excerpts from

Rival Hypotheses: Alternative Interpretations of Data Based Conclusions by Schulyer W. Huck & Howard M. Sandler

Part ONE

1. A Painful Look at Hunger: Correlation Isn't Causation

Current theories of hunger place the responsibility on a small part of the brain called the hypothalamus. The hypothalamus monitors the chemical content of the blood and triggers eating responses at appropriate times. It has even been shown that certain surgical lesions in the hypothalamus can lead to nonstop eating—the rats literally eat themselves to death. Early theories of hunger, however, were not so sophisticated. They were based mainly on the assumption that the stomach had a causal influence on the brain—that is, since hunger pangs were often reported by those who were hungry, it seemed logical to look at the influence of stomach contractions on hunger.

One of the earliest studies in this area was done by two researchers (Cannon & Washburn, 1912) who had human participants' swallow a small balloon that was then inflated. The air pressure in the "gastric balloon" was affected by stomach contractions that were transmitted to a recording device. The participants were also asked to indicate each time a hunger pang was felt. The researchers interpreted the strong positive correlation between stomach contractions and hunger pangs to mean that the contractions caused the pangs. **Where did they go wrong?**

2. Hypnosis and Biofeedback: Sampling (or Participant-Selection) Bias

A nineteenth-century physician named F. A. Mesmer was among the first to investigate hypnotism in a scientific manner. Although "mesmerized" is now used almost exclusively to mean "fascinated," it was at one time synonymous with the term "hypnotized." Many claims have been made over the years for the power of hypnosis, but our favorite involves the use of hypnosis to create a temporary state of unawareness in the hypnotic participant. We have never understood how the participant heard the command to come out of the hypnotic state if they were unaware!

Hypnotism was also involved in a recent study of biofeedback. The researchers were interested in seeing whether some individuals could exercise voluntary control over their peripheral skin temperature, as measured by the relative warmth of their hands. In this study, six college students were selected for their "hypnotic talent" (their ability to be hypnotized) and their "extensive" previous hypnotic training and experience. Four of the students had some experience in meditation; two were still actively pursuing meditation during the experimental period.

Participants were hypnotized before each experimental session and then asked to make one hand warmer than the other. Some began with the right hand and some with the left, the hand chosen alternated from one session to another. Headphones were used to carry a biofeedback tone to the participants. The tone would shift toward the right ear when the right hand was becoming relatively warmer, and the frequency of the tone would also increase as the temperature differential increased. At the end of the session the hypnotic state was lifted.

The results indicated that participants were able to create temperature differences between their hands. However, different approaches were used. For example, some would raise the temperature of the target hand, while others would lower the temperature of the non-target hand. Still others would raise the temperature of both hands at different rates and thus create the desired difference. One

thing the participants did share, however, was a "high degree of motivation and involvement" along with the belief that the experiment was both a "helpful and valuable experience."

In addition, all participants felt that the hypnosis had helped in their attempts to exercise control over their skin temperatures. The researchers concluded that "some individuals are capable of achieving voluntary control over the autonomic processes involved in peripheral skin temperature regulation" (Roberts el al., p. 168). We agree. They also pointed out a confounding factor in their study, involving their sample of participants. **How could "Sampling (or Participant Selection) Bias" account for their results?**

3. Required Textbooks: Researcher Bias

The cost of a college education is skyrocketing. And while we usually think of tuition, room, and board as accounting for the largest chunk of the student's expenses, the financial outlay associated with other little items typically adds up very quickly. For example, we have known students who went into their campus bookstore with a list of the texts required for their new courses and left the bookstore with a $1000 dent in their checking account.

Little wonder that someone finally conducted a study to see whether students could learn as much without a required text as they normally do when one is required. The research project took place over a two-year period at Marquette University in Milwaukee, with the focus of the study being an introductory course in psychological statistics. During the first semester of the 1974-75 academic year, 35 sophomores majoring in psychology signed up for this statistics course. On the first day of the term, the students were given a syllabus that stated, in part, that there was no text required for the course nor any specific book even recommended, and that they would be receiving various handouts containing formulas, problems, and the necessary statistical tables. A year later, the same course was again offered to 33 psychology majors. This second group of students also received a course syllabus on the first day of the term, and it was identical to the one used the previous year—except that it listed the required text and where it could be bought.

