RULES FOR REASONING

Edited by
Richard E. Nisbett

LEA

RULES FOR REASONING

Edited by

Richard E. Nisbett
University of Michigan

LEA LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1993 Hillsdale, New Jersey Hove and London

Lawrence Erlbaum Associates, Inc., Publishers
365 Broadway
Hillsdale, New Jersey 07642

Library of Congress Cataloging-in-Publication Data

Rules for reasoning / edited by Richard E. Nisbett.
p. cm.
Includes bibliographical references and indexes.
ISBN 0-8058-1256-3 (cloth). -- ISBN 0-8058-1257-1 (paper)
1. Reasoning. I. Nisbett, Richard E.
BC177.R76 1993
160--dc20 92-39042
CIP

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability.

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

For Matthew and Sarah

Acknowledgments

The editor wishes to acknowledge with thanks the many journals that gave their permission to reprint the following articles in this volume.

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

Thagard, P., & Nisbett, R. E. (1982). Variability and confirmation. *Philosophical Studies, 50,* 250–267.

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

Fong, G. T., Krantz, D. H. & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253–292.

Fong, G. T., & Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology, 120,* 34–45.

Cheng, P., Holyoak, K. J., Nisbett, R. E., & Oliver, L. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18,* 293–328.

Larrick, R. P., Morgan, J. N., & Nisbett, R. E. (1990). Teaching the use of cost-benefit reasoning in everyday life. *Psychological Science, 1,* 362–370.

Larrick, R. P., Nisbett, R. E., & Morgan, J. N., (In press). Who uses the normative rules of choice? *Organizational Behavior and Human Decision-Making.*

Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science, 238,* 625–631.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday life events. *American Psychologist, 43,* 431–443.

Lehman, D., & Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate education on reasoning. *Developmental Psychology, 26,* 952–960.

Smith, E. E., Langston, C., & Nisbett, R. E. (1992). The case for rules in reasoning. *Cognitive Science, 16,* 99–102.

Contents

1 Reasoning, Abstraction, and the Prejudices of 20th-Century Psychology

R. E. Nisbett
University of Michigan

Twentieth-century psychology has had a strong prejudice against abstraction, that is, against the view that the world is understood by means of rules that transcend the perception of a particular physical stimulus or the comprehension of a domain of related events. In the United States, the prejudice has been bound up with behaviorism and its successor positions. Behaviorists were determined to find the equivalent of the reflex arc in physiology—stimulus-response linkages that could be described with precision by a physical description of the stimulus, the response, and the conditions of their co-occurrence during learning. So complete was their dedication to such physical description that they felt confident that the study of animals could substitute for the study of humans in building a complete theory of behavior.

Early in the 19th-century, the behaviorist E. L. Thorndike performed a series of experiments that satisfied two generations of American psychologists that abstractions were not importantly involved in learning how to perform skilled tasks. He asked his subjects to perform a particular task for varying amounts of time (e.g., cancelling Os from a sentence, and then switched them to another task; cancelling adverbs from a sentence). He found that "transfer of training" effects were slight and unstable. Sometimes he found that performance of the first task enhanced the second, sometimes that it made it more difficult, and, often, that it had no effect at all. One would, of course, assume that performance on the second task would be improved if subjects learned something general from performance of the first task. Since they so often failed to show improved training, Thorndike inferred that people don't, in fact, learn much that is general when performing mental tasks. This meant that training was going to be very much a bottom-up affair, consisting of little more than slogging through countless stimulus-response associations.

This conclusion has suffused deeply into American psychology, cognitive science, and education. Newell (1980), based on some similar failed efforts to find training effects for reasoning tasks, has asserted that learned problem-solving skills generally are idiosyncratic to the task. Just as the earlier behaviorists took the evidence of weak transfer-of-training effects to buttress their case for the exclusive role of specific stimulus-response linkages, some modern cognitive scientists have used such evidence to support connectionism—the modern successor to behaviorism. To the connectionist, all learning is just a matter of adding strength values to an initially neutral "network" of highly specific elements. The connectionists have all the courage of the behaviorists' convictions—asserting that they can mimic the important details of learning and cognitive performance without the postulation of any rules whatever.

Other trends of modern psychology are opposed to abstract rules, though not necessarily to rules of all kinds. Devotees of case-based reasoning approaches to problem solving hold that people do little more when solving problems than perceive similarities between old and new problems and occasionally apply strategies of analogy construction. Biological and evolutionary theories of cognition are sympathetic to the notion that people operate using rules of limited generality, but these are usually assumed to be limited to relatively tight domains. Thus there are rules, even prewired rules, for language, or for physical causality, or for particular types of social relations, but these are limited to particular content domains and would never be used for understanding events outside those domains.

European psychology has never been so deeply antiabstractionist as American psychology. In fact, Jean Piaget, the European psychologist whose influence on world psychology has been greatest, explicitly endorsed the notion that there are abstract rules that guide thought and behavior. Piaget even thought that the very most abstract rules, those of formal logic, have their intuitive counterparts in the human cognitive repertoire. These rules, part of the equipment that Piaget called *propositional operations,* are used to acquire other, somewhat less abstract but still domain-independent rules in the course of development. These are the *formal operations,* which include the concept of proportionality, the notion that every action has a reaction, and what Piaget called the *probability schema,* but most people today would call the law of large numbers. Piaget believed that people possessed these highly abstract rules in a form in which they made contact with the most ordinary problems in everyday life. Indeed, many common problems could not be solved without the use of such rules, and it was the press of such requirements that pushed people toward their acquisition, by means of inductive procedures operating within the logical constraints of the propositional operations.

Despite his endorsement of abstract rules for reasoning, Piaget was quite firm in his opinion that such rules could not be explicitly taught—certainly not by abstract or formal means. Such rules are the common equipment of every adult, and everyone acquires them by virtue of being the kind of organism that each human is and by virtue of living in the kind of world that each human does. Given our native equipment and the kind of experiences we are going to have, we perforce learn the abstract rules we are going to need for purposes of deductive and inductive inference. But such rules are learned only by an inductive process of self-discovery; the day of their acquisition cannot be hastened either by abstract instruction in the rule system or by a forced inductive march through specific problems. Nothing, not even an abstract rule, is learned by abstract or top-down training procedures. Nor is there any point in trying to fool Mother Nature by excessive drilling on concrete problems. No rule will be learned before the organism is ready for it and learning is inevitable once the organism is ready—so long as it is not kept in a closet.

You will read Piaget in vain for any evidence for this extraordinarily influential theory of how rules for reasoning are learned. It was simply obvious to him that you cannot teach abstract rules of reasoning, and it became equally obvious to us largely because of his enormous prestige and persuasiveness. There are to be sure shreds of evidence available since Piaget's time that are consistent with his view: (a) solutions to the missionaries and cannibals problem do not generalize to formally identical problems; (b) accelerated learning of conservation of mass for clay does not seem to generalize to an understanding of conservation in general. And there remains Thorndike's work, showing that there can be little transfer of training even across tasks that require lower levels of cognitive skill than one would want to dignify with the term *reasoning*.

But what if Piaget and, even more so, the American psychological tradition, were mistaken about abstractions? What if you actually could teach people highly abstract rules of reasoning—and even do so by highly abstract and therefore efficient means? What if such instruction resulted in people being able to apply those rule systems potentially to the full range of problems in everyday life for which they are relevant? How would we think about the human mind then? How would we think about education?

This volume tries to answer those questions, coming up with some very surprising answers. Ten years ago, I held a version of the received views about reasoning. I was dubious that people had any abstract rules for reasoning and confident that even if they did, such rules could not be taught. Indeed, I had just completed 10 years of work that seemed to me to give substantial support to these views. I had worked on questions of reasoning about human social behavior, finding that people often violated the requirements of statistical, causal, and even logical rules of inference. This work was very much in the tradition of Kahneman and Tversky's research showing that people substitute simple judgmental heuristics for the more formal inferential rules that are necessary to solve the problems they gave their subjects. I believed not only that my subjects did not possess the necessary statistical rules, I believed that instruction in statistics resulted only in inserting a sterile set of formal rules that could make contact only with scientific problems or problems for which there existed some massive and probably ecologically uncommon cue triggering their use.

With Geoffrey Fong and David Krantz, I began what I thought would be a swift program of work establishing these points, namely that instruction in statistics does very little toward helping people to solve everyday problems that require a statistical solution. My very first attempt to look at this question showed me that I was wrong (or should have showed me—actually I didn't believe the implications at first). Kahneman and Tversky (1972) had developed a clever problem to show subjects' statistical weaknesses called the *maternity ward problem.* In this problem subjects are told that there is a town with two hospitals, one large and one small. At the large hospital, about 60 babies a day are born, and at the small hospital about 15. Subjects are then asked at which hospital they think there would be more days during the year in which 60% or more of the babies born would be boys. About one third of the undergraduates they studied believed it would be the larger hospital, about one third believed it would be the smaller, and about one third believed it would make no difference. The law of large numbers, of course, requires that it would be the smaller hospital, because deviant sample proportions are likely in inverse proportion to sample size. While teaching an upper level undergraduate class at the University of Michigan, I tried to duplicate these results in a classroom demonstration. To my surprise, most of the students got the problem right. I then asked students to indicate how much statistics they had had as well as their preferred answer. The results were clear-cut. The students who had had no statistics duplicated the pattern of the Kahneman and Tversky subjects, those who had had at least one course in statistics were unlikely to get the problem wrong.

