

CHAPTER 1:

Thinking Critically about Intelligence

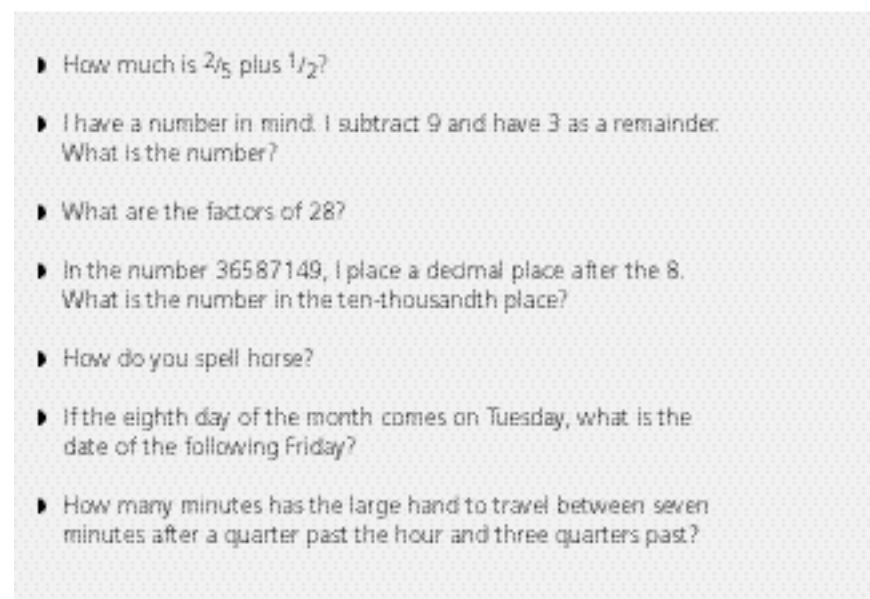
Mr. von Osten's Horse

For many years, people have wondered if animals other than humans are intelligent. Some, like the philosopher Rene Descartes, believed that animals are automata. That is, that they are biological machines that move about but have no intelligence--rather like mechanical wind-up toys. Others believe that some animals, often their own pets, have at least a degree of intelligence. Stories of animals that perform amazing feats appearing to require intelligence have caught the popular imagination. Sir William Kepler's bulldog became famous because it solved arithmetic problems by barking an appropriate number of times to indicate the answer. There was the reading pig of London. And Lady, the talking horse of Richmond could not only count and spell but could also locate missing objects and give financial advice. While enthusiasts saw evidence of remarkable animal intelligence in these cases, skeptics looked for other explanations, such as outright fraud or unintentional cueing of the animals to give the desired answer.

No case aroused more international attention than that of Clever Hans, Mr. Von Osten's horse. Hans could reliably perform a remarkable variety of tasks including arithmetic computations, reading, and spelling. He could answer questions about the calendar, the clock, and the musical scale. He answered the questions by tapping his foot. For answers requiring an integer, he tapped the appropriate number of times. For answers requiring fractions, he tapped first the

numerator and then the denominator. When the answer was a large number, Hans tapped rapidly but if the answer was a small number, he tapped more deliberately. For spelling, he used a two dimensional table devised by Mr. Von Osten. By tapping a number for the column and a number for the row, Hans was able to specify letters of the alphabet. Examples of the questions he was observed to answer correctly are shown in Table 1.1. In addition to answering questions such as these, Hans was able to pick out a cloth of a specified color from an array of cloths and even tell what note to drop from a dissonant cord to make it consonant.

Table 1.1 Questions asked of Mr. von Osten's horse.



▶ How much is $\frac{2}{5}$ plus $\frac{1}{2}$?
▶ I have a number in mind. I subtract 9 and have 3 as a remainder. What is the number?
▶ What are the factors of 28?
▶ In the number 36587149, I place a decimal place after the 8. What is the number in the ten-thousandth place?
▶ How do you spell horse?
▶ If the eighth day of the month comes on Tuesday, what is the date of the following Friday?
▶ How many minutes has the large hand to travel between seven minutes after a quarter past the hour and three quarters past?

Clever Hans became more famous than other performing animals because no obvious trickery was involved. Even when they were looking for them, visitors saw no cues that would prompt Hans to give the right answer. Mr. von Osten, himself, was thoroughly honest. He accepted no money for displaying his horse's talents and he readily agreed to let scientists and animal experts examine Hans' performances as carefully as they liked. Further, if Hans knew them well enough, he could correctly answer questions from people other than Mr. von Osten.

There was such interest in Clever Hans that a panel of 13 eminent experts, including naturalists, zoologists, teachers, and veterinarians carefully examined Hans' ability to answer questions both in the presence and in the

absence of Mr. von Osten. The panel concluded that Hans could answer questions whether Mr. von Osten were present or not and reported unanimously that they could detect no trickery. The zoologist, K. Mobius, said that he was convinced of the horse's ability to count and to do arithmetic.