The same instructor was used for both sections of this statistics course, and each time the course was taught "in a traditional manner which relied heavily on lectures and discussions, with occasional demonstrations" (Quereshi, p. 2). The two groups of students received the same set of handouts (containing the formulas, problems, and tables); they met for the same number of class sessions; and they were given the same number of quizzes and take-home exercises. Furthermore, the specific quiz and exercise items remained constant from one year to the next, with identical directions and credit being given for correct responses.

Because complete data (scores on the three quizzes and the take-home exercises) were not available for nine of the students in the first class or for eight of the students in the second class, the statistical comparison of student performance without the text versus performance with the text was based on sample sizes of 26 and 25, respectively. From the data provided by these 51 participants, no significant difference was found to exist between the two groups with respect to their take-home exercise scores. However, a significant difference was associated with the in-class quiz scores, with the students in the required-text group earning higher scores than students in the no-text group. With these results as ammunition, the researcher concluded that "using a textbook in conducting an undergraduate course in psychological statistics has a definite beneficial effect on performance on closed-book, subject-mastery tests or quizzes" (p. 2).

Whether or not the findings of this study generalize to the specific courses that you take or teach clearly depends upon several important considerations—for example, the nature of the text used, the type of students enrolled in the course, the degree to which the instructor's lectures correspond with the text, and so on. However, we would like you to put aside the issue of generalization for the moment. Instead, focus on the results obtained by this particular researcher at Marquette during the 1974 and 1975 fall terms. **How could experimenter bias explain the researcher's conclusion (which were drawn from his own students and research participants)?**

4. Angina Pectoris: Participant Bias

Angina pectoris is the medical term used to describe a severe pain that some people feel in their chests. The pain is caused by too little blood getting from the mammary arteries to the tissue around the heart. Many physicians believe that this condition, if allowed to worsen, will eventually bring about a heart attack. Consequently, several different remedies have been tried, ranging from doing nothing (and hoping that the problem will simply go away) to surgery.

The surgical techniques used to deal with angina pectoris vary. Sometimes the arteries are scraped out; other times, they are bilaterally ligated [tied or otherwise closed off]. This latter procedure involves tying the main mammary arteries to improve circulation to the cardiac region among the existing secondary routes and encourage the development of new routes. Because there are, in this case, alternative surgical techniques aimed at correcting the same problem, it is encouraging to know that research studies are sometimes conducted to assess the worth of each technique. Let us now consider one such investigation.

In this particular study, there were 50 patients suffering from angina pectoris. Each of these individuals was surgically treated by using the procedure wherein the internal mammary arteries are ligated. Between two to six months following the operation, the patients were contacted and questioned as to the status of their physiological problem. Of the 50 patients, 34 were clinically improved: Of those 34 patients, 18 reported no angina whatsoever following the operation, while 16 reported experiencing less severe angina. This rate of improvement (34 out of 50) is significant at the 0.05 level.

About a year after this first investigation was published, a very interesting follow-up article appeared in the medical literature. The physician-researchers associated with this second study were somewhat skeptical about the results of the first investigation; in particular, they had questions about the wisdom of using the 50 patients “as their own controls” and the presumption that their condition would remain unchanged unless treated. So, a new study was undertaken to reassess the value of ligating the mammary arteries.

This second study involved 17 patients who were suffering from angina pectoris. These participants were split, in a random fashion, into two groups: an experimental group of eight and a placebo control group of nine. The patients in the experimental group were operated on and had both mammary arteries ligated. The patients in the placebo control group were also operated on, but they did not receive the ligation; instead, they simply received the type of skin incision similar to that required to perform the ligation.

All patients, of course, were all under the impression that they had received the complete surgical procedure, and the doctors who conducted the postoperative evaluation of the patients' status were kept in the dark as to their experimental or placebo group affiliation. In other (more technical) words, this experiment was conducted in a “double-blind” fashion.

When the resulting evaluations of postoperative status were compared, there was no difference between the two groups of patients. Moreover, both groups of patients in the second study, the experimental group and the placebo/control group, improved as much as the participants in the first study. **What role could participant bias play in explaining why all three groups of patients (across the two studies) reported postoperative improvements?**

5. Brainstorming: History/Maturation Effect

Every so often, most groups of people — especially groups in work settings — encounter problems of one sort or another. Before selecting a course of action to deal with the problem, the group may engage in a “brainstorming” session. The purpose of such a session is to generate a lot of ideas, and the hope is that the best solution to the problem will appear among the suggested ideas (and then be recognized as the best idea by the group members).