Subsequent work showed there was no anomaly here. Problems that Kahneman and Tversky had looked at, as well as problems with more social content of the kind I had looked at, turned out to be highly influenced by statistical training. This was true even for problems without obvious statistical clues. For example, one problem we gave subjects asked why someone who had an excellent meal in a restaurant might be likely to complain about a less good meal the next time around. Untrained undergraduates almost always gave purely deterministic answers such as "maybe restaurants change their chefs a lot." But subjects who had had many courses in statistics usually gave statistical answers, such as "there are probably more restaurants where you can get an excellent meal some of the time than there are restaurants where you can get an excellent meal all the time, so a person who gets an excellent meal the first time has to assume it's likely that the next one won't be."

We then began seeing whether you could teach such statistical reasoning in short training programs. We found to our surprise that even very brief interventions could produce fairly pronounced effects on the sorts of answers subjects would give to problems of the Kahneman and Tversky type. Indeed, purely abstract training, in which we defined the terms *sample, population, parameter,* and *variance,* and explained the relations among *variability, N,* and sample-parameter accuracy, even had an effect on solution of problems with purely social content. Moreover, training in a given domain, for example, training on problems concerning sports transferred fully to another domain, for example, problems concerning ability tests. In several studies, we found literally no advantage for the trained domain over other domains so long as testing was immediate. Such results are consistent with the view that people can operate with very abstract rules indeed, and that the techniques by which they learn them can be very abstract. Abstract improvements to the preexisting intuitive rule system are passed along to the full range of content domains where the rules are applicable, and improvements in a given domain are sufficiently abstracted so they can be applied immediately to a very different content domain.

It is important to note that the sort of instructional effects we discovered in these studies are by no means limited to laboratory or academic settings. When subjects are contacted outside of such settings (e.g., in the context of an opinion poll), the trained subjects answer questions differently from the untrained. In one study, male college students who had either just begun or just finished their first statistics course were asked to participate in a poll on opinions about sports. After answering a number of questions about the National Collegiate Athletic Association rules and National Basketball Association salaries, they were reminded that the top batters in both baseball leagues typically have averages of .450 or higher at the end of the first two weeks of play, yet no one has ever finished the season with such a high average. The students were asked to explain why they thought this was the case. The students just beginning statistics nearly always responded with purely deterministic answers such as "the pitchers make the necessary adjustments." The students who had taken the course were twice as likely as novices to give a statistical answer, such as "two weeks isn't a very long time, so you get some atypically high (and low) averages; no one really has the ability to hit .450 over the long haul."

Over the next 10 years, I pursued the implications of these findings on trainability with different sets of colleagues who were expert in particular rule systems, including the self-selection concept critical to control procedures in the social sciences, "pragmatic reasoning schemas" for contractual relations such as permission and obligation, rules for assessing causality, and the cost-benefit rules of microeconomic theory. The generalizations below hold for these rule systems taken as a group. All of the generalizations have been tested on at least three different abstract rule systems; none are contradicted by any evidence I am aware of.

1. People have intuitive versions of these formal rule systems that they apply to at least some problems in everyday life. We know this because they solve problems that require use of the rule systems, because they articulate the rule systems in justifying their solutions, and because instruction in the rule systems increases the correct solution of the problems.
2. People at a given level of education, prior to formal instruction in a particular rule system, differ in the degree to which they understand the rule system and are able to apply it to solve concrete problems. Such individual differences are associated with verbal intelligence.
3. Formal education beyond secondary school produces dramatic differences in people's use of different rule systems. It is no exaggeration to say that people who have substantial knowledge of statistics, or of economics, view the world very differently from those who do not. All sorts of mundane problems are understood differently by people with differing levels of education in the relevant rule system.
4. The rule systems are embodied at a level of abstraction equal to that posited by Piaget for the so-called formal operations. The absence of domain specificity is a striking observation across training studies, as is the ability of investigators to "insert" the rules by purely formal and abstract instructional means.
5. Despite their abstract nature, the rules are not applied across all domains equally. The same student who has no trouble applying statistical rules to the behavior of random generating devices, such as dice, may apply statistical rules rarely or never to problems with social content. Problems differ a great deal in how transparent they are with respect to a rule system necessary for their solution.
6. A consequence of people's differential ability to apply rules in different domains is that training in coding a given domain in terms of the rule can have dramatic effects—making it possible for people to apply a rule they already have to a new domain where previously it was unlikely for them to use the rule.

The upshot of these findings is that modern cognitive science and modern educational theory must accommodate themselves to the existence of abstract inferential rules. Psychological theories that hold that there are no rules, or no domain-independent rules, for problem solving, are not tenable in the light of the work presented in this book. Educational positions that emphasize self-discovery and maturation must make room for the generalization that abstract techniques of instruction can be very powerful. Psychological and educational positions, as well as philosophical positions, that assume a universal adult competence with respect to reasoning must give way to the recognition that adult inferential competence is highly variable and highly dependent on educational history.

The rest of the volume presents work making these points in detail. Some of the chapters have been published before and some were written especially for this volume. (Reference styles are not consistent across the different papers. We chose to leave the previously-published material in the same form in which it appeared originally, minus their abstracts.)

Part I documents the existence of abstract, intuitive, and statistical rules. Chapter 2, by Nisbett, Krantz, Jepson and Kunda, shows that people without formal training in statistics solve problems using the law of large numbers and actually articulate the rule in justification of their answers. This chapter also shows that the presence of various cues about the partially random nature of the events in a given problem can dramatically affect the likelihood that people will apply the law of large numbers to the problem.

Chapter 3, by Thagard and Nisbett, proposes a solution to Hume's riddle of induction, namely "Why is a single instance, in some cases, sufficient for a complete induction, whereas in others myriads of concurring instances, without a single exception known or presumed, go such a very little way towards establishing a universal proposition?" The solution lies in the law of large numbers, coupled with real world knowledge about the variability of kinds of objects with respect to kinds of properties. A "single instance is sufficient for a complete induction" when we take it for granted that objects of the kind in question are invariant with respect to properties of the kind observed. For example, observing the color of a sample of a new chemical element leaves us in little doubt about the color of future samples. Myriads of concurring instances do not convince when we take it for granted that the kind of object is highly variable with respect to the kind of property observed. For example, observing a bird in the rain forest that is green does not convince us that the next bird we see of the same type will be green because we do not assume invariability of color for bird types.

Chapter 4 directly attacks the notion assumed by many philosophers and psychologists that there is a single human inferential competence. Some philosophers, notably Jonathan Cohen, have argued that empirical demonstrations of human inferential error are logically impossible since "Ordinary human reasoning—by which I mean the reasoning of adults who have not been systematically educated in any branch of logic or probability theory—cannot be held to be faultily programmed: It sets its own standards" (1981, p. 317). One wonders why it is untutored reasoning that Cohen presumes to be correct rather than tutored reasoning, but even if one accepts that untutored reasoning gets the normative nod, the position is untenable on empirical grounds. Untutored people differ dramatically in their preferred solutions for concrete inferential problems and in the abstract rules for inference that they endorse. Even within a given culture, people differ so much that it is impossible to identify a single competence and establish it as a normative standard.

Part II presents the evidence on alteration of statistical rules. It shows that both standard types of statistical training and various experimental versions can drastically change people's understanding of events characterized by random or partially random determination. Both top-down, abstract training and bottom-up training within a given domain have widespread effects on reasoning.

Parts III and IV of this volume address the question of whether people have abstract rules for deductive reasoning. More than 20 years ago, Wason and Johnson-Laird raised the possibility that people do not have deductive rules, at least not in a form that makes contact with everyday problems. They asked their subjects to perform a simple task. The subjects were shown four cards reading A, B, 4 and 7. Subjects were informed that the cards had letters on one side and numbers on the other and were directed to turn over as many cards, and only as many, as were necessary to find out whether the rule, "If a card has an A on one side, then it has a 4 on the other" is violated. Interpreting the "if-then" connective as the material conditional in standard logic, the correct answer in this example is to turn over the cards showing A and 7. The rule used in such problems is a conditional statement, "if *p* then *q*," and the relevant cases are *p* (because if *p* is the case it must be established that *q* is also the case) and *not-q* (because if it is not the case that *q* it must be established that it is also not the case that *p*). Fewer than 10% of college students can solve such problems. Yet, solving it merely requires the application of the material conditional—the cornerstone of standard logics.

Over the past two decades, many people have tried their hand at resolving the so-called *selection task* conundrum. The one favored in this book follows the lead of Cheng and Holyoak (1985), who proposed that people probably make little if any use of the rules of formal logic, certainly not of modus tollens (which states the equivalence of "if *p* then *q*" to "if *not-q* then *not-p*" and which is required to solve the abstract version of the selection task problem). What people do have are "pragmatic reasoning schemas"—highly generalized, domain-independent, but not purely syntactic rule systems. Such schemas include Piaget's formal operations such as the probability schema or the law of large numbers and generalized rule systems for analyzing causality (Kelley 1972, 1973). Cheng and Holyoak (1985) and Cheng and her colleagues in the present volume have argued that similar pragmatic reasoning schemas govern contractual relations such as permission and obligation. Thus the obligation schema ("if *p* occurs, one is obliged to do *q*") implies that one would have a violation in two of four possible cases: *p* occurs but *q* is not carried out and *q* is not carried out even though *p* occurred. These authors find that simply invoking the semantic notion of obligation allows people to solve with ease problems formally identical to the selection task. Moreover, formal instruction in how to solve contractual-schema problems is highly effective while instruction in formal logic is of no use for such problems.