After the panel had released its report, the psychologist Otto Pfungst performed a new set of experiments that created doubt about the nature of Hans' abilities. Pfungst found that Hans could not answer questions if no one in the room knew the answer. For example, if Hans was given a set of rings to count that no one else in the room could see, then Hans answered randomly. Further, if Hans wore a set of blinders that prevented him from seeing anyone in the audience, he could no longer answer questions correctly. Pfungst concluded that Hans was starting and stopping his tapping by noticing visual cues made by people who knew the answer to the question. Given this clue, Pfungst looked for very slight bodily movements by the questioners. He found that when a question was asked, the questioner would move his head very slightly forward to see Hans' hoof taps. This slight movement was Hans' signal to begin tapping. When the questioner expected a long answer, his head movement was a little more marked. The greater the head movement, the faster Hans tapped. When Hans approached the correct answer, the questioner involuntarily straightened up. This was the cue for Hans to stop tapping. Pfungst found that if he made these movements voluntarily, he could get Hans to start and stop tapping even though no question had been asked.

Clever Hans performance really was remarkable, even though it wasn't what people at first thought it was. Hans really couldn't do arithmetic or spell. However, in order to get rewards of carrot chunks or sugar cubes, he learned to detect movements so small that most people couldn't find them even when they were looking. The case of Clever Hans tells us two important things:

First, in looking at the behavior of an animal, it may be very difficult to identify the processes that lead to that behavior. What looked like calculation in Hans turned out to be perception of small movements. People may have attributed calculation to Hans because calculation is the way they, themselves, would have solved the problems that Hans faced. Hans

chose a different route. Perhaps because people are fond of animals (especially the fuzzy ones) they may tend to empathize with them; that is, they may attribute to them the same sorts of thought processes and feelings that people experience. This process of attributing human attributes to non-human animals and things is called anthropomorphizing. In the case of Clever Hans, and in other cases we will discuss later, anthropomorphizing can get in our way when we try to understand the nature of intelligence in other animals.

Second, the case of Clever Hans illustrates a very general problem that bedevils research whether it be with animals or humans. The problem is called "the researcher effect." Researcher effects are any unintentional effects that the observer has on the human or animal being observed. When Mr. von Osten unintentionally cued Clever Hans to give correct answers, that was a researcher effect.

Rosenthal & Fode (1963) conducted a study that illustrates dramatically how researcher effects can influence the results of research. A dozen psychology students served as researchers in a study of learning in rats. The students believed that the rats had been bred to be either bright or dull. Actually, all the rats were the same. Half of the students were told that their rats were bright and the other half, that their rats were dull. Over five days of testing, the "bright" rats outperformed the "dull" rats and the difference was statistically significant. Why the difference? Rosenthal believed that it might be a result of the way the researchers handled the animals. In experiments with rats, it is common practice to handle and pet the animals before an experiment in order to reduce the animals' anxieties in what for them may be a frightening situation. Students who believed that their rats were bright said afterwards that they handled the animals more often and more gently than did students who believed their animals were dullards. Thus, the researcher's attitude may unintentionally have influenced how calm and ready to learn the animals were.

Problems in Drawing Inferences from Observations

These examples and hundreds of others I could describe make it clear that it is easy to draw wrong conclusion from observations of behavior. For example, you probably know pet owners like the one shown in Figure 1.1 who believe that their pets understand English sentences. As we will see, even the most capable animals have very limited ability to understand language. Further, there is a great deal of misinformation that is widely believed. Legend has it that elephants have wonderful memories but, in fact, they don't (Carrington, 1959). If we are to draw solid conclusions about mental processes, we need to be able to think critically about our sources of information and about the conclusions that we and others may draw from observations. We need to be aware that a given set of observations may have more than one plausible interpretation.



Figure 1.1 Communication between man and dog.

The advice to think critically when drawing psychological conclusions may strike you as laudable but not very practical advice, as when people tell us to "be ready when opportunity knocks" but neglect to tell us exactly how to do that. Fortunately, there really is something that you can do that will help you to think more

critically. Through years of experience, researchers have identified some of the problems that most frequently lead to faulty reasoning about observations. The researcher effect that was a critical problem in interpreting the Clever Hans case is just one of many problems we need to be alert to in drawing conclusions about mental processes of animals (or people) from observing their behavior. I have listed several (but not all) of the major problems in Table 1.2. For convenience, I have organized the most common of these problems into three categories: reactivity, that is, the effects of the process of observing on the observed, accuracy of observation, and confounding.

In the following section, I give examples to illustrate each of the problems. An important part of thinking clearly about mental processes is being familiar with these problems. If you learn to recognize them and the situations in which they often occur, you will be better prepared to recognize and avoid some of the traps that you might otherwise fall into.

