-specificity effects of training but might be essential in order to produce any effects of training at all. In fact, it could be argued that all our "training" did was to increase the salience of subjects' statistical heuristics and did not teach them anything new at all. As we have known since Socrates' demonstration with the slave boy, it is always hard to prove whether we have taught someone something they did not know before or whether we have merely reminded them of something they already knew.

We have two main lines of defense, however, against the suggestion that our training effects in Experiments 1 and 2 were due simply to making the law of large numbers more salient to subjects. First, *reminding* subjects about the law of large numbers and encouraging them to use it had no effect either on the frequency or the quality of their answers. This is shown clearly by the fact that subjects in the demand condition were no higher than subjects in the control condition on either measure. Second, our training manipulations improved not only the frequency of statistical answers, which would be expected on the basis of a mere increase in salience, but the *quality* of answers, which would not be expected on the basis of a mere increase in salience.

The most effective response to the artifactual possibility of salience, however, would be to separate the time and context of training from the time and context of testing. We did this in two different experiments. In Experiment 3, we examined the effect of differing amounts of formal course training in statistics on subjects' tendencies to give statistical answers to problems. In Experiment 4, we examined the effect of course training in statistics, and we also disguised the context of testing as an opinion survey. In addition to helping rule out the salience and testing context alternatives, these experiments speak to practical questions about the effects of statistical training in formal courses on everyday inferential problems.

EXPERIMENT 3

In Experiment 3 we examined the effect of varying amounts of formal course training on the way people reasoned about two different versions of a problem from everyday life. The two versions were very similar, except that one had a powerful probabilistic cue. The study thus allows a comparison of the effects of training on both the likelihood of using statistical reasoning and the quality of statistical reasoning for both a problem for which statistical reasoning is relatively common and a problem for which it is relatively rare.

Subjects and Method

Four groups of subjects participated. These groups were chosen for their background, or lack of background, in formal statistical training. The *no statistics* group were 42 college undergraduates who were attending a lecture on attitudes; none had taken college level statistics. The *statistics* group were 56 students attending the same lecture who had taken an introductory statistics course. The *graduate* group were 72 graduate students in psychology, who were attending the first session of a course on statistical methods; all had taken at least one statistics course, and many had taken more than one. And the *tech* group were 33 technical staff members at a research laboratory who were attending a colloquium on probabilistic reasoning. Nearly all were Ph.D. level scientists who had taken many statistics courses.

Subjects were presented with a problem about restaurant quality. There were two versions. In the *no randomness cue* version, a traveling businesswoman often returns to restaurants where she had an excellent meal on her first visit. However, she is usually disappointed because subsequent meals are rarely as good as the first. Subjects were asked to explain, in writing, why this happened.

The *randomness cue* version included a random mechanism for selection from the menu. In this version, the protagonist was a businessman in Japan who did not know how to read the language. When eating at a restuarant, he selected a meal by blindly dropping a pencil on the totally unreadable menu and ordering the dish closest to it. As in the other version, he is usually disappointed with his subsequent meals at restaurants he originally thought were superb. Why is this?

Answers were classified as "statistical" if they suggested that meal quality on any single visit might not be a reliable indicator of the restaurant's overall quality (e.g., "Very few restaurants have only excellent meals; odds are she was just lucky the first time"). "Non-statistical" answers assumed that the initial good experience was a reliable indicator that the restaurant was truly outstanding, and attributed the later disappointment to a definite cause such as a permanent or temporary change in the restaurant (e.g., "Maybe the chef quit") or a change in the protagonist's expectation or mood (e.g., "Maybe her expectations were so high on the basis of her first visit that subsequent meals could never match them"). Explanations that were statistical were coded as to whether they merely referred vaguely to chance factors ("poor statistical") or whether they also articulated the notion that a single visit may be regarded as a small sample, and hence as unreliable ("good statistical"). Thus, the coding system was essentially the same as the one used in Experiments 1 and 2.

Results

Figure 7 shows the frequency and quality of answers as a function of training and type of problem. The left side of Fig. 7 demonstrates clearly that the frequency of statistical answers increased dramatically with level of statistical training, $\chi^2(6) = 35.5, p < .001$. Almost none of the college students without statistical training gave a statistical answer to the version without the randomness cue, whereas 80% of Ph.D. level scientists did so.

Inclusion of the randomness cue markedly increased the frequency of statistical answers, $\chi^2(4) = 27.1, p < .001$. For the untrained college students, for example, the presence of the randomness cue increased frequency from 5 to 50%. The randomness cue thus apparently encourages

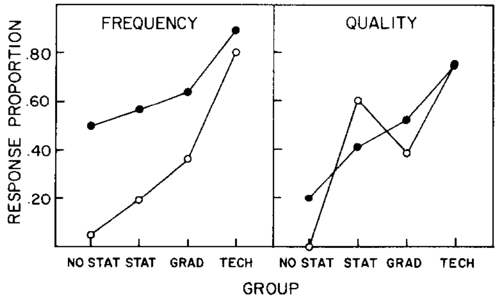


FIG. 7. Frequency and quality of statistical answers as a function of group and problem version in Experiment 3. Closed points (●) represent responses to the Randomness Cue version. Open points (○) represent responses to the No Randomness Cue version.

the subject to code restaurant experiences as units that can be sampled from a population.