The two chapters in Part IV try to extend the notion of pragmatic reasoning schemas to the case of rules for causality. Though Kelley proposed the existence of such schemas and many theorists have taken it for granted that they exist, rigorous evidence in support of such schemas has not been provided. Cheng and Nisbett (see Chap. 8) find some evidence that deductive reasoning about causal relations is influenced by pragmatic considerations concerning expectations about the relative probability of an effect given that a cause is or is not present. Morris and Nisbett (see chap. 10) find that graduate instruction in psychology, which emphasizes assessment of causality, improves students' ability to reason deductively about causal relations. Instruction in other graduate fields has little or no effect on students' ability to reason about causal relations. This suggests that deductive rules specifically governing causal relations exist and can be formally taught. Taken together, the two sets of studies suggest that, when deductive reasoning centers on specifically causal relations, specifically causal rules are invoked. These rules are highly abstract, in that they are completely independent of domain, but they are not as abstract as the rules of formal logic which are indifferent not merely to type of entities under consideration but to type of relationship.

Part V of this volume deals with rules for choice. Economists have long argued that all choice makes use of cost-benefit rules, which require people to assess values of possible outcomes as well as their probabilities, to note "opportunity costs" of their behavior (i.e., value lost by continuing a course of action rather than switching) and to ignore "sunk costs" (i.e., never to carry out some action simply because value has already been expended—tickets bought, etc.). Psychologists are of course quite predisposed to doubt the existence of such an abstract rule system. Moreover, many clever psychologists from Herbert Simon to Kahneman and Tversky have found it easy to show that people's choices often depart grossly from those that would be dictated by an application of cost-benefit rules.

Nevertheless, I present evidence here that people do possess a version of cost-benefit rules. It is just that, as with the law of large numbers, the rule system they possess merely overlaps with the formal, prescriptive one and is not identical to it. Indeed, people can be shown to endorse, and even to articulate spontaneously, a choice rule that is diametrically opposed to the sunk-cost rule. However, as with the other pragmatically useful rule systems discussed in this volume, cost-benefit rules can be taught. When they are, people subsequently reason differently about a huge range of choice problems. Economists make different choices than do biologists, undergraduates who have taken a course in economics reason differently than do those who have not, and even a brief session of teaching the sunk-cost rule causes people to make very different choices than those who have not had such training.

Part V also presents evidence that people are better off using the microeconomic rules of choice, a claim made from the beginning by economists but never tested by them. Professors who are more likely to apply the microeconomic rules of choice in their daily lives make more money than those who are less likely to apply them. I believe this is the case because sound choice principles make people more effective in their work, which is recognized by their higher salaries. Similarly, college students who are more likely to apply the microeconomic rules of choice have higher grade point averages than those who are less likely to apply them. This is not the case simply because brighter students are both more likely to know and use the rules and to get higher grade point averages. The relation between rule use and Grade Point Average is actually higher when intelligence (verbal Scholastic Aptitude Test) is partialled out of the relationship. Use of the microeconomic rules is thus associated with overachievement.

Part VI spells out the implications of the research for higher education. Twenty-five hundred years ago, Plato enunciated the educational doctrine that held sway in the west until this century. This was the view that instruction in formal rule systems improved people's ability to reason. "Even the dull," he said, "if they have had an arithmetical training . . . always become much quicker than they would otherwise have been" (cited in Jowett, 1875, p. 785). The Romans added the study of grammar to the study of arithmetic and geometry; the medieval scholastics added the syllogism and the humanists added the study of Latin and Greek, and this formed the core of the curriculum until well into the nineteenth century. The rationale was Plato's "formal discipline" theory: The study of abstract rules improves reasoning.

The first policy victory of modern psychology was to destroy the intellectual basis for the classical curriculum. William James mocked the theory as recommending mere exercise for the nonexistent "muscles of the mind." Thorndike's transfer-of-training findings provided all the empirical evidence that was needed against the notion that teaching one kind of rule in one kind of context with one kind of material could have the slightest effect in another context with another kind of material (and probably a different rule to boot).

The critique was entirely successful, and deserved to be. Learning Latin, in fact, probably does nothing for reasoning, arithmetic probably does nothing for any mental operations except the purely arithmetical, and even training in syllogisms probably does little good. Bertrand Russell (1960) had this to say of the syllogism:

> The inferences that we actually make in daily life differ from those of syllogistic logic in two respects, namely, that they are important and precarious, instead of being trivial and safe. The syllogism may be regarded as a monument to academic timidity: if an inference might be wrong, it was dangerous to draw it. So the medieval monks, in their thinking as in their lives, sought safety at the expense of fertility. (p. 83)

So the behaviorists probably were throwing out bathwater when they insisted that there was no general inferential benefit of much consequence from study of the classical curriculum. But there was also a baby in that bathwater, a baby unknown to pedagogues of yore, from Plato onward, but a baby nonetheless. That baby was the set of all pragmatically useful inferential rule systems whose use can be increased by explicit instruction. What rules are in that set? We don't know them all yet, but we can certainly identify some: the law of large numbers, the confounded-variable principle, causal schemas, social contract schemas, and cost-benefit rules of choice.

What we know about this list of rules is fairly impressive and more than justifies their inclusion in the curriculum:

1. People can make better inferences if they know these rules than if they don't.
2. Some people have a better grasp of each of these rules than others.
3. Everyone's grasp of these rules can be improved by instruction. Different graduate courses, and even different undergraduate majors, emphasize certain of these rule systems, and change students' inferential behavior differentially.
4. The instruction can be relatively economical. Perhaps precisely because of the abstract nature of these rules, abstract instruction is effective by itself.
5. Notwithstanding their abstract nature, the rules can be made more accessible by teaching examples of their use, and especially by teaching people how to decode the world in ways that make it more accessible to the rule system.
6. We can do a much better job of teaching these rule systems than we do.

Statistics is taught to most people, in most courses, as if the instructor were determined to prevent its escape from the narrow world of formal data analysis. Examples are restricted if at all possible to IQ tests and agricultural plots. To a greater or lesser extent, the same is true of most of the other pragmatically useful rule systems; when they are taught at all they are taught with little imagination or sense of conviction about their general relevance. I believe that we have only begun to scratch the surface both of the number of pragmatically useful rule systems that can be taught and the means by which they can be most effectively taught.

Finally, Part VII brings home the relevance of the work to the field of cognitive science. In a word, it is a mistake, at this date, to try to found a theory of mental life on mere associations or connections between concretely-defined elements. Organisms make use of rules of some generality; humans at least make use of highly abstract rules that are completely independent of any particular domain of events. Beyond that, new rules are incorporated gracefully into pre-existing systems of rules, something that is difficult to achieve even with most rule-based artificial-intelligence models, especially when, as this book shows to be empirically the case, those rules are inserted from the top rather than "grown" from the bottom. Holland, Holyoak, Nisbett, and Thagard (1986) presented a sketch of a system that gracefully accepts new rules. Somewhat ironically, the systems return to old devices of reinforcement characteristic of behaviorist models in attempting to account for rule modification. Whether this is the best route to go, I don't pretend to know. But any artificial intelligence model that purports to rest on a realistic theory of mind will have to deal with the facts presented in this book: highly general rules exist, and can even be inserted in top-down and highly abstract fashion.

REFERENCES

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391-416.

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences, 4,* 317-331.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. T. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: Bradford Books/The MIT Press.

Jowett, B. (1875). (Ed.) *The dialogues of Plato,* Oxford: Oxford University Press.

Kahneman, D. & Tversky, A. (1972). Subjective probability: a judgment of representativeness. *Cognitive Psychology, 3,* 237-251.

Kelley, H. H. (1972). Causal schemata and the attribution process. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior.* Morristown, NJ: General Learning Press.

Kelley, H. H. (1973). The process of causal attribution. *American Psychologist, 28,* 107-128.

Newell, A. (1980). One final word. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research.* Hillsdale, N.J.: Lawrence Erlbaum.

Russell, B. (1960). *An outline of philosophy.* New York: Meridian.

I ESTABLISHING THE EXISTENCE OF RULES FOR REASONING

2 The Use of Statistical Heuristics in Everyday Inductive Reasoning

Richard E. Nisbett
University of Michigan

David H. Krantz
Bell Laboratories

Christopher Jepson
Ziva Kunda
University of Michigan

It can be argued that inductive reasoning is our most important and ubiquitous problem-solving activity. Concept formation, generalization from instances, and prediction are all examples of inductive reasoning, that is, of passing from particular propositions to more general ones or of passing from particular propositions to other particular propositions via more general ones.

Inductive reasoning, to be correct, must satisfy certain statistical principles. Concepts should be discerned and applied with more confidence when they apply to a narrow range of clearly defined objects than when they apply to a broad range of diverse and loosely defined objects that can be confused with objects to which the concept does not apply. Generalizations should be more confident when they are based on a larger number of instances, when the instances are an unbiased sample, and when the instances in question concern events of low variability rather than high variability. Predictions should be more confident when there is high correlation between the dimensions for which information is available and the dimensions about which the prediction is made, and, failing such a correlation, predictions should rely on the base rate or prior distribution for the events to be predicted.