One researcher hypothesized that group brainstorming might be equally or more effective than individual brainstorming if the technique of "synectics" is used. With this technique, each group member is required to act out the central object associated with the problem. For example, if the problem or task is coming up with brand-names for a new type of hot dog that's ready to be marketed, the group members are asked to lie down on the floor, one at a time for one minute, and play the part of a hot dog. During each minute, all members of the group, including the hot dog on the floor, suggest as many new brand-names as they can.

To determine whether the technique of synectics was worthwhile, the researcher who suggested its use conducted an experiment. The purpose of this study was to compare regular brainstorming with brainstorming-plus-synectics. The 88 male participants came from an introductory psychology course. The participants were divided into four-person groups, with the regular brainstorming groups being run through the experiment during the first half of the academic term, and the synectics groups during the second half of the term.

All students received course points for their experimental participation plus $7.50 an hour. Each of the four-person groups had the same eight problems to work on, and these problems were presented in the same order to each of the brainstorming and synectics groups. These problems were to devise (1) brand-names for a cigar, (2) uses for old bricks, (3) possible problems of people being tall, (4) brand-names for a winter hat, (5) uses for newspaper, (6) procedures for getting more tourists to come to the United States, (7) brand-names for an automobile, and (8) uses for old coat hangers. (It certainly would have been interesting to observe the synectics group act out some of these problems!) On each of these eight problems, each of the four-person groups was measured in terms of the total number of different ideas generated. A closed-circuit television was used to record all suggestions from all groups. The average number of ideas for the synectics groups was computed for each problem, as was an average for the regular brainstorming groups.

A statistical comparison of these two brainstorming techniques was performed on the data associated with each problem. Hence, eight comparisons were made. On five of these comparisons, a significant difference was found between the two averages — and in each case a significantly greater number of ideas was associated with the brainstorming-plus-synectics groups. Even on the four problems that did not yield a statistically significant result, the synectics group mean was higher. The researcher concluded that the synectics strategy for group problem solving was more effective than typical brainstorming. This conclusion is based, of course, on the differences between the average number of ideas generated by groups using these two approaches. **How could "history" or the "maturation effect" account for the results?**

6. Sensory Deprivation: Mortality Effect

We know a few people who have the ability to concentrate on a task regardless of the auditory and visual distractions intruding upon them. If you're like us, however, even minor distractions often cause you to lose your train of thought. A dog barking in a neighbor's yard, the shadow of a person walking by your desk, an ambulance siren, even the ticking of a clock – these small sights and sounds can sometimes make us wish that we could simply flip a switch and completely shut off all auditory and visual input. But even if such a switch did exist, we probably wouldn't use it very often. Why? Because we've read how a complete absence of environmental stimuli can have temporary or lasting negative effects on a person's mental equilibrium.

A few years ago, a study was conducted that seemed to show that sensory deprivation may actually have a beneficial influence on the ability to perform simple tasks. In this investigation, 60 men enlisted in the Navy were randomly selected from a group of volunteers who had been told that if selected, they would be serving in one of two "relaxation" conditions. From this group, 40 participants were randomly assigned to an experimental group and 20 to a control group. Those in the experimental condition were put into separate cubicles that were sound reducing and light-tight; hence, these individuals received almost no visual or auditory stimulation while participating in the investigation. The participants in the control condition, on the other hand, could adjust the lights

inside their cubicles, watch television, listen to the radio, read books, write, and even talk and text with one another on the phone.

The experiment lasted for seven consecutive days. Twenty-four hours before the start and at three points during the seven-day interval (after 25, 73, and 145 hours), all participants were given a "vigilance" test. The test involved listening to a series of 60 short tones (beeps) through a loudspeaker in each participant's cubicle. In each 90-minute testing period, these tones were randomly spaced, with a half-minute to two-and-a-half-minute interval between tones. The participant's task during the vigilance test was to pull a lever (located near the bed in each cubicle) immediately, after hearing each tone. To equalize the testing conditions during the four vigilance tests, the lights and all entertainment equipment in the control group cubicles were shut off.