The right side of the figure indicates that degree of statistical training was also associated with *quality* of statistical answers, $\chi^2(3) = 12.3, p < .001$. Only 10% of the statistical answers by untrained college students were rated as good, whereas almost 80% of the statistical answers by Ph.D level scientists were rated as good.

Although the presence of the randomness cue was very important in determining whether subjects would think statistically at all, it did not affect the *quality* of statistical answers for subjects at any level of training. This duplicates the findings of Experiments 1 and 2, showing that problem difficulty does not affect the quality of answers, given that the answers are statistical. Apparently cues about randomness can trigger the use of statistical rules, but they do not necessarily produce good statistical answers. Such cues can only trigger rules at whatever level of sophistication the subject happens to possess them. This correlational study thus buttresses our assertion that, whatever plausibility the salience alternative has for the frequency results, it has very little plausibility for the quality results.

Discussion

These data indicate that, when one examines people who represent a broad range of statistical expertise, one can find very marked differences in the tendency to approach certain kinds of problems statistically. The data also indicate that, even when statistical approaches are preferred by untutored subjects, as for the version of the problem having the randomness cue, the quality of answers given by such subjects will be markedly inferior to that which more expert subjects can give.

But, while suggestive, the data do not show to precisely what degree formal rule training per se is effective. First, statistical training was undoubtedly confounded with intellectual ability, and perhaps even with experiences in superb restaurants. Second, more extensive training in statistics is normally associated with more extensive training in content disciplines that teach the use of statistical and methodological rules in at least an informal way, across a variety of domains. Thus statistical training is also confounded with other types of potentially relevant training.

In Experiment 4, we removed these sources of confounding and also provided a testing context that would not be expected to cue subjects into using statistical rules. We conducted Experiment 4 in order to examine the effects of formal statistical training in a setting completely outside of the context of training. Students enrolled in an introductory statistics course were contacted at home and were asked to participate in a telephone survey on "students' opinions on sports." Some of the questions could be analyzed with reference to statistical concepts such as the law of large numbers and the regression principle. None of the students was aware that the survey was related to the statistics class they were enrolled in. If training has an effect in this situation, this would provide very strong evidence for the formal view that statistical heuristics are represented at a highly abstract level and that statistical training provides inferential tools that are quite domain and context independent.

EXPERIMENT 4

Subjects

The subjects were 193 randomly selected males at the University of Michigan who were enrolled in an introductory statistics course. The course had a total enrollment of over 600 students.

Method

We obtained the class list from the instructor and randomly selected half of the males to be contacted during the first week of the semester, and the other half to be contacted during the last week of the semester.

The protocol we used was designed to convince subjects that we were conducting a genuine opinion survey. The interviewer introduced herself in the following way:

> I am calling from the Research Center for Group Dynamics at the University of Michigan. We're conducting a campus survey about students' opinions on sports. Some of the questions in this survey ask for opinions on current events in professional and collegiate sports; other questions ask for general opinions about sports. The whole survey takes only about 10 to 15 min. Would you have time now to answer our questions?

After asking for some demographic information[9], the interviewer went on to the questions. To enhance the idea that this was a legitimate opinion survey, the first two questions indeed asked subjects to give their real opinions about certain sports controversies (e.g., what colleges should do about recruiting violations). Respondents were quite unaware that the survey was really designed to test their statistical knowledge—none voiced any suspicion.

Following the filler items, subjects were asked a series of questions that could be answered with reference to statistical concepts. This was the first such question:

> In general, the major league baseball player who wins Rookie of the Year does not perform as well in his second year. This is clear in major league baseball in the past 10 years. In the American League, eight Rookies of the Year have done worse in their second year; only two have done better. In the National League, the Rookie of the Year has done worse the second year 9 times out of 10. Why do you suppose the Rookie of the Year tends not to do as well his second year?

[9] In order to ensure that subjects had enough knowledge of sports to be able to understand the survey questions, they were asked to rate their knowledge about sports. Those who rated themselves as having little or no knowledge of sports were not used in this experiment.

Responses to this regression question were tape-recorded and coded for the presence of statistical reasoning and for whether a statistical response was a good one. A typical nonstatistical response for this question would be, "The Rookie of the Year doesn't do as well because he's resting on his laurels; he's not trying as hard in his second year." A good statistical response would be, "A player's performance varies from year to year. Sometimes you have good years and sometimes you have bad years. The player who won the Rookie-of-the-Year award had an exceptional year. He'll probably do better than average in his second year, but not as well as he did when he was a rookie."

Results and Discussion

Results indicated that training in a standard statistics course had a significant effect in enhancing the use of statistical explanations for this question. For those contacted at the beginning of the term, 16% gave statistical answers. For those contacted at the end of the term, over twice as many (37%) gave answers that utilized statistical thinking. This increase in frequency was significant, $z = 3.23$, $p < .005$. In addition, the statistics course also enhanced the quality of statistical responses, from .12 to .38, though this was only marginally significant, $z = 1.77$, $p < .10$.