Because inductive reasoning tasks are so basic, it is disturbing to learn that the heuristics people use in such tasks do not respect the required statistical principles. The seminal work of Kahneman and Tversky has shown that this is so and, also, that people consequently overlook statistical variables such as sample size, correlation, and base rate when they solve inductive reasoning problems. (See surveys by Einhorn & Hogarth, 1981; Hogarth, 1980; Kahneman, Slovic, & Tversky, 1982; Nisbett & Ross, 1980.)

The above research on *nonstatistical heuristics* has been criticized on several grounds. Some critics have maintained that evolution should be expected to produce highly efficacious and generally correct principles of reasoning and that the research may therefore be misleading in some way (Cohen, 1979; Dennett, 1978, 1981, Note 1; Lycan, 1981). Others have maintained that the research does not demonstrate that people fail to apply correct inferential rules but rather that (a) it is the researchers themselves who are mistaken about the correct inferential rules (Cohen, 1981), (b) subjects have been misled by illusionary circumstances of little general significance beyond the laboratory (Cohen, 1981; Lopes, 1982; Dennett, Note 1), or (c) people's general inferential goals are such that at least some violations of statistical principles should be regarded as a form of *satisficing,* or cost-effective inferential shortcuts (Einhorn & Hogarth, 1981; Miller & Cantor, 1982; Nisbett & Ross, 1980).

We offer a different perspective on the incorporation of statistical principles into inductive reasoning, one that rejects the preceding criticisms but is, at the same time, fairly sanguine about people's statistical reasoning. Workers in the Kahneman and Tversky tradition have focused primarily on (a) establishing that people fail to respond to important statistical variables for a wide range of problems and (b) examining the inferential principles that people seem to rely on in solving such problems. There has been no comparable systematic effort to determine whether people do respond to statistical variables, either for problems that are easier than those examined to date or for problems of a different kind than those examined.

If it could be shown that people sometimes do reason using explicitly statistical principles, then the work to date on inductive reasoning, and the criticism of that work, would be cast in a different light. Rather than asking why the failures occur or whether the failures are real, it would seem more fruitful to ask questions such as the following. What factors encourage statistical reasoning and what factors discourage it? For what kinds of events and for what kinds of problems is statistical reasoning most likely to be used? Does purely formal training modify the untutored heuristics of everyday inductive reasoning? In addition, accusations that the work to date rests on a kind of experimental sleight of hand or that people are deliberately and advisedly setting aside statistical principles in favor of quicker and generally satisfactory procedures would seem less plausible. Instead, it would seem more likely that there are just difficulties—surprisingly severe difficulties to be sure, but difficulties merely—in people's use of statistical principles for inductive reasoning.

In this article we first summarize the recent work establishing failures to reason statistically. We then review anecdotal and experimental evidence indicating that people do sometimes reason statistically. Next we present original experimental work indicating some of the factors that influence statistical reasoning. Then we summarize research suggesting that people's ability to reason statistically about everyday life problems is affected by training in formal statistics. Finally, we speculate on the normative implications of people's ability and trainability for statistical reasoning.

STATISTICAL PROBLEMS AND NONSTATISTICAL HEURISTICS

In a succession of studies over the past decade, Kahneman and Tversky have shown that much inductive reasoning is nonstatistical. People often solve inductive problems by use of a variety of intuitive heuristics—rapid and more or less automatic judgmental rules of thumb. These include the representativeness heuristic (Kahneman & Tversky, 1972, 1973), the availability heuristic (Tversky & Kahneman, 1973), the anchoring heuristic (Tversky & Kahneman, 1974), and the simulation heuristic (Kahneman & Tversky, 1982). In problems where these heuristics diverge from the correct statistical approach, people commit serious errors of inference.

The representativeness heuristic is the best studied and probably the most important of the heuristics. People often rely on this heuristic when making likelihood judgments, for example, the likelihood that Object A belongs to Class B or the likelihood that Event A originates from Process B. Use of the heuristic entails basing such judgments on "the degree to which A is representative of B, that is, by the degree to which A resembles B" (Tversky & Kahneman, 1974, p. 1124). In one problem, for example, Kahneman and Tversky (1972) asked subjects whether days with 60% or more male births would be more common at a hospital with 15 births per day, or at a hospital with 45 births per day, or equally common at the two hospitals. Most subjects chose the latter alternative, and the remainder divided about evenly between 15 and 45. The law of large numbers requires that, with a random variable such as sex of infant, deviant sample percentages should be less common as sample size increases. The representativeness heuristic, however, leads subjects to compare the similarities of the two sample proportions to the presumed population proportion (50%); because the two sample proportions equally resemble the population proportion, they are deemed equally likely. The data indicate that, for this problem at least, most subjects used the representativeness heuristic and very few subjects used the law of large numbers.

In another demonstration, Kahneman and Tversky (1973) studied the prediction of an outcome for a target person based on various characteristics of that person or based on scores from various predictor tests. Subjects used the representativeness heuristic: In general, they predicted whichever outcome was most similar to the target person's characteristics or scores.

For instance, in predicting the grade point average (GPA) for a target person who is in the 90th percentile on a predictor test, about the same results are obtained—that is, prediction of a GPA well above average—whether the predictor is the score on a test of sense of humor (which subjects do not regard as very diagnostic of GPA), the score on a test of mental concentration, or the GPA itself (!). Such predictions diverge from those that would be obtained from statistical considerations in which the average accuracy of prediction would be taken into account. Subjects do not seem to realize that if accuracy is very limited, then it is far more probable that the target person's outcome will be equal to the modal outcome (or near the mean of the unimodal symmetric distribution) than that it will take some relatively unusual value that happens to match the characteristics on the predictor. This is the statistical principle of regression to the mean, or base rate.

Other investigations have confirmed and expanded the list of statistical failings documented by Kahneman and Tversky. The failings seem particularly clear and particularly important in people's reasoning about social behavior. Nisbett and Borgida (1975), for example, showed that consensus information, that is, base rate information about the behavior of a sample of people in a given situation, often has little effect on subjects' attributions about the causes of a particular target individual's behavior. When told that most people behaved in the same way as the target, subjects shift little or not at all in the direction of assuming that it was situational forces, rather than the target's personal dispositions or traits, that explain the target's behavior. In a typical experiment, Nisbett and Borgida (1975) told subjects about a study in which participants heard someone (whom the participants believed to be in a nearby room) having what seemed to be an epileptic seizure. Subjects' predictions about whether a particular participant would quickly help the "victim" were unaffected by the knowledge that most participants never helped or helped only after a long delay. Similarly, subjects' causal attributions about the behavior of a participant who never helped the "victim" were unaffected by consensus information. Subjects were just as likely to say that the participant's personality was responsible for his behavior when they knew that most other participants were similarly unhelpful as when they assumed that most other participants helped with alacrity.

Nisbett and Ross (1980) maintained that people fail to apply necessary statistical principles to a very wide range of social judgments. They claimed that people often make overconfident judgments about others based on small and unreliable amounts of information; they are often insensitive to the possibility that their samples of information about people may be highly biased; they are often poor at judging covariation between events of different classes (e.g., "Are redheads hot-tempered?"); and both their causal explanations for social events and their predictions of social outcomes are often little influenced by regression or base rate considerations.

STATISTICAL HEURISTICS

Selective Application of Statistical Reasoning

The foregoing work indicates that nonstatistical heuristics play an important role in inductive reasoning. But it does not establish that other heuristics, based on statistical concepts, are absent from people's judgmental repertoire. And indeed, if one begins to look for cases of good statistical intuitions in everyday problems, it is not hard to find some plausible candidates.

Even when judgments are based on the representativeness heuristic, there may be an underlying stratum of probabilistic thinking. In many of the problems studied by Kahneman and Tversky, people probably conceive of the underlying process as random, but they lack a means of making use of their intuitions about randomness and they fall back on representativeness. In the maternity ward problem, for example, people surely believe that the number of boys born on any particular day is a matter of chance, even though they rely on representativeness to generate their subjective sampling distributions. But consider the following thought experiment: If someone says, "I can't understand it; I have nine grandchildren and all of them are boys," the statement sounds quite sensible. The hearer is likely to agree that a causal explanation seems to be called for. On the other hand, imagine that the speaker says, "I can't understand it; I have three grandchildren and all of them are boys." Such a statement sounds peculiar, to say the least, because it seems transparent that such a result could be due just to chance—that is, there is nothing to understand. Such an intuition is properly regarded as statistical in our view.

The contrast between the statistical intuition in our anecdote and subjects' use of the representativeness heuristic in the maternity ward problem illustrates the selectivity with which people apply statistical concepts. The failure to do so in the maternity ward problem may be due to the use of "60%" in the problem, which evokes comparison between 60% and 50% and thence the dependence on the similarity judgment in choosing an answer. It may also be due to lack of concrete experience in thinking about samples in the range 15–45. As Piaget and Inhelder (1951/1975) put it, people seem to have an intuitive grasp of the "law of *small* large numbers," even though they may not generalize the intuition to large numbers.