Table 1.2 Problems in drawing inferences from observations.

Reactivity: Observing Influences What is Observed

If observing something changes what is being observed, then we say that the act of observing is reactive. Different methods of observing differ in how reactive they are. A method in which a person of one race interviews a person of another race about racial attitudes is likely to be highly reactive. In contrast, an anonymous questionnaire about political attitudes will probably be less reactive. Some observational techniques are not reactive at all. For example, assessing the popularity of various types of reading by examining library circulation records probably has no effect on the public's reading behaviors. Historical studies, of course, are completely nonreactive. There are two major sources of reactivity in empirical studies: researcher effects and participant effects.

Researcher effects: The researcher influences the observations.

The case of Clever Hans and the case of the student experimenters discussed above illustrate that researcher effects can be both important and subtle. In neither case did the observers know that they were influencing the observations. Researchers can do a number of things to reduce such effects on the outcomes of studies. For example, they can restrict the opportunities they have for treating participants in different conditions differently. Instructions can be scripted or, in some cases, tape-recorded or videotaped to insure that everyone in the study is treated in the same way. The experimenter's responses to questions might be scripted as well. In many studies, participants' questions are answered simply by referring them to the appropriate section of the instructions. In conducting studies, researchers should be as unobtrusive as possible. For example, the researcher can try to stay out of the participant's line of sight and intervene as little as possible during data collection. In some studies, called experimenter-blind studies, the person running the study doesn't know whether the participant is in the experimental or control condition. Techniques such as these can greatly reduce researcher effects.

An example of an experimenter-blind study is described by Payne (1998). Elephants are able to produce sounds well below the range of human hearing. The researchers wanted to know if elephants could respond to these sounds over long distances. The study involved two groups of researchers: one group photographed the elephants during an agreed on ten minute interval and the other group, a mile or more away, broadcast a low frequency elephant call during one of the ten minutes. Researcher's examined the films without knowing when in the ten minute interval the sound had occurred and looked for evidence of response such as raised ears, head turning, etc. The observations provided clear evidence that the elephants did respond to distant low-frequency calls.

Participant effect: The experience of being observed influences those being observed.

People who participate in research studies respond in many different ways to being observed. They may feel honored to be singled out for

study, they may feel anxious to perform well, or they may feel spied on. In any case, they probably feel and act somewhat differently than they would if they weren't being studied. The probable extent of these differences must be taken into account in interpreting the results of studies.

Two special cases participant effects are called the Hawthorne effect and the placebo effect. The Hawthorne effect was discovered by social psychologists who were interested in finding out what made workers productive. They changed a variety of work conditions including lighting, ventilation, and rest periods. Each change they made improved productivity. Finally, they changed everything back to the way it was at the beginning of the study and productivity went up again! The researchers concluded that the changes in productivity happened because the workers enjoyed the attention that they were getting from being in the study.

The placebo effect is the tendency of patients to feel better if they believe they are receiving treatment for what ails them. Thus, patients with colds are likely to report less coughing and sneezing if they are given inert pills (called placebos) that they believe they are being given medication. Because the placebo effect is well known to medical researchers, it is typically taken into account in testing the effectiveness of new medications. In such tests, half of the participants will be given the new medication and the other half, inert pills. None of the participants know which kind of pill they have received. To be considered effective, the new medication must at least be more effective than the placebo.

Participant effects can be reduced in a number of ways. For example, it may be possible to design an experimental study so that all of the participants undergo both the experimental and control treatments. In a study of the impact of computers on writing, each participant might write some essays with a computer (the experimental treatment) and other essays with pen and paper (the control treatment). This way, if a participant writes better (or worse) while being observed, at least he will do so in both conditions. Participants may respond differently, for example, be less motivated, if they believe that they are in the control condition rather

than the experimental condition. A strategy to reduce this effect is to make it appear that all participants are in the same condition even though they are not. Suppose that a study is to be conducted by distributing packets of experimental and control materials to students in the same classroom. If the researcher does not call attention to the differences in the packets being distributed, the participants may not notice them. Incorporating such strategies in the design of the study can often reduce participant effects.

A double-blind experiment is a technique for managing both researcher and participant effects at the same time. Imagine a study designed to measure the effectiveness of a new medicine on the common cold. Half of the patients are given the new medicine and half are given a placebo (a sugar pill). Both physicians and patients are asked to judge the severity of the symptoms before and after medication. However, neither the physicians nor the patients are told who has gotten the medicine and who has not. Thus, if symptoms are reduced more in patients who took the medicine than in those who took the placebo, that result can't be attributed either to researcher or participant effects.

Accuracy of Observation

Sampling Bias: All members of the relevant group not represented equally.