Of the 40 participants in the experimental group, 21 remained in their sensory deprived cubicles for the full seven-day period of the investigation. In the control group, 19 of the 20 participants completed the week-long confinement in their cubicles.

The participants' responses to the 60 tones comprising each of the four vigilance tests were scored in the following manner. If a participant pulled the response lever within two seconds of a beep, one point of credit was given; if the level was pulled after a two-second interval or not pulled at all, no credit was earned. Since there were 60 beeps at each testing period, up to 60 points could be earned. The data from the two groups of 15 men were subjected to a statistical comparison, and the results indicated no difference between the groups at the pretest period (12 hours prior to confinement), but significant differences in each of the three tests during treatment. And at each of these testings (after 25, 73, and 145 hours), it was the experimental group who made fewer errors.

The researchers who conducted this study concluded that their results strongly corroborated previous findings that individuals who experience sensory deprivation outperform non-deprived participants on vigilance tasks. **How could a "mortality effect" account for the results of this seven-day confinement study?**

7. How to Stop Smoking: Self-Report Bias

Of the people who smoke cigarettes, cigars, or pipes, many want to continue their habit. On the other hand, there are smokers who would like to quit, but few can do so with apparent ease. For the majority of smokers, it is exceedingly difficult to kick the habit.

Several techniques have been used to help people quit smoking, such as filters that progressively screen out more and more of the smoke, group get-togethers patterned after Alcoholics Anonymous, and the substitution of gum or candy for tobacco whenever one experiences the onset of a "nicotine fit." In addition, a multitude of books and articles have been written on the subject. Unfortunately, most attempts to help smokers become nonsmokers have met with limited success. Consequently, we expect there to be a great deal of interest in a study that seems to show that an entirely different approach produces quick and lasting results.

The participant used in this investigation was a 27-year-old female who volunteered to be in the experiment. She had been smoking an average of 30 cigarettes per day for several years, and she had previously made five attempts to quit smoking. However, each of these attempts ultimately failed, and the longest interval of self-imposed abstinence was a four-month period.

The new approach to help this participant kick the habit was called "aversive smoking," and the treatment consisted of nine 30-minute sessions held on the Monday, Wednesday, and Friday of three consecutive weeks. During these treatment sessions, the participant was involved in three activities. First, she was asked to light up a cigarette and smoke at her normal rate while handling cigarette litter. Here, the participant was asked to run her hands through a five-pound bag of cigarette litter, and she was also encouraged to put her head right over the bag to smell the stink as she smoked her cigarette. The second activity required that the participant smoke at a faster-than-normal rate while a machine blew warm, stale, smoky air at her face and body. In the third activity,

the participant was encouraged to drink water (as a substitute activity) whenever she felt an urge to light up a new cigarette at home, at work, or at social events.

Before, during, and after the three-week period of treatment, the participant maintained a record of how many cigarettes she smoked each day. During the baseline period (the week prior to treatment), the participant reported smoking an average of 30 cigarettes per day. During the treatment phase of the study, the participant's smoking rate decreased steadily, and there was absolutely no smoking reported for the final eight days of this three-week period. Throughout the six-month follow-up period (which included a "booster" treatment session once each month), the participant reported complete abstinence. During the final follow-up booster session, the participant indicated that she had no desire to smoke and that the sight of other people smoking was quite distasteful to her.

According to the researchers, the dramatic change in the participant's smoking behavior was most likely attributable to the litter part of the treatment. While she was handling the cigarette litter in the plastic bag, ashes quickly found their way under the participant's fingernails; this was considered to be especially disgusting to the participant, and she reported nausea and at least once actually vomited. Even though she washed her hands thoroughly following the litter-handling episodes, the participant found it quite horrible to smell the ash residue on her arms and hands throughout the day and to see small particles of the cigarette litter lodged under her fingernails.

This study appears to document an amazingly strong cause-and-effect relationship. The effect is a drastic change in smoking habits, while the causal variable is a series of short sessions involving cigarette litter, smoky air, and talk of water. However, before accepting the researchers' conclusions about the causal role of their "aversive smoking" treatment, we need to answer a questions: **What role could self-report bias could play in producing these results?**

8. Air Force Officer School and Dogmatism: Regression to the Mean

According to a personality theorist named Milton Rokeach, dogmatism can be defined as "a relatively closed cognitive organization of beliefs about reality, organized around a central set of beliefs about absolute authority which, in turn, provides a framework for patterns of intolerance and qualified tolerance toward others." In less formal terms, the person who is highly dogmatic tends to be closed-minded, inflexible, and often more concerned about the status of a communication source than about the substance of what's being communicated. We are confident that you know at least one person who fits this description.