People also seem to have an ability to use base rates for selected kinds of problems. Consider the concepts of *easy* and *difficult* examinations. People do not infer that a student is brilliant who received an A+ on an exam in which no one scored below A− nor that the student is in trouble who flunked a test that was also failed by 75% of the class. Rather, they convert the base rate information (performance of the class as a whole) into a location parameter for the examination (easy, . . ., difficult) and make their inference about the particular student in terms of the student's relative position compared to the difficulty of the exam. Indeed, laboratory evidence has been available for some time that base rates are readily utilized for causal attributions for many kinds of abilities and achievements (Weiner et al. 1972).

As Nisbett and Ross (1980) suggested, one suspects that many lay concepts and maxims reflect an appreciation of statistical principles. It seems possible, for example, that people sometimes overcome sample bias by applying proverbs such as "Don't judge a book by its cover" or "All that glitters is not gold." Perhaps people sometimes even manage to be regressive in everyday predictions by using concepts such as "beginner's luck" or "nowhere to go but up/down."

There is one inductive reasoning task in particular for which there is good reason to suspect that statistical intuitions are very frequently applied. This is *generalization from instances*—perhaps the simplest and most pervasive of everyday inductive tasks. People surely recognize, in many contexts at least, that when moving from particular observations to general propositions, more evidence is better than less. The preference for more evidence seems well understood as being due to an intuitive appreciation of the law of large numbers. For example, we think that most people would prefer to hold a 20-minute interview rather than a 5-minute interview with a prospective employee and that if questioned they would justify this preference by saying that 5 minutes is too short a period to get an accurate idea of what the job candidate is like. That is, they believe that there is a greater chance of substantial error with the smaller sample. Similarly, most people would believe the result of a survey of 100 people more than they would believe that of a survey of 10 people; again, their reason would be based on the law of large numbers.

As we shall see, there is reason to believe that people's statistical understanding of the generalization task is deeper still. People understand, at least in some contexts, that the law of large numbers must be taken into account to the degree that the events in question are uncertain and variable in a statistical sense. Thus they realize that some classes of events are very heterogeneous; that is, the events differ from one another, or from one occasion to another, in ways that are unpredictable, and it is these classes of events for which a large sample is particularly essential.

Randomizing Devices and the Ontogeny of Statistical Reasoning

Where do people's selective statistical intuitions come from? An extremely important series of studies by Piaget and Inhelder (1951/1975) suggests that the intuitions may arise in part from people's understanding of the behavior of random generating devices. Statistical reasoning is of course very commonly applied in our culture to the behavior of such mechanisms. Piaget and Inhelder showed that statistical intuitions about random devices develop at an early age. They conducted experiments in which children were shown various random generating devices and then were asked questions about them. The devices included different-colored marbles on a tilt board, coin tosses, card draws, a spinner, and balls dropped through a funnel into a box with a varying number of slots. Children were shown the operation of these devices and then were asked to predict outcomes of the next operation or set of operations and to explain why particular outcomes had occurred or could or could not occur. The work showed that even children less than 10 years old used the concept of chance and understood the importance of sequences of repeated trials.

In one study, for example, Piaget and Inhelder (1951/1975) spun a pointer that could stop on one of eight different-colored locations. The young children they studied (in general, those less than 7 years old) did not initially recognize their complete inability to predict the pointer's stopping place.

> He knows quite well that he is not likely to be able to predict the color on which the bar will stop, but he does believe in the legitimacy of such a prediction and tries to guess the result. . . . The child oscillates quickly between two solutions. . . . Either the bar will have the tendency to come back to a color on which it has already stopped, or it will, on the contrary, stop on the colors not yet touched (p. 61).

At this stage the children did not recognize the equivalent chances of the various stopping places, and when the pointer was made to stop at one color repeatedly (by using a magnet) they found nothing unusual in this. A satisfactory causal explanation usually was forthcoming: for example, "the pointer got tired."

By around the age of 7, the Piaget and Inhelder subjects began to understand the chance nature of the pointer's behavior. After a few demonstrations, they quickly came to doubt the predictability of single trials and came to see the distribution of possibilities and their equivalence. Between the ages of 7 and 10, their subjects came to understand the importance of repeated trials and long run outcomes.

E: If I spun it ten or twenty times, could there be one color at which it never stopped?
S: (age 7): Yes, that could happen. That would happen more often if we did it only ten times rather than twenty (p. 75).
E: Will it hit all the colors or not?
S: (age 10 years, 7 months): It depends on how long we spin it.
E: Why?
S: Because if we spin it often, it will have more chances of going everywhere (p. 89).

How does the child come to have an understanding of the concept of chance during this period? Piaget and Inhelder argue that *the child's understanding of uncertainty grows out of the child's understanding of physical causality.* To the very young child with little understanding of the causal mechanisms that produce outcomes in a physical system, every outcome is a "miracle"—that is, unanticipated—and, paradoxically, once the outcome has occurred, the child believes that it can be explained. As the child comes to understand, in terms of concrete operations, the causal mechanisms that produce outcomes, the child begins to recognize which sorts of outcomes are predictable (and explainable) and which are not. The outcomes that are not predictable are gradually understood to obey certain *non-causal rules.* In particular, the child comes to recognize some cases of the law of large numbers, for example, that the likelihood of any given outcome occurring is greater with a large number of trials than with a smaller number.

By the age of 11 or so, many children have—in addition to a clear conception both of fully deterministic systems and of random generating devices—a good understanding of non-uniform probability distributions. These are partially random systems in which causal factors are at work making some of the possible outcomes more likely than others. The child comes to learn that even though individual events are uncertain in such a system, aggregate events may be highly predictable. In such a probabilistic system, the child grasps the relevance to prediction of the base rate, that is, the distribution and relative frequency of the various outcomes.

This latter point is well illustrated by children's understanding of a device that allows balls to be dropped through a hole into one of a number of slots or bins beneath. Here the chances of a ball dropping into one slot versus another can be made quite unequal by the physical set-up. It is easy to build the device, for example, so that most balls drop into middle bins and fewer drop in the side bins, generating a crude bell curve. Children under 7 generally fail to use this distribution as a basis of prediction. Although they slowly come to recognize that central positions will collect more balls than peripheral ones, they cannot generalize this fact from a box with a particular number of slots to another box with a different number; they do not expect symmetry between slots that are equidistant from the center; and they do not recognize the role of the law of large numbers in making the central slots particularly favored over a long series of trials. All of these intuitions, in contrast, come easily to many 12-year-olds.

We may speculate that the older child's statistical conceptualization of the behavior of randomizing devices serves as the basis for a similar conceptualization of other kinds of events that may be seen as variable and uncertain. We discuss later just what characterizes events where an analogy to randomizing devices can be seen versus those where it cannot be seen.

The Intellectual History of Statistical Reasoning

The cultural history of statistical reasoning appears to parallel in some interesting respects the developmental course described by Piaget and Inhelder (1951/1975). This history has been traced by Hacking in his book *The Emergence of Probability* (1975). Hacking points out that although random generating devices have been used at least since Biblical times, the modern concept of probability was invented rather suddenly in the 17th century. This was true despite the popularity of games of chance in antiquity and the existence of sophisticated mathematics. (Hacking notes that someone with only a modest knowledge of modern probability could have won all Gaul in a week!)

Paradoxically, the major factor underlying the sudden emergence of the modern concept of probability was the change to a deterministic understanding of the physical world. In the Renaissance, the task of science was understood not primarily as a search for the causal factors influencing events but as a search for signs as to the meaning of events. These signs were clues and portents strewn about by the benign Author of the Universe. This sort of understanding of events encouraged a heavy reliance on the representativeness heuristic. The Renaissance physician, for example, adhered to the *doctrine of signatures.* This was the "belief that every natural substance which possesses any medicinal virtue indicates by an obvious and well-marked external character the disease for which it is a remedy, or the object for which it should be employed" (John Paris, cited in Mill, 1843/1974, p. 766). The representativeness heuristic thus could be derived as a rule of inference from the principle that the Author of the Universe wanted to be helpful in our attempts to understand the world.

A quite different way of understanding events became predominant in the 17th century. This was a new "mechanistic attitude toward causation" (Hacking, 1975, p. 3). Just as the development of concrete operations helps the child to recognize the irreducible ignorance and uncertainty that is left as a residue after causal analysis of a randomizing device, so the new attitude toward causation helped 17th century scientists appreciate the nature of uncertainty in probabilistic systems. "Far from the 'mechanical' determinism precluding an investigation of chance, it was its accompaniment . . . this specific mode of determinism is essential to the formation of concepts of chance and probability" (Hacking, 1975, p. 3).

Summary

In short, there is good reason to believe that people possess *statistical heuristics* – intuitive, rule-of-thumb inferential procedures that resemble formal statistical procedures. People apply these heuristics to the behavior of random generating devices at a fairly early age. The formal understanding of statistical principles – that is, of the rules governing the behavior of randomizing devices – increases at least until adolescence. The use of such heuristics, both individually and culturally, seems related to the growth of causal understanding of the physical world and to attempts to extend this causal understanding, by analogy, to wider domains. Although we know little at present of the growth in the child's or adolescent's ability to apply statistical heuristics to events other than those produced by randomizing devices, it seems clear that such growth does take place. Adults who are untutored in formal statistics seem to reason statistically about a number of events other than those produced by randomizing machines – such as performance on tests, sports, weather, and accident and death risks. In addition, it is hard to imagine that people could conduct the most basic of inferential tasks, namely, generalization from instances, without the application of at least a rudimentary version of a law-of-large-numbers heuristic.