In many cases, considerations of cost and time make it impractical to study all the members of a population we are interested in (e.g., all of the students in a large college or all the children in an urban school system). Instead, we often have to study a sample of the population that we hope will represent the whole. How that sample is chosen is important because some ways of choosing samples may systematically bias the results of the study. For example, suppose that a researcher wanted to know about teachers' attitudes toward computers in education. One very natural thing for the researcher to do would be to interview friends who teach since they are easy to access. Unfortunately, one's friends are usually not a representative sample of the population. Friends tend to resemble one another in age, gender, race, and professional specialization. If attitude toward computers in education varies with age, as seems likely, or any of the other factors in which friends tend to resemble each

other, then samples of friends are likely to provide biased results.

Another way to get a sample would be to post a bulletin asking teachers to volunteer to be interviewed. Unfortunately, samples of volunteers are usually not representative of the population either. Teachers who are hostile to computers in education or who are uninterested in the topic are less likely to volunteer for interviews than are those who are very enthusiastic about it. Thus, results based on a volunteer sample are likely to be biased.

The best way to pick a sample is through randomization. A random sample is one in which every member of the population has an equal chance of being included. Whether a given individual is included in the sample or not depends on a random choice, such as one made by picking a number out of a hat or rolling a die. For example, suppose that we wanted to pick a sample of 20 teachers from the 400 teachers employed at a college. First, we would assign a number from 1 to 400 to each teacher. Then, we could write the numbers on slips of paper, put the slips in a bowl, and stir them up. The sample of teachers would be those corresponding to the first 20 numbers chosen from the bowl. Alternatively, one could choose numbers from a table of random numbers. Choosing samples by randomization is the surest and simplest way to avoid bias.

A dramatic example of the importance of unbiased sampling was provided in the attempt to predict the winner of presidential election of 1936. The contestants were Roosevelt, the democrat, and Landon, the republican. At that time, the most widely accepted public opinion poll was conducted by a popular magazine, the *Literary Digest*. To predict the election outcome, the *Literary Digest* mailed out about 10 million ballots contacting about 1/3 of all households in the United States. The ballots were sent to individuals identified through telephone directories and automobile registration lists. On the basis of the returns, *Literary Digest* predicted that Landon would win the election with 57% of the vote. In fact, Roosevelt won receiving 61 % of the vote.

What went wrong? Even though the *Literary Digest* collected a very large sample, it was a biased sample. In 1936, many poor families had neither automobile nor telephone. In 1960, only 80% of

households had a telephone. Thus, the sample was biased toward wealthier families. In 1936, during the depression years, poorer families tended to vote for democrats and wealthier families for republicans.

In contrast to the *Literary Digest*, George Gallup used polling to predict the election outcome correctly. Gallup used a technique called quota sampling. First Gallup used the census to determine what proportion of voters were male and female, rural and urban, lower, middle, or upper class, and so on. He then set quotas for his interviewers so that they interviewed people in each of these groupings so that the proportion of people in his sample who were, say, middle class women matched the proportions of middle class women in the county. Thus, Gallup's sample was representative sample. On the basis of his sample, Gallup predicted that Roosevelt would win with 54% of the votes. Gallup's prediction wasn't perfect but it was much more accurate than the *Literary Digest's* prediction.

It is interesting to note that Gallup sampled only about 3000 people in contrast to the 10 million sampled by *Literary Digest*. To insure accuracy, it is much more important to have an unbiased sample than to have a large one. Currently, national polls are typically based on between 1000 and 3000 respondents.

Sampling bias can apply to materials and contexts as well as to participants. For example, a researcher who was interested in demonstrating that students have trouble understanding their textbooks might be inclined to study the worst of the available textbooks. In fairness, the results of such a study would have to be qualified to indicate how the texts were chosen (e.g., "Students have trouble understanding really bad textbooks"). If the results are to be applied to textbooks generally, then the texts studied should be a random sample of those available. Similarly, if we are to draw reasonable conclusions about how students solve problems in college, we should not base those conclusions on observations of students in a single context, for example, psychology classes. Rather, any such conclusion should be based on observations of students in a random sample of the various and often very different contexts in which college students are required to solve problems (e.g.,

biology labs, school newspaper offices, philosophy exams, etc.).

Observer reliability: Do independent observers agree?

Imagine that a school is planning to institute the following placement policy: all incoming students are to be placed into basic, average, or honors writing classes on the basis of their performance in writing an impromptu theme. For reasons of fairness, all of the students' essays are to be graded by the same English instructor, Mr. Brown. At a faculty meeting, another member of the English department, Ms. Jones, voices concern about the placement policy. "Suppose," she says, "that these essays were graded independently by two English instructors and their judgments didn't agree at all. Wouldn't that mean that by the proposed policy we would be assigning students to basic and honors courses pretty much arbitrarily? Why not just draw lots?" Clearly, Ms. Jones has raised the question of observer reliability.