Many people believe that the structure of the armed services and the inherent chain-of-command basis of communication attract highly dogmatic volunteers. Military commanders, of course, disagree; they claim that they have a need for officers who are open-minded, tolerant, and able to win the respect and loyal cooperation of the personnel they direct. A researcher recently wondered whether a 14-week stint at officer training school would affect the participants' degree of dogmatism. Would this training program cause the junior officers to become more or less dogmatic, or would it have no effect on dogmatism? And would the influence of the 14-week program be the same for those participants who began with high levels of dogmatism as it was for those who began with low levels?

The participants in this investigation came from a pool of 764 officers who completed the three-and-a-half-month Squadron Officer School (SOS) at Maxwell Air Force Base in Alabama. As the researcher saw it, there were several facets of the SOS program that might have made for a change in dogmatism. For example, each student was given extensive feedback from peers, the opportunity to discuss the personality characteristics of other trainees, a chance to deal with unstructured situations, and experience in planning military strategy in areas other than their field of expertise. These and other similar activities might, according to the researcher's hypothesis, cause the office trainees' dogmatism levels to decrease over the 14-week time interval.

During the first and last weeks of training, all students in the SOS program were administered the Rokeach Dogmatism Scale, Form E. (In this study, it was titled the Rokeach Opinion Scale.) This measuring instrument is made up of 40 statements, each of which is rated on a -3 to + 3 scale to indicate the extent of one's disagreement or agreement. Two of the statements from this scale are the following: "Most people just don't know what's good for them," and "A group which tolerates too much difference of opinion among its own members cannot exist for long."

From among the SOS students who completed and returned the Rokeach Scale at both the pre-test and post-test periods, 250 were randomly selected. Then, based upon an examination of the pre-test scores, the 250 participants were subdivided into five groups of 50 participants each. In terms of the dogmatism continuum, these subgroups were described as high, above average, average, below average, and low.

The data were statistically analyzed in two ways. First, the pre-test mean for all 250 participants was compared to the overall post-test mean. Results indicated no significant difference. Next, a two-way analysis of variance was used to see whether the five subgroups changed in a similar fashion between the beginning and end of the SOS program -- or didn't change at all. The pre- and post-test means for the five subgroups turned out as follows:

Grouping Based on Pre-Test	Pre-Test Dogmatism Score	Post-Test Dogmatism Score
High Dogmatism	170.04	161.14
Above Average Dogmatism	151.02	145.42
Average Dogmatism	138.12	137.26
Below Average Dogmatism	126.12	128.78
Low Dogmatism	108.24	117.32

The statistical analysis indicated a significant interaction between subgroups and pre-post trials. (Such an interaction means that the change from pre-test to post-test is not the same for all subgroups.) Based on the subgroup means presented in the above table and the significant statistical finding, the researcher stated that "Subjects high in dogmatism on the pretest tended to become less dogmatic by the last week of training while those scoring below the mean tended to become more dogmatic" (Gleason, p. 35).

It looks as if the SOS training program causes SOS students to become more similar to each other in terms of dogmatism. **But, before drawing such a conclusion, what role could "regression to the mean" play in accounting for this statistically significant result?**

9. Professional Socialization in Nursing School: Cross-Sectional vs Longitudinal Effect

One of the major goals of faculty members and administrators in the health professions is to help students adopt realistic views of their future jobs. The rationale behind this goal is very simple, and

it is based upon empirical research. Students who have accurate expectations of their work roles tend to be more satisfied and more likely to remain in the profession.

The term “professional socialization” has been used to describe the process by which students acquire appropriate perceptions of their professional roles. As you might suspect, faculty members play an important role in this socialization. In fact, it has been hypothesized that this process can be identified by seeing whether the students come to adopt more and more of their mentors' attitudes and values as they progress through their formal training program.

Several researchers have attempted to document the phenomenon of professional socialization in both medical schools and nursing schools. But these efforts have been hindered by the unavailability of well-designed, standardized, measuring instrument. However, two individuals joined efforts to develop such a scale for use with student nurses.