Similar results were obtained on another problem, which asked subjects to explain why the top batting average after 2 weeks of the season is around .450, when such a high average has never been obtained over an entire season. Frequency increased from .50 to .70, $z = 2.87$, $p < .01$, and quality increased from .24 to .50, $z = 2.74$, $p < .01$.

The statistics course did not have any effect on two other problems that we included in the sports survey. One problem asked whether a more talented squash player should choose a five-point or a one-point tie breaker. The other asked subjects to critique a large sample study about whether marriage has an adverse effect on a professional athlete's performance. We have no explanation for why a statistics course failed to enhance statistical reasoning for these two problems.

This study indicates clearly that statistical training can enhance the use of statistical rules in reasoning about everyday life and can do so completely outside the context of training.

GENERAL DISCUSSION

The experiments presented here demonstrate that statistical training serves to enhance the use of statistical principles in reasoning. The effects of training are impressive in their generality across method, context, type of subject, and event domain. Statistical training conferred benefits whether the training consisted of several statistics courses, a single semester-long course, or even a 25-min training session. Training effects occurred not only when the testing context was identical to the training context, but also when the testing context was completely different from the training context in time and situation. Training enhanced statistical thinking not only for college students enrolled in introductory psychology, but also for high school students and adults. Training enhanced both the frequency and quality of statistical thinking not only for events commonly associated with uncertainty and probability, but also, to the same extent, for events rarely associated with such concepts.

A qualification that must be placed on the present results is that the effects at least of relatively brief training sessions may be limited to problems for which some untrained subjects are able to give a statistical answer. Many previous demonstrations of people's difficulties with statistical principles are based on problems to which no subjects, or almost no subjects, apply statistical reasoning (e.g., Hamill, Wilson, & Nisbett, 1980; Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1983). Quite deliberately, we avoided such difficult problems in the present investigations. Even for the subjective problems in Experiment 1 and 2, the average rate of statistical answers for untrained subjects was slightly in excess of 20%.

It is indeed striking that statistical training enhances statistical thinking for subjective judgments, such as those made about the social world. Social judgments such as attributions of success or failure, or judgments of a person's traits based on a first impression, are those that, by their very nature, have a critical impact on our lives. At the same time, social judgments are those for which unexplained variation plays a major role.

But because social events are difficult to code, and because the sample space for such events is typically difficult to define, social judgments are also those that are least likely to be made with reference to statistical considerations. This is shown by the domain-specificity effects found in the experiments presented here and by Jepson et al., (1983) on the use of statistical thinking for probabilistic, objective, and subjective events. It is a disturbing state of affairs that the domain where statistical thinking is most necessary on an everyday basis is the one where it is least likely.

Our training studies, however, suggest that people are able to understand and accept the applicability of statistical principles for social events as well as for nonsocial events. The lack of an interaction between training and problem domain indicates that statistical training enhances statistical thinking for social events just as much as it does for nonsocial events. This domain independence of statistical training makes us optimistic that people can indeed be taught to understand the role of uncertainty and sample size in making social judgments.

More generally, the studies reported here make an important point concerning pedagogy in statistics. In the early 1800s, Laplace wrote, "the theory of probabilities is at bottom nothing but common sense reduced to calculus." It seems to us that courses in statistics and probability theory today concentrate almost entirely on the calculus, while often ignoring its commonsense roots. Experiments 3 and 4 clearly demonstrate how classroom training in statistics can potentially have a significant effect on how people make judgments. If introductory statistics courses were to incorporate examples of how statistical principles such as the law of large numbers can be applied to judgments in everyday life, we have no doubt that such courses would have a more far-reaching effect on the extent to which people think statistically about the world.

These studies suggest very strongly that people make use of abstract inferential rules in the form of statistical heuristics. We also know this because training on the purely formal aspects of the law of large numbers improves statistical thinking over a broad range of content, and because showing subjects how to use the rule in a given content domain generalizes completely to quite different content domains. We are aware of no more convincing evidence, in fact, for the existence of abstract rules of reasoning than the present work.

What is the origin of abstract inferential rules about the law of large numbers? Why do people develop such high-level representations of the law of large numbers? We suspect the answer comes, in large measure, from the ubiquity of the principle. The basic notion that large samples are more reliable than small samples underlies concept formation and generalization. It can be argued that during cognitive development, the child learns, through repeated exposure to the law of large numbers across many domains, a highly abstract representation of the principle.

The experiments presented here demonstrate the dual usefulness of inferential training studies. Such studies are important for pragmatic reasons because they provide information about how everyday reasoning might be improved. It is heartening to discover that a 25-min session on the law of large numbers can serve to significantly enhance people's use of statistical thinking, and that a formal course in introductory statistics can lead to a greater appreciation of variability in judgments, even those made outside the context of the classroom or laboratory. In addition, such studies are important for theoretical reasons not only because of what they tell us about how inferential rules are utilized, but also about how they are represented and how they can be modified.

APPENDIX A

The Eighteen Test Problems Used in Experiments 1 and 2

Probabilistic—Structure 1

At Stanbrook University, the Housing Office determines which of the 10,000 students enrolled will be allowed to live on campus the following year. At Stanbrook, the dormitory facilities are excellent, so there is always great demand for on-campus housing. Unfortunately, there are only enough on-campus spaces for 5000 students. The Housing Office determines who will get to live on campus by having a Housing Draw every year: every student picks a number out of a box over a 3-day period. These numbers range from 1 to 10,000. If the number is 5000 or under, the student gets to live on campus. If the number is over 5000, the student will not be able to live on campus.