FACTORS THAT AFFECT STATISTICAL REASONING

Despite ontogenetic and historical growth in the ability to reason statistically, contemporary adults do not reason statistically about a wide range of problems and event domains that require such reasoning, and they often do not do so even if they have substantial training in formal statistics (Tversky & Kahneman, 1971). Why is this? What factors make it difficult to apply statistical heuristics when these are required, and what factors can make it easier? Three factors that seem important are implicit in the preceding discussion.

Clarity of the Sample Space and the Sampling Process

Randomizing devices are usually designed so that the sample space for a single trial is obvious and so that the repeatability of trials is salient. The die has six faces and can be tossed again and again; the pointer can stop on any of eight sectors and can be spun over and over. Clarity of sample space makes it easier to see what knowledge is relevant. For randomizing devices, the most relevant knowledge is often just the observation of symmetry of the different die faces, spinner sectors, and so forth. The salience of repeatability makes it easier to conceptualize one's observations as a sample.

In the social domain, sample spaces are often obscure, and repeatability is hard to imagine. For example, the sample space consisting of different degrees of helpfulness that might be displayed by a particular person in a particular situation is quite obscure, and the notion of repetition is strained. What is it that could be repeated? Placing the same person in *different* situations? Or *other* people in the same situation? The probability that Person P will exhibit Behavior B in Situation S is abstract and not part of the inductive repertoire of most people most of the time. Even though people recognize the possibility of errors in their judgments of social situations, they do not try to construct probability models; rather, they rely on the representativeness heuristic.

Recognition of the Operation of Chance Factors

A second major factor encouraging the use of statistical heuristics is the recognition of the role of chance in producing events in a given domain or in a particular situation. We have already seen how Piaget and Inhelder (1951/1975) describe the recognition of chance in the operation of randomizing devices. The child comes to recognize the limitations of causal analysis for a spinner and the consequent residual uncertainty about the production of events. Something like the same transparent indeterminism exists for other sorts of events as well, even those involving human beings. For example, statistical understanding of some types of sports is undoubtedly facilitated by the manifestly random component in the movement of the objects employed: "A football can take funny bounces." The random component probably does not have to be physical in order for people to recognize it. It is possible to recognize the unpredictability of academic test performance by repeated observations of one's own outcomes. Even with one's own efforts and the group against which one is competing held constant, outcomes can vary. One may even recognize that one's performance on particular occasions was particularly good or poor because of accidents: "I just happened to reread that section because Jill never called me back"; "It was very noisy in the study area that night so I didn't get a chance to review my notes."

In contrast, cues as to randomness in the production of events are much subtler for other kinds of events, especially for many social ones. When we interview someone, what signs would let us know that a particular topic got explored just by chance or that the person seems dour and lackluster because of an uncharacteristic attempt to appear dignified rather than because of a phlegmatic disposition? In addition, as Einhorn and Hogarth (1978) have pointed out, the gatekeeping function of the interview may serve to prevent us from recognizing the error variance in our judgments: The great talent of some people not hired or admitted may never be observed. Daniel Kahneman (Note 2) has suggested to us that the "interview illusion" exists in part because we expect that brief encounters with a living, breathing person ought to provide a "hologram" of that person rather than merely a sample of the person's attributes and behaviors. In most situations, cues as to the fact that an interview ought to be regarded as a sample from a population, rather than a portrait in miniature, are missing. The same may be true for visits to a city, country, or university. One of us long believed that reports of raininess in England were greatly exaggerated because he once stayed in London for 10 days and it only drizzled twice!

Cultural Prescriptions

A third factor that may contribute to the use of statistical heuristics is a cultural or subcultural prescription to reason statistically about events of a given kind. Although Piaget and Inhelder focused on developmental changes in the ability to reason statistically about randomizing devices, from a historical perspective it is the young child's ability to reason statistically at all about such devices that is remarkable. It seems implausible that a medieval European child would have reasoned in such a sophisticated way as the Piaget and Inhelder subjects. Statistical reasoning is the culturally prescribed way to think about randomizing devices in our culture, and this general approach undoubtedly trickles down to children. Similarly, statistical reasoning has become (or is becoming) the norm for experts in many fields—from insurance to medical diagnosis—and is rapidly becoming normative for the lay novice as well in such domains as sports and the weather. Models of statistical reasoning abound for sports in particular, as the two examples below indicate.

> Baseball's law of averages is nothing more than an acknowledgement that players level off from season to season to their true ability—reflected by their lifetime averages. A .250-hitter may hit .200 or .300 over a given period of time but baseball history shows he will eventually level off at his own ability ("Law of Averages," 1981).

> The musky tends to be a deep water fish. Most fishing success is in shallow water, but . . . this misleading statistic [is probably accounted for in part by the fact that] sheer statistical chance dictates that fish will come from the waters receiving the most man hours of fishing pressure. Shallow water fishing for muskies is very popular, and very few fishermen work them deep (Hamer, 1981).

The statistical spirit embodied in these quotations reaches many fans. Thus, it is commonplace to hear lay people endorse the proposition that "On a given Sunday any team in the NFL can beat any other team." (Compare with "On a given Sunday, any parishioner's altruism can exceed that of any other parishioner"!)

In our view, these three factors—clarity of the sample space and the sampling process, recognition of the role of chance in producing events, and cultural prescriptions to think statistically—operate individually and, perhaps more often, together to increase people's tendencies to apply statistical heuristics to problems that require a statistical approach. If these factors are genuinely important determinants of people's ability to reason statistically, then it should be possible to find support for the following predictions.

In cases where the sample space is clear and the possibility of repetition is salient, people will respond appropriately to statistical variables. In particular, in the task of generalizing from instances, where the sample space is a clear dichotomy and the sampling process is just the observation of more members of a clearly defined population, (a) people will generalize more cautiously when the sample size is small and when they have no strong prior belief that the sampled population is homogeneous, and (b) people can be influenced to generalize more or less readily by manipulations that emphasize the homogeneity or heterogeneity of the sampled population.

The following predictions should hold both for generalization and for other, more complex, inferential tasks: (a) Manipulations designed to encourage recognition of the chance factors influencing events should serve to increase statistical reasoning. (b) People who are highly knowledgeable about events of a given kind should be more inclined than less knowledgeable people to apply statistical reasoning to the events—because both the distributions of the events and the chance factors influencing the events should be clearer to such people. (c) People should be disinclined to reason statistically about certain kinds of events that they recognize to be highly variable and uncertain—notably social events—because the sample spaces for the events and the chance factors influencing the events are opaque. (d) Training in statistics should promote statistical reasoning even about mundane events of everyday life because such training should help people to construct distributional models for events and help them to recognize "error," or the chance factors influencing events.

Generalizing From Instances

Generalization from observed cases is the classic concern of philosophers and other thinkers who are interested in induction. A number of instances of Class A are observed, and each of them turns out to have Property B. Possible inferences include the universal generalization *all A's have B,* or the near universal *most A's have B,* or at least the relinquishing of the contrary generalization, namely, *most A's do not have B.*

The untrammeled employment of the representativeness heuristic would lead people to make the above inferences from quite small numbers of instances, and, indeed, this is often found, both anecdotally and in laboratory studies (Nisbett & Ross, 1980, pp. 77–82). On the other hand, philosophers since Hume have puzzled about how these generalizations can be logically justified, even when very large numbers of instances are observed. The puzzle has been compounded by the fact that sometimes it seems correct to generalize confidently from a few instances. Hume (1748/1955) wrote, "[Often, when] I have found that . . . an object has always been attended with . . . an effect . . . I foresee that other objects which are in appearance similar will be attended with similar effects" (p. 48). The problem is that only sometimes do we draw such a conclusion with confidence. "Nothing so like as eggs, yet no one, on account of this appearing similarity, expects the same taste and relish in all of them" (p. 50). Mill (1843/1974), a century later, phrased the problem like this: "Why is a single instance, in some cases, sufficient for a complete induction, while in others myriads of concurring instances, without a single exception known or presumed, go such a very little way towards establishing a universal proposition?" (p. 314).

The statistical advances since Mill's time make it clear that a large part of the answer to his question has to do with beliefs about the variability or *homogeneity* of certain kinds or classes of events (cf. Thagard & Nisbett, 1982). Generalization from a large sample is justified in terms of one's beliefs that the sampling itself is homogeneous (i.e., that the distribution of possible sample statistics is the same as would be predicted by random sampling). And generalization from a small sample or resistance to generalization, even from a large sample, are justified in terms of prior beliefs about the homogeneity or heterogeneity of objects or events of a certain kind with respect to a property of a certain kind. If, for example, the object is one of the chemical elements and the property is electrical conductivity, then one expects homogeneity: All samples of the element conduct electricity or none do. But if the object is an animal and the property is blueness, one's prior belief does not favor homogeneity so strongly; color may or may not vary within a particular species.

In other words, there are cases where use of the representativeness heuristic is justified in terms of beliefs about homogeneity, which in turn may be soundly based on individually or culturally acquired experience with kinds of objects and kinds of properties. For other cases, simple representativeness cannot be justified, and there are indeed cases, as Mill claimed, in which a strong prior belief in *heterogeneity* properly prevents acceptance of a generalization even after quite large numbers of instances have been observed.