The issue here is not Mr. Brown's qualifications. Even if all faculty members were considered good judges of writing, they still might not see the same things or value the same things in the students' essays. Thus they may not agree in placing students into basic, average, and honors sections. To estimate how well instructors agree in placing students, it would be appropriate to ask two or more instructors to judge the essays independently. Their judgments could then be compared. The usual way to make the comparison would be to calculate the correlation between the judgments of the pair of instructors. A correlation of +0.80 or better is generally taken as indicating a satisfactory level of observer reliability.

Observer bias in reporting observations.

Bias in reporting the facts is something we learn to take into account everyday. We expect salespeople to tell us the good things about their products but not the bad. We know that people often play up their successes and play down their failures. However, bias can enter in subtle ways. People giving biased judgements may not know that their judgments are biased.

An interesting study of observer bias concerned the phenomenon of psychokinesis—the supposed

ability of some individuals to move physical objects by willing them to move. In this study, individuals who claimed psychokinetic abilities attempted to influence the number that would appear on a pair of dice in a series of rolls. For example, on a particular roll, they would try to make the dice come out "nine" or "seven." The observers in the study were people who believed in psychokinesis, people who didn't believe in psychokinesis, and a camera. The job of each observer was to record whether the "influencer" succeeded in making the dice turn up in the desired way. The people who believed in psychokinesis found that the influencer did better than chance. The people who did not believe in psychokinesis found that he did worse than chance and the camera found that the results were what one would expect by chance.

A special form of observer bias is called the "halo" effect. In the halo effect, judges tend to score individuals who have performed well in the past better than their current performance would warrant. Thus, a student who received an "A" on the first paper in a course is likely to be graded more leniently on the second paper than a student who received a "D."

Even if the observers are trained researchers, that doesn't mean that their observations will be unbiased. Researchers usually care a great deal about how their research comes out. They spend much of their time thinking up clever theories and hypotheses, and they are, quite naturally, eager to see these theories confirmed in their data. The result is that researchers can be quite biased observers. Knowing this, good researchers take special care to reduce and control biases in evaluating data. A very effective way to do this is to use "blind scoring." In blind scoring, care is taken so that the person who does the scoring doesn't have information that would allow for bias. Thus, the scorer shouldn't know whether the individual being scored was in the experimental or control group, whether a particular measurement was made pre-treatment or post-treatment, or, in some cases, whether two performances were made by the same or different individuals. Blind scoring serves a somewhat different purpose than experimenter-blind studies. Experimenter-blind studies prevent the experimenter from treating experimental and control participants differently while they are participating in the study. Blind scoring is used to insure that the data obtained from the

experimental and control groups is scored in an unbiased way.

Self-reporting bias: People often do not report accurately about themselves.

Suppose that a college committee designing a humanities course wants to know what students read on their own outside of class. One way to try to get that information would be to ask students to fill out a questionnaire in which they are asked to report how frequently they read novels, poetry, non-fiction, plays, biography, science fiction, etc. The problem with such self-reporting is that respondents may tend to overreport high prestige reading, e.g., novels and poetry, and under-report low prestige reading, e.g., Playboy and The National Enquirer. Researchers for the National Assessment of Educational Progress attempted to reduce this kind of bias by asking students to name the books they had read. The assumption here is that students' claims to have read novels are more believable if they are actually able to name some.

Another way to deal with bias in this situation would be to make use of multiple sources of information about student's reading habits. For example, one might get information about what sells at student bookstores, what books are checked out of college libraries, and what reading materials students carry around with them. Using many sorts of measures to investigate the same question from different angles is called triangulation. In this case, triangulation can help to control for self-reporting bias because it would allow the researchers to compare survey data that may be biased with bookstore sales data that are less likely to be biased.

Confounding: Confusion about What Causes What

When two events occur together routinely, that is, when they are strongly correlated, we often infer that one causes the other. Sometimes the inference can be seriously in error. Consider the case of a researcher interested in finding practices in the lifestyles of the elderly that lead to long life. Among other questions, the researcher asked residents of retirement homes whether or not they currently had active sex lives. When he returned to the retirement homes a

year later, he found that the survival rate of those who had reported active sex lives was substantially greater than of those who did not. He concluded that sexual activity leads good health.

Now, while sex, like any exercise, may promote health to some degree, it seems likely that the researcher in this case had the causal relation backward. It seems more likely that the individuals who were leading active sex lives were generally healthier than those who were not and that the causal relation was that good health leads to sexual activity rather than the reverse.

In general, when A is correlated with B, one can't conclude that A causes B. Correlation does not necessarily imply causation. It may be that A causes B but it is also possible that B causes A or A and B are both caused by some third factor, C. In the following section, we describe four such third factors that often lead to confounding when they are overlooked: chance, drop-outs, order effects, and maturation.

Chance: Did It Happen by Accident?