On the first day of the draw, Joe talks to five people who have picked a number. Of these, four people got low numbers. Because of this, Joe suspects that the numbers in the box were not properly mixed, and that the early numbers are more favorable. He rushes over to the Housing Draw and picks a number. He gets a low number. He later talks to four people who drew their numbers on the second or third day of the draw. Three got high numbers. Joe says to himself, "I'm glad that I picked when I did, because it looks like I was right that the numbers were not properly mixed."

What do you think of Joe's reasoning? Explain.

Probabilistic—Structure 2

For his vacation, Keith decided to drive from his home in Michigan to California to visit some of his relatives and friends. Shortly after crossing the border into Nevada, Keith pulled into a gas station and went inside to buy a state map. There, in a corner of the gas station, were two slot machines. Keith had heard about slot machines before, but had never actually seen one. He went over to the slot machines and looked at them, trying to figure out how they worked. An old man who was sitting close to the machines spoke to Keith. "There ain't no winning system for slot machines. It's all luck. You just put in a coin, pull the lever, and hope that you'll win. But let me tell you this: some machines are easier to lose on than others. That's because the owners can change the mechanism of the slots so that some of them will be more likely to make you lose. See those two slot machines there? The one on the left gives you about an even chance of winning, but the one on the right is fixed so that you'll lose much more often than you'll win. Take it from me—I've played them for years." The old man then got up and walked out of the gas station.

Keith was by now very intrigued by the two slot machines, so he played the machine on the left for a couple of minutes. He lost almost twice as often as he won. "Humph," Keith said to himself. "The man said that there was an even chance of winning at that machine on the left. He's obviously wrong." Keith then tried the machine on the right for a couple of minutes and ended up winning more often than he lost. Keith concluded that the man was wrong about the chances of winning on the two slot machines. He concluded that the opposite was true—that the slot machine on the right was more favorable to the player than the machine on the left.

Comment on Keith's conclusion and his reasoning. Do you agree? Explain your answer.

Probabilistic—Structure 3

Bert H. has a job checking the results of an X-ray scanner of pipeline welds in a pipe factory. Overall, the X-ray scanner shows that the welding machine makes a perfect weld about 80% of the time. Of 900 welds each day, usually about 680 to 740 welds are perfect. Bert has noticed that on some days, all of the first 10 welds were perfect. However, Bert has also noticed that on such days, the overall number of perfect welds is usually not much better for the day as a whole than on days when the first 10 welds show some imperfections.

Why do you suppose the number of perfect welds is usually not much better on days where the first batch of welds was perfect than on other days?

Probabilistic—Structure 4

Joanna has a large collection of pennies with dates in the 1970s. Donny admires her collection and decides to start his own collection of pennies, but decides to collect only 1976 pennies because he wants to commemorate the Bicentennial. Looking through his pockets, he discovers he has only a dime. Examining it carefully, he finds that it is a 1971 dime, with a "D" (Denver) mint mark. Donny thinks it would be fun to collect 1976 pennies with the same initial as his name and asks Joanna what proportion of the 1976 pennies in her collection have a "D" mint mark on them. She doesn't know, but they decide to find out. They take the huge jar of her pennies out. Since the jar has thousands of pennies in it, Donny shakes the jar and then reaches into it and picks out a handful from the middle of the jar. Donny finds all the 1976 pennies that he scooped out (four of them) and finds that two of them have "D" mint marks. Because of this, he estimates that around 50% of all Joanna's 1976 pennies have the "D" mint mark. But Joanna looks through the other 36 pennies they have scooped out (dated 1970–1975 and 1977–1979) and discovers that only 2 of them have the "D" mint mark. She argues that only 4 of 40 pennies altogether have the "D" mark, and estimates that around 10% of the 1976 pennies in her collection are "D" pennies.

Comment on the validity of Joanna's and Donny's reasoning. Whose conclusion about the 1976 pennies in Joanna's collection is more likely to be correct? Explain.

Probabilistic—Structure 5

An auditor for the Internal Revenue Service wants to study the nature of arithmetic errors on income tax returns. She selects 4000 Social Security numbers by using random digits generated by an "Electronic Mastermind" calculator. And for each selected social security number she checks the 1978 Federal Income Tax return thoroughly for arithmetic errors. She finds errors on a large percentage of the tax returns, often 2 to 6 errors on a single tax return. Tabulating the effect of each error separately, she finds that there are virtually the same number of errors in favor of the taxpayer as in favor of the government. Her boss objects vigorously to her assertions, saying that it is fairly obvious that people will notice and correct errors in favor of the government, but will "overlook" errors in their own favor. Even if her figures are correct, he says, looking at a lot more returns will bear out his point.

Comment on the auditor's reasoning and her boss's contrary stand.