We attempted to demonstrate, in a laboratory study of judgment, that people do in fact temper the use of representativeness to a greater or lesser degree depending on beliefs about the variability of a kind of object with respect to a kind of property.

Study 1: Beliefs About Homogeneity and Reliance on the Law of Large Numbers

In this study, we simply guessed at the prevailing beliefs about homogeneity. We tried to obtain different degrees of heterogeneity by using conductivity of metals, colors of animals, and so on. Subjects were told of one instance or of several instances of a sampled object having a particular property and were asked to guess what percentage of the population of all such objects would have the property. The sample sizes used were 1, 3, or 20; in the latter cases, all 3 or all 20 of the objects had the property in question. We anticipated that subjects would generalize more readily from a given number of instances when the kind of object was perceived as homogeneous with respect to the kind of property than when the kind of object was perceived as heterogeneous with respect to the kind of property.

Method

Subjects were 46 University of Michigan students of both sexes who were enrolled in introductory psychology. (As sex did not affect any of the dependent variables in this or any of the other studies, it will not be discussed further.) Eighty-five percent of the subjects had taken no statistics courses in college. The questionnaire was presented as one of several in a study on judgment. It read as follows for the $N = 1$ condition:

> Imagine that you are an explorer who has landed on a little known island in the Southeastern Pacific. You encounter several new animals, people, and objects. You observe the properties of your "samples" and you need to make guesses about how common these properties would be in other animals, people or objects of the same type.
>
> Suppose you encounter a new bird, the shreeble. It is blue in color. What percent of all shreebles on the island do you expect to be blue?

(This and the subsequent questions were followed by

> "___ percent. Why did you guess this percent?")
>
> Suppose the shreeble you encounter is found to nest in a eucalyptus tree, a type of tree which is fairly common on the island. What percent of all shreebles on the island do you expect to nest in eucalyptus trees?

> Suppose you encounter a native, who is a member of a tribe he calls the Barratos. He is obese. What percent of the male Barratos do you expect to be obese?
>
> Suppose the Barratos man is brown in color. What percent of male Barratos do you expect to be brown (as opposed to red, yellow, black or white)?
>
> Suppose you encounter what the physicist on your expedition describes as an extremely rare element called floridium. Upon being heated to a very high temperature, it burns with a green flame. What percent of all samples of floridium found on the island do you expect to burn with a green flame?
>
> Suppose the sample of floridium, when drawn into a filament, is found to conduct electricity. What percent of all samples of floridium found on the island do you expect to conduct electricity?

The questionnaires for the $N = 3$ condition and the $N = 20$ condition were identical except that they specified larger samples of each object. For example, the first shreeble item for the $N = 3$ condition read as follows:

> Suppose you encounter a new bird, the shreeble. You see three such birds. They are all blue in color. What percent of all shreebles on the island do you expect to be blue?

The reasons subjects gave for guessing as they did were coded as to their content. There were three basic sorts of answers: (a) references to the homogeneity of the kind of object with respect to the kind of property, (b) references to the heterogeneity of the kind of object with respect to the kind of property—due to the different properties of subkinds (e.g., male vs. female), to some causal mechanism producing different properties (e.g., genetic mistakes), or to purely statistical variability (e.g., "where birds nest is sometimes just a matter of chance"), and (c) other sorts of answers that were mostly based on representativeness or that were mere tautologies. Two independent coders achieved 89% exact agreement on coding category.

Results

Any one element is presumed by scientists to be homogeneous with respect to most properties. At the other extreme, most human groups are highly heterogenous among themselves in many attributes, including body weight. If educated lay people share these beliefs and if they reason statistically, then (a) they should exercise more caution in generalizing from single cases when heterogeneity is expected than when homogeneity is expected and (b) large N should be important primarily in the case of populations whom subjects believe to be heterogeneous with respect to the property in question.

Figure 2.1 presents subjects' estimates of the percentage of each population having the property associated with the sample as a function of sample size presented. It may be seen that subjects are quite willing to generalize from even a single instance of green-burning or electricity-conducting floridium and also from a single, brown, Barratos tribesman. The modal estimate for $N = 1$ (as well as for $N = 3$ and $N = 20$) in all of these cases is 100%. In contrast, generalizations are less extreme for even 20 instances of blue shreebles or eucalyptus-nesting shreebles or 20 obese Barratos. The $t(31)$ contrasting $N = 1$ for floridium attributes and Barratos color with $N = 20$ for shreeble attributes and Barratos obesity is 3.00; $p < .01$.[1]

Subjects' explanations for their estimates fully justify this pattern of inferences. It may be seen in Table 2.1 that subjects reported believing that elements are homogeneous with respect to color and conductivity and that tribes are homogeneous with respect to color. In contrast, subjects rarely expressed the belief that there is homogeneity for the other kinds of populations and properties and instead expressed belief in heterogeneity of

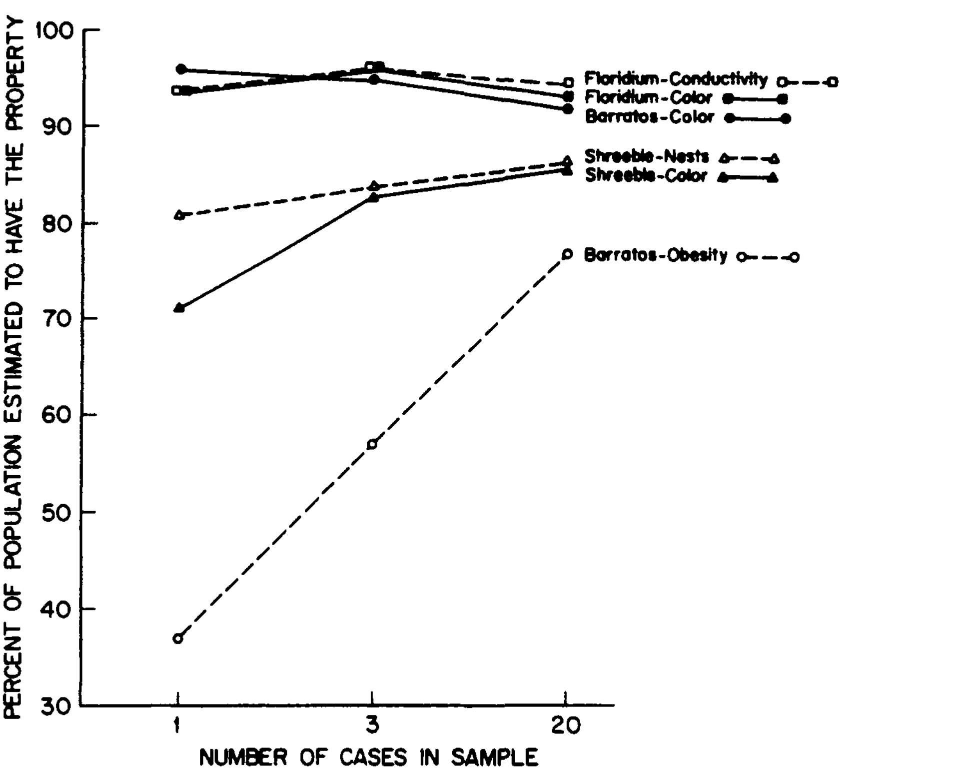

FIG. 2.1 Percentage of each population estimated to have the sample property as a function of number of cases in the sample.

TABLE 2.1
Number of Subjects Giving Each Type of Reason and Percentage of Population Estimated to Have the Property

	Reason					
	Homogeneity		*Tautology*		*Heterogeneity*	
Object and property	*n*	%	*n*	%	*n*	%
Shreeble						
Color	6	95	17	83	22	75
Nests	8	96	19	84	19	78
Barratos						
Obesity	5	79	10	62	31	53
Color	31	98	7	94	8	80
Floridium						
Color	31	97	9	91	6	82
Conductivity	31	98	7	92	8	82

one sort or another for these objects and properties. Figure 2.1 shows that it is only for these latter cases that subjects reasoned statistically in the sense that they were more willing to assume that the population resembles the sample when N is larger. N affects the estimates of the obesity of Barratos and the color of shreebles ($p < .001$ and $p = .11$, respectively). In addition, a total of 10 subjects complained on one or more problems that the N was too small to give a good estimate. For nine of these subjects, the complaints were about one or more of the three problems where populations were presumed to be heterogeneous with respect to the property in question, whereas for only one subject was the complaint about a problem for which subjects in general believed populations to be homogeneous with respect to properties (exact $p = .02$).

Finally, an internal analysis of the Table 2.1 data for each question showed that those subjects who believed the population to be homogeneous with respect to the property estimated that a higher percentage of the population was like the sample than did those subjects who believed the population was heterogeneous with respect to the property. The lowest t resulting from the six comparisons yielded $p < .05$.

Study 2: Manipulating the Salience of Distribution Parameters

Study 1 established that people can apply statistical reasoning to one of the most basic of inferential tasks. It also established that beliefs about variability of the class of events in question can mediate the statistical reasoning. One other study in the literature made similar points. Quattrone and Jones (1980) proposed a version of the present view that beliefs about variability influence inductive generalizations in their important study on perception of ingroups versus outgroups. They hypothesized that "an observer's tendency to generalize from the behavior of a specific group member to the group as a whole is proportional to the observer's perception of the group's homogeneity" (p. 141). Because people are more familiar with the members of groups to which they happen to belong, they will recognize "the group's general variability, the extent to which its members . . . differ from one another when viewed over all dimensions" (p. 141). Because people are less familiar with outgroups, they are at liberty to assume that their members are relatively uniform. Thus people may generalize more readily from observations of the behavior of outgroup members than from observations of the behavior of ingroup members.