Whenever we observe a correlation between two events, it is always possible that the correlation occurred by chance. If a friend tells us that he can beat us at Ping-Pong and then actually whips us ten games in a row, we are impressed and may be convinced of his superior Ping-Pong skill. However, there is obviously a chance that we are exactly equal in skill (equally likely to win a game) and that the lopsided outcome of the games was accidental. In this case the chance is quite small (one in 1024) so we may not be inclined to regard it as a very serious alternative. However, there are two situations in which it is very important to consider chance as a confounding factor. One situation involves very small samples and the other involves what is called "data fishing."

Small samples: Is the number of observations sufficient to warrant the conclusion? Imagine that you are testing a new method for teaching dogs to sit on command. You choose two dogs at random and teach one by the new method and the other by the standard method. Suppose further that on a subsequent test, the dog that was taught by the new method learns to sit more quickly than the other dog. Assuming that your teaching and the test were fair to both methods, is it time for

celebration? Of course not! Even if the two teaching methods were equally good, one of the dogs would have done better on the test than the other. The chances that the dog who did better would happen to be the dog trained by the new method would be 50%. You would need to repeat your observations with many dogs before you could be confident that the new method was an improvement over the standard one.

Data fishing: Seek and ye shall find. When we believe strongly in an idea, we may be inclined to search far and wide to find evidence supporting the idea and to ignore evidence that may contradict the idea. If one measure or one statistical test doesn't provide support, perhaps others will. The probability that some evidence will provide support if we look long enough may be very high. Consider this example. Suppose that one day you are sitting in a large class with perhaps 100 people on the left side and 100 people on the right side of the room. To protect yourself against hearing what the lecturer is saying, you begin to fantasize about why people have sorted themselves out in the room the way they have. Are the people on the left left-handed and the people on the right right-handed? Is it a male/female split? Perhaps it is the length of the earlobes. In fact, people have enough traits so that if the lecture lasts long enough you will be able to find a trait or a combination of traits that differentiates all the people on the left from all the people on the right. Let's say that the trait you found was a feature of the left thumbprint. "Isn't that amazing," you might think, "that all the people over here have this kind of thumb print and all the people over there have that kind." No, it isn't amazing and that is the problem with data fishing. If you search long enough you are very likely to find something that accidentally differentiates between the two groups.

Some year ago, I was involved in a study in which my colleagues and I were trying to find a test that would predict who was and who was not likely to be a "creative" engineer. We asked hundreds of engineers to answer hundreds of test questions and we asked supervisors to rate the engineer for creativity. We then performed a statistical analysis that searched for a pattern of answers to the test questions that would differentiate the creative from the less creative engineers. Because we tested a large number of engineers with a large number of questions, it was

inevitable that we would find some pattern of answers that differentiated the creative from the less creative engineers in our sample. However, when we asked if the pattern we had found in our first sample would predict creativity in a new sample, the answer was a clear "no." The pattern we found in the first sample was accidental—simply the result of data fishing.

Every presidential-election year, attention is focused on so-called "belle weather" counties. These are counties that, over the last several elections, have voted for the candidate who actually won. Polls in these counties are watched carefully because they are thought to be especially predictive of the election outcome. Are they? I doubt it. There are enough counties in the country that we should be able to find some that voted for the winning candidate over the last several elections. Does that mean that they will vote for the winner this time? Not necessarily.

Effects of dropouts: The population may change during the study.

In a study designed to evaluate the effectiveness of a remedial reading course, researchers compared the average score on a reading test achieved by students at the beginning of the course with the average achieved at the end of the course and found a substantial increase. On the basis of this result, the researchers concluded that the course was very effective. In deciding whether or not to accept this conclusion, it is important to ask how many students' scores were involved in each average. If there were many more students at the beginning of the course than at the end, it is possible that dropouts may contribute heavily to the result. Suppose for example, that the students who dropped out were the ones who had greatest difficulty with the course. These may well also be students with relatively low reading scores. If this were the case, it might be that the reported gains occurred just because these students' scores were not included in the end-of-course average.

A good way to protect against drop-out effects would be to base the evaluation only on those students who completed the course, eliminating from consideration the pre-test scores of the students who dropped out. If big gains were found

for these students, then we would have evidence that the course actually helped some students.

Order effects: The order of testing may influence the results.

Suppose that a researcher was interested in determining whether children learned better from projects they designed themselves or from projects that the teacher chose. To answer this question, the researcher might have the children work first on the teacher's project and then on their own project. If the researcher did this, though, the study would be open to the criticism that any difference found might be due to the order in which the projects were done. Perhaps the second project would benefit from the students' practice or suffer because of their fatigue or boredom. To deal with this problem, the researcher can require half of the children to work on the teacher's project first and the other half of the children to work on their own project first. In this way, each project will, on average, be equally advantaged or disadvantaged by the order in which the projects are executed. This process of balancing the testing order is called counterbalancing.

Maturation: The effect may result simply from the passage of time.