Probabilistic—Structure 6

A brewery buys nearly all of its reusable glass bottles from a local glass manufacturer. One summer, however, the local company is unable to deliver enough bottles, and the brewery orders a shipment from a large glass manufacturer that distributes its products nationwide. On the first day that these new bottles are used, however, the bottle-filling machinery has to be stopped four times because of jamming, and, as a result, production for the day is unusually low. (Ordinarily the brewery does not experience more than one jamming stoppage per day and frequently there are none at all.) The foreman is worried about the new bottles. He decides to test the new bottles produced by the national manufacturer carefully. He randomly selects 300 cases of these new bottles and instructs the bottle-filler operators to record carefully each jamming incident. Meanwhile, company mechanics carefully lubricate and check adjustments on the bottle-filling machinery. When they are finished, the bottle-filling machinery is running more smoothly than it has for years. During the next 2 days, the 300 cases of new bottles are fed to the machine. There are only two jamming incidents, one each day. The foreman concludes that there is in fact little or no real disadvantage of the new bottles with respect to jamming of the bottle-filling machinery.

Comment on the foreman's reasoning. Is it basically sound? Can his procedure be criticized?

Objective—Structure 1

A talent scout for a professional basketball team attends two college games with the intention of observing carefully the talent and skill of a particular player. The player looks generally excellent. He repeatedly makes plays worthy of the best professional players. However, in one of the games, with his team behind by 2 points, the player is fouled while shooting and has the opportunity to tie the game by making both free throws. The player misses both free throws and then tries too hard for the rebound from the second one, committing a foul in the process. The other team then makes two free throws, for a 4-point lead, and goes on to win by 2 points.

The scout reports that the player in question "has excellent skills, and should be recruited. He has a tendency to misplay under extreme pressure, but this will probably disappear with more experience and better coaching."

Comment on the thinking embodied in the scout's opinion that the player (a) "has excellent skills" and that the player has (b) "a tendency to misplay under extreme pressure." Does the thinking behind either conclusion have any weaknesses?

Objective—Structure 2

The Caldwells had long ago decided that when it was time to replace their car they would get what they called "one of those solid, safety-conscious, built-to-last Swedish cars"—either a Volvo or a Saab. As luck would have it, their old car gave up the ghost on the last day of the closeout sale for the model year both for the Volvo and for the Saab. The model year was changing for both cars and the dollar had recently dropped substantially against European currencies; therefore, if they waited to buy either a Volvo or a Saab, it would cost them substantially more—about $1200. They quickly got out their *Consumer Reports* where they found that the consensus of the experts was that both cars were very sound mechanically, although the Volvo was felt to be slightly superior on some dimensions. They also found that the readers of *Consumer Reports* who owned a Volvo reported having somewhat fewer mechanical problems than owners of Saabs. They were about to go and strike a bargain with the Volvo dealer when Mr. Caldwell remembered that they had two friends who owned a Saab and one who owned a Volvo. Mr. Caldwell called up the friends. Both Saab owners reported having had a few mechanical problems but nothing major. The Volvo owner exploded when asked how he liked his car. "First that fancy fuel injection computer thing went out: $250 bucks. Next I started having trouble with the rear end. Had to replace it. Then the transmission and the clutch. I finally sold it after 3 years for junk."

Given that the Caldwells are going to buy either a Volvo or a Saab today, in order to save $1200, which do you think they should buy? Why?

Objective—Structure 3

Howard was a teacher in a junior high school in a community known for truancy and delinquency problems among its youth. Howard says of his experiences: "Usually, in a class of 35 or so kids, 2 or 3 will pull some pretty bad stunts in the first week—they'll skip a day of class, get into a scuffle with another kid, or some such thing. When that kind of thing happens, I play it down and try to avoid calling the class' attention to it. Usually, these kids turn out to be no worse than the others. By the end of the term you'll find they haven't pulled any more stunts than the others have." Howard reasons as follows: "Some of these kids are headed toward a delinquent pattern of behavior. When they find out nobody is very impressed, they tend to settle down.''

Comment on Howard's reasoning:

(a) Do you agree that it is likely that the students who pull a ''pretty bad stunt in the first week'' are ''headed toward a delinquent pattern of behavior?''

(b) Do you agree that it is likely that the students who initially pull a ''pretty bad stunt'' turn out to be no worse than the others because they find no one is impressed with their behavior?

Objective—Structure 4

The psychology department of the University of Michigan keeps records on the performance of all its graduate students and relates this performance score to all kinds of background information about the students. Recently there was a debate on the admissions committee about whether to admit a particular student from Horace Maynard College. The student's scores on the GRE and his GPA were marginal—that is, almost all students actually admitted to the department have scores as high or higher, while most rejected students have lower scores. The student's letters of recommendation were quite good, but none of the writers of the letters were personally known to any of the Michigan faculty.

One member of the admissions committee argued against admission, pointing out that department records show that students who graduate from small, nonselective colleges like Maynard perform at a level substantially below the median of all Michigan graduate students. This argument was countered by a committee member who noted that 2 years ago Michigan had admitted a student from Maynard who was now among the three highest ranked students in the department.

Comment on the arguments put forward by these two committee members. What are their strengths and weaknesses?