To test this hypothesis, Quattrone and Jones (1980) showed Princeton and Rutgers University undergraduates videotapes of male students who were allegedly serving as participants in psychology experiments. These students were asked to make choices such as to wait for a few minutes by themselves versus in the company of others or to listen to rock music versus classical music. Half of the subjects at each campus believed they were viewing Princeton men, and half believed they were viewing Rutgers men. After seeing the choice of one participant, subjects were asked to predict what the 100 participants in the study did. Quattrone and Jones found greater generalization from the participants' behavior to outgroup members than to ingroup members. Thus, Princeton subjects generalized more strongly to the behavior of the Rutgers population after observing the choice of the "Rutgers" participant than they did to the Princeton population after observing the choice of the "Princeton" participant.

If, as both we and Quattrone and Jones assume, generalizations about groups from the behavior of its members are mediated by assumptions about variability of group members, then it should be possible to manipulate those assumptions and therefore to influence the degree of generalization. People are inclined to think of (their own) university populations as being immensely variable—what with caftans here and exotic accents there, football players here and budding physicists there. In fact, however, university populations are not as heterogeneous as one might casually presume. Most students, even at multiversities, are, after all, bright young middle-class people of fairly homogeneous geographic and ethnic backgrounds. It seems possible that, if subjects were required to contemplate the central tendencies of their university populations before observing choice behavior like that presented to Quattrone and Jones's subjects, they might generalize more. This possibility was examined in Study 2.

Method

The procedure used by Quattrone and Jones (1980) was followed almost exactly, except that subjects were told that the videotapes were either of

University of Michigan or of Ohio State University students, and half of the subjects were exposed to a central-tendency manipulation before viewing the videotapes. Subjects were 115 University of Michigan undergraduates of both sexes enrolled in introductory psychology. They participated in small groups, seated around a table facing a .53-m (21-inch) video monitor. Subjects were told that the investigators were "studying how people make judgments about people—working from actual information they have about people to guesses about other aspects of people. One of our major interests is in how students perceive students at (their own/another) university."

At this point the central-tendency manipulation was delivered to experimental subjects, who were told that "we will be asking you several questions about students at (the University of Michigan/Ohio State University)" and were given the appropriate central-tendency questionnaire. Control subjects began viewing videotapes immediately.

The central-tendency questionnaire consisted of three questions that we expected would influence subjects' conceptions of the variability of a student population. Subjects were asked to "please list what you would guess to be the 10 most common majors at (the University of Michigan/Ohio State University)" and next to list the five most common ethnic group backgrounds and the five most common religious backgrounds at that university. Answering these questions might be expected to prompt subjects to recognize that the student body is not all that heterogeneous: Most students are, after all, white Protestants concentrated in a limited number of relatively popular majors.

Subjects viewed the Quattrone and Jones videotapes.[2] They were introduced as having been made during psychology experiments conducted at the University of Michigan or at Ohio State University. In each of the three tapes a male participant was shown being confronted with a decision, and he then chose one of two alternative behaviors offered. In the first scenario, a target person had to choose between waiting alone or waiting with other subjects while his experimenter fixed a machine. In the second scenario, the choice was between listening to classical music or listening to rock music during an experiment on auditory perceptual sensitivity. In the third scenario, the choice was between solving mathematical problems or solving verbal problems during an experiment on the effects of noise on intellectual performance. As the order in which scenarios were presented had no effect in the Quattrone and Jones study, it was held constant in our study.

The procedure was the same for each scenario. Subjects watched the target person being given instructions and being asked to make his decision. At this point the tape was turned off and subjects were asked to predict the target person's decision on a 21-point scale that had endpoints labeled with the two relevant options. The tape was then turned on again and subjects observed the participant make his decision. Half of the subjects saw the participants in the three scenarios make one set of decisions, and half saw the complementary set. Thus, subjects in Set A saw the target persons choose (a) to wait alone, (b) to listen to classical music, and (c) to solve mathematical problems. Subjects in Set B saw targets choose (a) to wait with others, (b) to listen to rock music, and (c) to solve verbal problems.

The dependent variable of interest consisted of the subjects' estimates of how many out of 100 participants in each of the three experiments chose each of the two options. (For the sole purpose of replication, subjects were also asked to indicate what they would have done and who they liked as people more—those who would prefer Option A or those who would prefer Option B.)

Results

Figure 2.2 presents subjects' generalizations about the University of Michigan and Ohio State University populations for control subjects and for subjects exposed to the central-tendency manipulation. *Generalization* is defined as the difference between population estimates for subjects presented with Set A choices versus those for subjects presented with Set B choices. The higher this index is, the more a group of subjects was influenced in their estimates by the behavior of the particular subject they witnessed. The index sums across all three types of choices, but the trends were the same for each of the three problems.

The difference between the control groups exposed to Ohio State

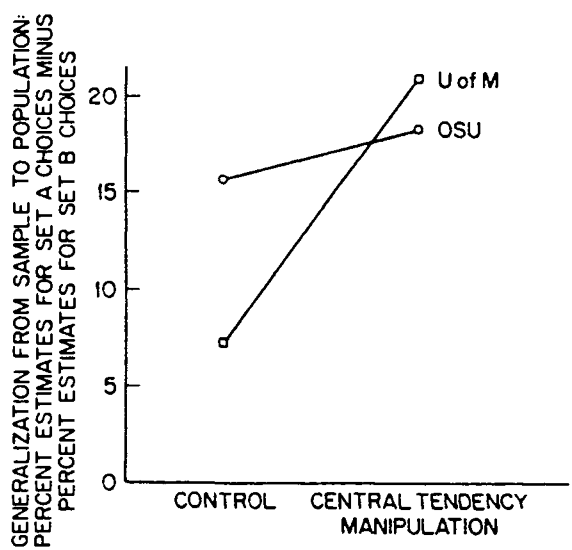

FIG. 2.2 Generalization from sample to population as a function of campus population and central-tendency manipulation. (U of M = University of Michigan; OSU = Ohio State University.)

University participants versus those exposed to University of Michigan participants provides a replication of the Quattrone and Jones finding. The magnitude of the difference is very similar to that found by them, though for our smaller sample it is only marginally significant, $F(1, 50) = 2.76$, $.05 < p < .10$).

The effect of the central-tendency manipulation was to increase the degree of generalization from the sample, $F(1, 107) = 4.23$, $p < .05$). It may be seen that the effect was largely due to the behavior of the University of Michigan group. This is not surprising because the judgments about the Ohio State students may have already incorporated central tendencies in the form of an outgroup stereotype. This explanation should be viewed with caution, however, inasmuch as the interaction failed to reach statistical significance.

Both findings provide support for the contention that concurrent representations of population variability mediate inductive generalizations. Familiarity with one's own group results in less willingness to generalize for them than for another group, although forced contemplation of central tendencies results in more willingness to generalize, at least for the familiar ingroup.

One other study, by Silka (1981), shows the importance for inductive reasoning of people's focus on variability versus central tendency. She asked subjects to examine a series of numerical values that were said to represent the mental health of several individuals. Some subjects were asked to remember the average of the values, and some were asked to remember the range. When subjects were asked, 1 week later, to assess the degree of change represented by a new value, subjects who had been asked to remember the average were more likely to infer that there had been a genuine change than those who had been asked to remember the range. The implication of Silka's finding, together with those of Study 2, is that inferences about continuity and change, and inductive reasoning generally, may be in part a function of arbitrary encoding and retrieval factors that accidentally emphasize either the homogeneity or the heterogeneity of events.

Study 3: Manipulating the Salience of Chance Factors

Study 2 establishes that manipulations of the salience of distributional parameters can influence subsequent generalizations. It should also be possible to influence generalizations by manipulating the salience of chance factors. One potentially interesting way of doing this would be to highlight for subjects the degree to which evidence about an object should properly be regarded as a sample from the population of the object's attributes. Such a reminder ought to prompt subjects to reason more statistically, deemphasizing evidence from smaller samples and placing greater weight on evidence from larger samples.

Borgida and Nisbett (1977) argued that people often ignore the judgments of others when choosing between two objects and substitute their own initial impressions of the objects as the sole basis of choice. People do this in part because they do not recognize the relevance of the law of large numbers when reasoning about events of the personal preference kind. When the objects are multifaceted and complex, however, the law of large numbers is applicable in two ways: (a) The reactions of other people to the object, especially if they are based on more extensive contact with the object than one has had oneself, generally should be a useful guide to choice (though, of course, it is possible to construct cases where other people's reactions would not be useful). (b) One's own experience with the object, especially if it is brief or superficial, may be a poor guide to choice because of the error that plagues any small samples, even those that happen to be our own.

It seemed likely that if people were made explicitly aware of the role of chance in determining the impression one may get from a small sample, they might place less faith in a small personal sample and more faith in a large sample based on other people's reactions.

Method

Subjects were 157 University of Michigan students of both sexes who were enrolled in introductory psychology classes. Eighty-seven percent had taken no statistics courses in college. Subjects participated in small groups. They were presented with two versions of the following problem.

> 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. 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 similar to his. The 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 smaller 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