People, especially young ones, get better at a lot of things as they get older. For example, we would expect kindergarten students' vocabularies to increase over the course of time whether we gave them special vocabulary lessons or not. Thus, if a researcher reported that kindergarten students increased their vocabularies by 15% when they were exposed to an enrichment program for a year, we wouldn't know whether or not to be impressed. After all, between ages four and five, children increase their age by 25%. Perhaps vocabulary could be expected to increase by a similar amount. To assess the effectiveness of the enrichment program, we would have to compare the reported improvement to the amount students would have improved their vocabularies over the same period of time without the enrichment program. That is, we would need to measure vocabulary gains over the same period in a control group of students who did not have the enrichment program.

In an article recently submitted for publication, the author focused on the relation between

writing ability and personality traits. The author had found a strong relation between writing ability and the presence or absence of a particular personality trait. What is relevant for us here is that the personality trait was strongly related to age. Older children had it, and younger children didn't. Thus, what the author of the article had interpreted as a rather surprising effect of personality on writing ability may well have been a not very surprising effect of age. In this study, age was a confounding factor because it varied along with the variable that the researcher was attending to. As such, it provided opportunities for alternative interpretations of the author's claims about the meaning of the data.

Detecting Hidden Abilities

In this chapter, we have focused on the importance of critical thinking for understanding intelligence. Indeed, it is an essential part of the research process. When we are given relevant observations, we need to be careful in interpreting them as we saw in the case of "Clever Hans." However there is a complementary problem in understanding intelligence. There are things we may want to know about an animal's mind but have no idea what to observe in order to answer them. The problem is not figuring out how to interpret what we see but rather figuring out what evidence to look at. Figure 1.2 illustrates the problem. How would you find out if your dog understood commands she didn't respond to? How would you find out if a chimp recognized that the interesting face staring at him from the mirror was actually his own? How would you decide if apes can tell lies? How would you tell if a newborn infant sees the world in 3-D? Often, the hard part of the researcher's task is the detective's task of finding clues that will help to answer these questions. The task of the researcher is to think of clever ways of observing that will shed light on questions such as these. In major part, the rest of this book is devoted to outlining the inventions of numerous scientists that allow us to look inside the minds

of animals and infants to answer questions such as these.



Figure 1.2 Private communication between dog and dog. (Reprinted from The New Yorker, July 8, 2002, p47.)

Critical Thinking Cases

Below are 15 cases in which people have drawn inferences from observations. In at least some of the cases, there are one or more plausible alternatives to the interpretation that was drawn. Solving these cases can help you to acquire critical thinking skills that are useful in interpreting observational data. Your task is to:

1. Read each case carefully,
 2. Find the alternative interpretations, if any, in each case, and
 3. Identify the problem type in Table 1.2 that may have led the person in the case to overlook the alternative interpretation.
-
1. Fred, an undergraduate, has been told by his psychology professor that goals and plans are very important for successful performance in school and in life generally. Fred wanted to see if this odd notion actually applied to success in school. He interviewed all of his fraternity brothers to identify those with clear professional goals, that is, those that had chosen a profession such as chemist or writer. He found that those with professional goals earned significantly better grades than those who were uncertain about their career choice. Further, he found that to a great extent, the better grades of those with professional goals were earned in courses relevant to their career goals (e.g., aspiring chemists did especially well in chemistry courses). This evidence makes a lot of sense if you believe that goals are important for success. However, there are some other interpretations that Fred should consider before he draws conclusions from his data. What are they? What sorts of data could Fred collect that would help him to evaluate the various hypotheses?
 2. A researcher is interested in the relation between shoe size and problem solving ability. For the purpose of his study, he obtains permission to test students in a local junior high school. He asks for and obtains 40 volunteers. He then measures the size of each student's shoes and administers a problem-solving test. His result shows a moderate but significant positive correlation between shoe size and problem solving ability. If your little brother in junior high school has enormous feet, should you be optimistic about his problem solving ability on the basis of this study?
 3. It has been claimed that formal training exercises in topics such as Latin or mathematics strengthen the mind. A researcher believes that this maxim applies to college studies. In particular, he believes that training in the sciences will improve scientific intelligence and that training in the humanities will improve verbal intelligence. To test this belief, he paid 100 randomly selected college seniors in the class of 1999 (50 in the Science College and 50 in the Humanities College) to take the SATs. He then compared the 100 seniors' SAT scores to the average SAT scores of the class of 1999 when they entered the Science and Humanities Colleges as freshmen in 1995. He obtained the following results:

	SAT Math		SAT Verbal	
	Science	Humanities	Science	Humanities
Freshmen (1985)	602	541	535	620
Seniors (1989)	636	539	532	654

The researcher claims that these data supports his hypothesis because the science students gain in SAT MATH while the humanities students gain in SAT VERBAL. Are there other plausible ways to explain these results?