The instrument, called the Nurses' Professional Orientation Scale contained 112 items that described behaviors often displayed by nurses when on the job. (Two of the items were “Question instructions when the reason for them is not clear,” and “Accept the death of a patient with no overt emotional signs.”) Each respondent was asked to rate the importance of each of these behaviors for the practicing nurse, and a five-point scale was provided with the possible choices extending from “extremely important” to “undesirable.

Scoring weights for the item responses were established by administering the scale to 94 nursing faculty members at three universities. For each item, it was possible to determine the percentage of faculty who selected each of the five response options, and these percentages (rounded to the nearest 10 percent) became the weights associated with the five options. Thus, if a student chose a response that had previously been endorsed by 90 percent of the faculty respondents, the student earned 9 points on that item; if the student chose a response that had been selected by only 80 percent of the faculty, the student would earn only 8 points, and so on. By means of the scoring system, students would earn high total scores on the Profession Orientation Scale if they rated the 112 traits in a fashion similar to the responses of the faculty members.

Once constructed, the 112-item scale was administered to 488 students enrolled in four-year baccalaureate nursing programs at Michigan State University, Wayne State University, and the University of Illinois. This participant pool was randomly split into two halves, and an item analysis was conducted (using data from the 244 students in the first half) to identify the better items. (An item was considered to be good if more seniors agreed with faculty than juniors, more juniors than sophomores, and more sophomores freshmen.) As a result of this item analysis, 52 items were thrown out, leaving 60 items in the final form of the instrument. Then, total scores on the shortened 60-item scale were computed for the 244 students in the second half of the participant pool.

To validate the new scale, the students in the second half of the participant pool were subdivided on the basis of year in school. Then, the average score on the 60-item scale was computed for each of these four subgroups. These average scores were 132.9 for freshmen, 154.3 for sophomores, 176.1 for juniors, and 193.5 for seniors. When these students’ scores were compared statistically, it was shown that the mean for each class was significantly greater than the mean for the class below. In the technical report of this study, the researchers stated that their results indicated:

> that the nurses' professional orientation scale measures a shift in the student's view of the profession that was positively related to the length of the training experience. The congruence between student and faculty ratings of items increased significantly with each year of training experience. (Crocker and Brodie, p. 234).

In general, the researchers felt that their data supported the contention, expressed and held by believers of professional socialization, that there exists a definite trend for students to gravitate toward the faculty way of viewing the profession as graduation nears.

Based on their data, the researchers are asking us to believe that nursing students come to agree more with the faculty view of the profession because of the contact the two groups have had with

one another. But did the study show that nursing students actually change? **How could the limitations of the study's cross-sectional design confound these results?**

10. Imaginary Friends: Correlation Isn't Causation

Occasionally, in the past, our children reported to us on their imaginary playmates (who were often held responsible for whatever mischief occurred in our homes). We always listened attentively and wondered about the expense of psychotherapy for young children. Having failed to come to a reasonable conclusion on the latter point, we were naturally attracted to an article entitled "Imaginary Companions and Creative Adolescents." Thinking that there may yet be hope for our pre-teenage children, we dove into the article.

What we found was a study in which 800 adolescents from schools with "outstanding records" in "creative student achievement" were classified into eight groups of 100 according to the students' sex, creativity (more versus less), and field of interest (science, art, writing).

All students were asked if they had ever had an imaginary companion. Comparisons were then made between the more creative and less creative groups (based on teachers' judgments, as well as on two tests of creative thinking) and on the frequency of reported imaginary companions. Significant differences were found between the more- versus less-creative boys for whom art was their field of interest and between the more- versus less-creative girls for whom writing was their field of interest.

The authors concluded tentatively that the presence of an imaginary companion was related to creativity in the area of writing (for girls) and art (for boys). This conclusion is made all the more appealing when we consider that many authors and artists make a fortune by sharing with the rest of us a particularly vivid imaginary character. Should we write up lucrative contracts for our children? **Or is there a caution to be held?**

11. Intelligence and Strangeness: Sampling (or Participant-Selection) Bias

Physicists use a number of constructs such as angular momentum, spin, baryon number, and parity in describing elementary particles. Our favorite construct in physics, however, is that of strangeness. Elementary particles are assigned strangeness numbers that occasionally follow the law of conservation of strangeness, depending upon the type of interaction in which the particles are engaged. It may be that many of you believe that the IQ score already serves a function similar to that which strangeness serves in physics.