Objective—Structure 5

The superintendent of schools was urging the school board to make an expensive curriculum shift to a ''back-to-basics'' stress on fundamental learning skills and away from the electives and intensive immersion in specialized arts and social studies topics that had recently characterized the secondary schools in the district. He cited a study of 120 school systems that had recently begun to emphasize the basics and 120 school systems that had a curriculum similar to the district's current one. The ''back-to-basics'' school systems, he said, were producing students who scored half-a-year ahead of the students in the other systems on objective tests of reading, mathematics, and science. Of the 120 ''back-to-basics'' school systems, 85 had shown improved skills for students in the system vs only 40 with improved skills in the 120 systems which had not changed. One of the school board members took the floor to argue against the change. In her opinion, she said, there was no compelling reason to attribute the improved student skills in the ''back-to-basics'' systems to the specific curriculum change, for two reasons: (1) school systems that make curriculum changes probably have more energetic, adventurous administrators and faculty and thus the students would learn more in those school systems no matter what the curriculum was. (2) *Any* change in curriculum could be expected to produce improvement in student performance because of increased faculty interest and commitment.

Comment on the reasoning of both the superintendent and the board member. On the basis of the evidence and arguments offered, do you think it is likely that the ''back-to-basics'' curriculum is intrinsically superior to the district's current curriculum?

Objective—Structure 6

An economist was arguing in favor of a guaranteed minimum income for everyone. He cited a recent study of several hundred people in the United States with inherited wealth. Nearly 92% of those people, he said, worked at some job that provided earned income sufficient to provide at least a middle-class life style. The study showed, he said, that contrary to popular opinion, people will work in preference to being idle. Thus a guaranteed income policy would result in little or no increase in the number of people unwilling to work.

Comment on the economist's reasoning. Is it basically sound? Does it have weaknesses?

Subjective—Structure 1

Gerald M. had a 3-year-old son, Timmy. He told a friend: "You know, I've never been much for sports, and I think Timmy will turn out the same. A couple of weeks ago, an older neighbor boy was tossing a ball to him, and he could catch it and throw it all right, but he just didn't seem interested in it. Then the other day, some kids his age were kicking a little soccer ball around. Timmy could do it as well as the others, but he lost interest very quickly and started playing with some toy cars while the other kids went on kicking the ball around for another 20 or 30 min."

Do you agree with Gerald's reasoning that Timmy is likely not to care much for sports? Why or why not?

Subjective—Structure 2

David L. was a senior in high school on the East Coast who was planning to go to college. He had compiled an excellent record in high school and had been admitted to his two top choices: a small liberal arts college and an Ivy League university. The two schools were about equal in prestige and were equally costly. Both were located in attractive East coast cities, about equally distant from his home town. David had several older friends who were attending the liberal arts college and several who were attending the Ivy League university. They were all excellent students like himself and had interests that were similar to his. His friends at the liberal arts college all reported that they liked the place very much and that they found it very stimulating. The friends at the Ivy League university reported that they had many complaints on both personal and social grounds and on educational grounds. David initially thought that he would go to the liberal arts college. However, he decided to visit both schools himself for a day. He did not like what he saw at the private liberal arts college: several people whom he met seemed cold and unpleasant; a professor he met with briefly seemed abrupt and uninterested in him; and he did not like the "feel" of the campus. He did like what he saw at the Ivy League university: several of the people he met seemed like vital, enthusiastic, pleasant people; he met with two different professors who took a personal interest in him; and he came away with a very pleasant feeling about the campus.

Which school should David L. choose, and why? Try to analyze the arguments on both sides, and explain which side is stronger.

Subjective—Structure 3

Janice is head nurse in a home for the aged. She says the following of her experiences: "There is a big turnover of the nursing staff here, and each year we hire 15–20 new nurses. Some of these people show themselves to be unusually warm and compassionate in the first few days. One might stay on past quitting time with a patient who's having a difficult night.

Another might be obviously shaken by the distress of a patient who has just lost a spouse. I find though that, over the long haul, these women turn out to be not much more concerned and caring than the others. What happens to them, I think, is that they can't remain open and vulnerable without paying a heavy emotional price. They usually continue to be considerate and effective but they build up a shell."

Comment on Janice's reasoning. Do you think it is likely that she correctly identifies the nurses who are unusually warm and compassionate? Do you agree it is likely that most of the ones who are unusually warm at first later build up a shell to protect themselves emotionally?

Subjective—Structure 4

The director of a Broadway production of Shakespeare's *As You Like It* had just finished auditions for the female lead in the show. Two of the candidates gave readings for the part he liked a great deal. Another was an actress whom the director had worked with before in three Shakespeare comedies. The director thought she had been superb in each. Unfortunately, of her three readings for the lead in this play, one had been fairly good, but two had been quite flat. This third actress had to know immediately whether she was going to be chosen for the part. If not, she would take a minor role in a movie that would keep her on the West Coast for the next 6 months.

What should the director do—hire the third actress or hire one of the two whose readings he liked better? Why?