4. To test the effectiveness of a creative thinking course for executives, the Personnel Manager of a company asked 25 executives who had just finished the course and 25 other executives to keep a record of their "successes" over a one-week period. Those who had taken the course reported 25% more successes than those who had not. On the basis of this result, the Personnel Manager recommended that the course would be beneficial to all of the executives in the company. Could he be making a mistake?
5. A group of researchers collected five protocols of students writing essays. One of the researchers reads through the protocols to identify instances of planning and then evaluates the essays for quality. The researchers find that the higher rated essays were written by students whose protocols showed more planning. They conclude that more planning results in better essays and that students should be taught to plan in order to improve the quality of their essays. Are there other ways to interpret these data?
6. A researcher is interested in the effectiveness of review sessions in preparing students for exams. When he compares attendance rates at review sessions and grades, he finds that frequent attenders get much higher grades than do infrequent attenders. On the basis of these observations, he recommends the use of review sessions. Could he be overlooking something?
7. A biologist examined the reactions of young geese to a cut-out silhouette of a bird that resembled a hawk (short neck, long tail) when moved on a string in one direction and a goose (long neck, short tail) when moved in the other direction. The birds behaved as if they were trying to escape when the silhouette was in the first position, but not when it was in the second. Since the wings on the silhouette were symmetrical from front to back, the difference in response could not have been based on wing shape. The researcher concluded that young geese respond to shape in relation to direction of movement and not just shape alone. Can you find any weaknesses in this argument?
8. Imagine that you are the assistant to the PA Commissioner of Education. The Commissioner has been very impressed by a study in which several hundred 12th graders across the country were observed while they were writing 50-minute in-class essays. The main findings of the study were:
 - (1) students differed widely in the amount of planning which they spontaneously engaged in before they started to write (some began writing immediately while others spent up to 10 minutes of the available time in planning, and
 - (2) students who did more planning wrote better essays (as judged by a panel of experts) than those who did less planning.

On the basis of this study, the Commissioner is considering requiring high school seniors to take a 3-week course designed to increase the amount of planning they do. Do you think that the study justifies this decision? Identify relevant alternative interpretations that you think the Commissioner should consider.

9. In a survey of elementary schools around the country, a researcher finds that schools providing computers for students have a much lower rate of behavior problems than schools that do not provide computers. He concludes that schools that do not provide students with computers could reduce their rate of behavior problems by providing computers. Are there other ways of thinking about this?
10. Smith is the president of a collection agency that employs many people to write threatening letters to deadbeat clients. He believes that some people are much better at writing collection letters than others. To identify the more talented collection letter writers, he develops a psychological test that includes questions such as "People are swine -- true or false" and "When did you last kick your dog?" He had this test administered to all new employees. To check its effectiveness, Smith provided the test scores to his supervisors and asked them to report to him about the performance of the above and below average scoring letter writers. Smith was pleased with his test because the supervisors reported that the 50 high scorers wrote much better collection letters than the 50 low scorers. Is Smith thinking critically enough?
11. In a "brainstorming" session, a group of people works together to generate ideas to solve a shared problem. A researcher notes that brainstorming groups produce more ideas than do single individuals attempting to solve the same problem. On the basis of this evidence, he concludes that people in brainstorming groups work more efficiently than do individuals solving problems alone. Are there plausible alternatives to this conclusion?
12. Sir Francis Galton examined the biographies of eminent British writers, scientists, artists, and politicians. He found that the eminent individuals were much more likely to be related to each other than one would expect on the basis of chance. On the basis of these observations, he concluded that the tendency of an individual to be eminent is a trait inherited from relatives. Do we have to believe Galton's conclusion?
13. *Weekly World News* December 10, 1985:

"STUDY IS A WASTE OF TIME," Professor claims. School kids can study 'til they're blue in the face -- but they're wasting their time because the practice has little or nothing to do with making good grades! "We never would have predicted it, but studying may not pay off in high grades," said Dr. Edward J. Walsh, associate professor of sociology at Pennsylvania State University.

Walsh reached this shocking conclusion on the basis of a series of studies which showed very little correlation between how much students study and how good their grades are. Is his conclusion warranted by these observations?
14. A survey organization mailed a questionnaire to 4200 people treated at a chiropractic clinic in the last two years and received 2040 replies (not bad for a mail survey). Nearly 90% of the respondents said that they had been helped or helped greatly by their treatment. An advertisement based on this survey said "90% helped by chiropractic treatment." Is there anything questionable about this ad?

15. In *The Psychopathology of Everyday Life*, Freud claims that slips of the tongue are caused by peoples' personal problems. To illustrate his point, he describes the case of a young man he talked to on a train who made a slip of the tongue. By asking the man to free-associate to the slip of the tongue, Freud soon found an association to an important personal problem that was bothering the man. If Freud had been able to find an association to a personal problem for every slip of the tongue he observed (suppose there were 1000 of them) would that prove his point?