Subjective—Structure 5

Two New Yorkers were discussing restaurants. Jane said to Ellen, "You know, most people seem to be crazy about Chinese food, but I'm not. I've been to about 20 different Chinese restaurants, across the whole price range, and everything from bland Cantonese to spicy Szechwan and I'm really not very fond of any of it." "Oh," said Ellen, "don't jump to conclusions. I'll bet you've usually gone with a crowd of people, right?" "Yes," admitted Jane, "that's true. I usually go with half a dozen people or more from work." "Well, that may be it," said Ellen. "People usually go to Chinese restaurants with a crowd of people they hardly know. I know you, you're often tense and a little shy, and you're not likely to be able to relax and savor the food under those circumstances. Try going to a Chinese restaurant with just one good friend. I'll bet you'll like the food."

Comment on Ellen's reasoning. Do you think there is a good chance that if Jane went to a Chinese restaurant with one friend, she'd like the food? Why or why not?

Subjective—Structure 6

Martha was talking to a fellow passenger on an airplane. The fellow passenger was on his way to Hawaii for a month's vacation. "I don't like vacations myself," Martha said. "I've always worked. I put myself through college and law school and now I have a full-time legal practice. Frequently, of course, I've had slow periods when I wasn't working at all, but I never liked those times. For example, there would usually be a week or two between the end of school and the beginning of a summer job and another week or two of enforced idleness at the end of the summer. And there were many occasions when I was getting started in my career when I had no real work to do for fairly long periods. But I never enjoyed the leisure. I know there are some people who talk about using vacations to "recharge" themselves. But I suspect many of these people don't really enjoy their work or don't have a very high energy level. I do have a lot of energy, and I do enjoy my work, and I guess that's why I don't really like vacations."

Analyze Martha's reasoning. Do you think she had good evidence for feeling she doesn't like vacations?

APPENDIX B

The Three Objective Example Problems Used in Experiment 1, Also Used in the Objective Examples Training Condition of Experiment 2

Example 1 (Structure 1)

A major New York law firm had a history of hiring only graduates of large, prestigious law schools. One of the senior partners decided to try hiring some graduates of smaller, less prestigious law schools. Two such people were hired. Their grades and general record were similar to those of people from the prestigious schools hired by the firm. Although their manners and "style" were not as polished and sophisticated as those of the predominantly Ivy League junior members of the firm, their objective performance was excellent. At the end of 3 years, both of them were well above average in the number of cases won and in the volume of law business handled. The senior partner who had hired them argued to colleagues in the firm that, "This experience indicates that graduates of less prestigious schools are at least as ambitious and talented as graduates of the major law schools. The chief difference between the two types of graduates is in their social class background, not in their legal ability, which is what counts."

Comment on the thinking that went into this senior partner's conclusion. Is the argument basically sound? Does it have weaknesses? (Disregard your own initial opinion, if you had one, about graduates of nonprestigious law schools, and concentrate on the thinking that the senior partner used.)

Please consider this problem for a few moments. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.

The senior partner is trying to draw a conclusion about a certain population. We can think of the members of this *population* as newly graduated lawyers, from nonprestigious law schools, who otherwise meet the law firm's hiring standards. If we divide the members of this population into *two categories,* "excellent" and "mediocre or worse," we can think of the *population distribution* as the percentage in each category. The senior partner has concluded that the percentage in the "excellent" category is very high, or anyway, just as high as in another population, involving graduates of prestigious law schools. This conclusion was based on observing a *sample* of *size = 2,* in which the *sample distribution* was 100% "excellent," 0% "mediocre or worse."

Apart from any other considerations, however, the *sample distribution* for size 2 is apt to be quite different from the *population distribution:* the latter could be only 60 or 50% or even perhaps as low as 40% "excellent," and a 2–0 sample split would not be so unusual; just as one would not be at all amazed to draw two out of two red gumballs from an urn with only 40% reds. So the senior partner's attitude is quite unwarranted: a larger sample is needed.

Example 2 (*Structure 3*)

Susan is the artistic director for a ballet company. One of her jobs is auditioning and selecting new members of the company. She says the following of her experience: "Every year we hire 10–20 young people on a 1-year contract on the basis of their performance at the audition. Usually we're extremely excited about the potential of 2 or 3 of these young people—a young woman who does a brilliant series of turns or a young man who does several leaps that make you hold your breath. Unfortunately, most of these young people turn out to be only somewhat better than the rest. I believe many of these extraordinarily talented young people are frightened of success. They get into the company and see the tremendous effort and anxiety involved in becoming a star, and they get cold feet. They'd rather lead a less demanding life as an ordinary member of the corps de ballet."

Comment on Susan's reasoning. Why do you suppose that Susan usually has to revise downward her opinion of dancers that she initially thought were brilliant?

Please consider this problem for a few moments. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.

We can analyze this problem using the law of large numbers by thinking of each ballet dancer as possessing a *population* of ballet movements. Susan is interested in excellence, so we can divide the members of each population into two categories: "brilliant movements" and "nonbrilliant, or other movements." We can think of the *population distribution* as the percentage or proportion in each category. For many dancers, the population distribution is actually 0% brilliant and 100% other: these dancers simply lack the talent to perform a brilliant movement. For many other dancers, there is a small or moderate percentage of "brilliant movement" gumballs in their urn. A true ballet star would therefore have a population distribution with a greater percentage of "brilliant" movements than an ordinary member of the corps de ballet.

By conducting auditions, Susan is observing *samples* of each dancer's population distribution. An audition, however, is a very small sample of a dancer's movements. We know from the law of large numbers that small samples are very unreliable estimates of the population. When a dancer performs some brilliant moves during an audition, it is often because the dancer has happened to draw a couple of the "lucky gumballs" that day: it does not prove that the population distribution for that dancer consists of a large percentage of "brilliant movements." It is reasonable to think that there are really very few dancers that have population distributions with a large percentage of brilliant movements; and so when Susan sees a dancer performing brilliantly at audition, the chances are it is just a lucky draw from a dancer who is capable of performing some, but not necessarily a great number of "brilliant movements." Therefore, when Susan hires such dancers and evaluates them after seeing a much larger sample of their movements, it is not surprising that she finds that many of these dancers that were brilliant at audition turn out to be only somewhat better than the rest.

Example 3 (*Structure 5*)

Kevin, a graduate student in sociology, decided to do a research project on "factors affecting performance of major league baseball players" in which he gathered a great amount of demographic data on birthplace, education, marital status, etc., to see if any demographic factors were related to the performance of major league baseball players (e.g., batting average, pitching victories). Kevin was unable to use data for all the major league teams because information for some of the players was unavailable, but he was able to obtain data for some 200 players in the major leagues.

One finding that interested Kevin concerned the 110 married players. About 68% of these players improved their performance after getting married, while the remainder had equal or poorer performance. He concluded that marriage is beneficial to a baseball player's performance. At a social hour sponsored by the Office of the Commissioner of Major League Baseball, he happened to mention his finding to a staff member of the office. The staff member listened to Kevin's results and then said, "Your study is interesting but I don't believe it. I'm sure that baseball performance is worse after a marriage because the ball player suddenly has to take on enormous responsibilities: taking care of his spouse and children. Plus the factor of being stressed by having to be on the road so much of the time and therefore away from the family. The player will no longer be able to devote as much time to baseball as before he was married. Because of this he will lose that competitive quality that is necessary for good performance in baseball.

What do you think of the staff member's argument? Is it a sound one or not? Explain your reasoning.

Please consider this problem for a few moments. After you have considered the problem and analyzed it for a minute or two, turn the page for our analysis.

Kevin is trying to find out how performance in major league baseball is affected by being married. To do this, he obtained data for 200 players in the major leagues and discovered that out of the 110 that had gotten married, 68% had improved performance after the wedding (and 32% had equal or poorer performance). According to the law of large numbers, which states that the larger the sample, the better it is in estimating the population, there is substantial evidence that marriage is beneficial to baseball players' performance. Recall that in the gumball demonstration, samples of size 25 were very good estimates of the population: these samples did not differ much from population. Extending the argument, samples of size 110 are extremely accurate estimates of the population. Thus, it can be concluded that, in general, marriage is beneficial to baseball players' performance.

What about the staff member's theory that baseball performance is worse after a marriage because the ball player assumes enormous responsibilities and will no longer be able to devote as much time to baseball as before? Although this argument may have some intuitive appeal, it should be discounted because it is not supported by any data and is, in fact, contradicted by Kevin's large sample of 110 players.

REFERENCES

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge, MA: MIT Press.

D'Andrade, R. (1982, April). *Reason versus logic*. Paper presented at the Symposium on the Ecology of Cognition: Biological, Cultural and Historical Perspectives, Greensboro, North Carolina.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology*, 32, 53–88.

Evans, J. St. B. T. (1982). *The psychology of deductive reasoning*. London: Routledge & Kegan Paul.

Golding, E. (1981). *The effect of past experience on problem solving*. Paper presented at the Annual Conference of the British Psychological Society, Surrey University.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect In Wason's selection task. *British Journal of Psychology*, 73, 407–420.

Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology*, 39, 578–589.

Haberman, S. J. (1972). Log-linear fit for contingency tables—Algorithm AS51. *Applied Statistics,* **21,** 218–225.

Hogarth, R. M. (1980). *Judgment and choice: The psychology of decision.* New York: Wiley.

Jepson, C., Krantz, D. H., & Nisbett, R. E. (1983). Inductive reasoning: Competence or skill? *Behavioral and Brain Sciences,* **6,** 494–501.

Johnson-Laird, P. N., Legrenzi, P., & Sonino-Legrenzi, M. (1972). Reasoning and a sense of reality. *British Journal of Psychology,* **63,** 395–400.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge Univ. Press.

Kahneman, D., & Tversky, A. (1971). Subjective probability: A judgment of representativeness. *Cognitive Psychology,* **3,** 430–454.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review,* **80,** 237–251.

Manktelow, K. I., & Evans, J. St. B. T. (1979). Facilitation of reasoning by realism: Effect or non-effect? *British Journal of Psychology,* **70,** 477–488.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Fong, G. T. (1982). Improving inductive inference. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge Univ. Press.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review,* **90,** 339–363.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Reich, S. S., & Ruth, P. (1982). Wason's selection task: Verification, falsification and matching. *British Journal of Psychology,* **73,** 395–405.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science (Washington, D.C.),* **185,** 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review,* **90,** 293–315.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.). *New horizons in psychology I.* Harmondsworth, England: Penguin.

(Accepted January 14, 1986)