One well-known longitudinal study of this belief was started in 1921 by Lewis Terman. The youngest student in each class (because they were assumed to have skipped a grade), along with the three "brightest" students in each class (as determined by their teacher), constituted the sample, which totaled 1528 children. These children's IQ scores were later determined to be above 140 (i.e., above the ninety-eighth percentile by today's standards), and they were larger, healthier, socially adept, and generally wonderful. As adults, this group made remarkable achievements in any number of areas. They also continued to display exceptionally good health, both psychological and otherwise.

These data have been used extensively to show that gifted children are not strange creatures who sit in corners and read until their eyes are ruined. **How could sampling (or participation selection bias) produce these results?**

12. Charity Begins at Home: Researcher Bias

At one time or another we all receive solicitations for charitable organizations. Maybe these solicitations are sort of randomly distributed throughout the year, with only a small amount of

seasonal variation. It often seems, however, that the first solicitation of spring signals an onslaught of similar requests. Since we already have aggregate nouns for geese ("gaggle") and larks ("exaltation"), perhaps it would be appropriate to use a "solicitation of charities" to cover these springtime activities. Regardless of when they send out their representatives, we strongly suspect that charitable organizations would be interested in research on the factors that influence the probability and the amount of giving by an individual.

In one such study, two researchers collected the responses of 240 adults contacted in the course of an actual door-to-door fund-raising drive, in an effort to study the "effect of verbal modeling on contributions to charity." Participants in the study were randomly assigned to one of 16 groups, each of which received a different combination of verbal information. One of the two researchers, who served as the experimenter, gave the participants information about the percentage of their neighbors who donated (more than three-fourths or less than one-fourth), the amount the neighbors donated (more than $10.00 or less than $5.00), the reason that the participant should donate (to fulfill their social responsibility or just to feel good), and the level of need of those helped by the charity (desperate as opposed to "could use your help").

The amount donated was used as the dependent variable in a 2 x 2 x 2 x 2 analysis of variance. The results of this analysis indicated a significant difference between the two "percentage of their neighbors who donated" conditions, with people giving more when told that over three-fourths of their neighbors had previously contributed than when told that less than one-fourth of their neighbors had contributed. Also, a significant interaction was found between the "percentage of their neighbors who donated" factor and the factor dealing with the reason for giving (participants who had been told that it would fulfill their social responsibility versus participants who were told that it would just feel good).

Additional analyses showed that combining social responsibility with the information that their neighbors had given a large amount resulted in the collection of more money. This led the researchers to conclude that giving verbal information about the behavior of others was sufficient to elicit modeling behavior" (Catt and Benson, p. 83).

How could researcher bias driven the fundraising results?

13. Fore!: Participant Bias

Although we consider ourselves to be relatively athletic and sports-minded individuals, we do not play golf very often. The primary reason for our infrequent appearance on the links is simply our inability to get that seemingly innocent little ball from the tee to the green in a reasonable number of shots. Almost every time we swing the club, the ball hooks to the left or slices to the right — assuming, of course, that we haven't whiffed. It sure would be nice if some sporting goods company could invent a ball that wouldn't hook or slice, or simply a ball that would travel a long distance, because most of our shots (even the few straight hits) trickle along the ground and only occasionally make it to the fairway.

For golfers who share our concern about finding a ball that gets good distance, a recent magazine advertisement by the Rawlings Golf Company has probably caught our eyes. Rawlings claimed to have invented a new ball — called the Toney Penna DB — that travels further than other balls when hit. (Incidentally, the letters "DB" stand for the phrase "distance ball.") The reasons provided for the new ball's distance quality are somewhat technical. We suspect that most avid golfers are in the dark, along with us, when it comes to terms such as "high-rebound core," "a rugged Surlyn cover" "weight distribution point," "centrifugal action," and "longer spin."

However, the comparative test results provided in the advertisement were not difficult to understand. To substantiate their claims, Rawlings had the Nationwide Consumer Testing Institute conduct a little experiment in which the new Toney Penna DB was compared to five other brand-name balls. In this experiment, each of 51 golfers hit 18 balls (three of each brand) off a driving tee,

using a driver and each of the six brand-name balls. The average distance that each brand-name ball ended up being from the tee was reported as follows:

1. Toney Penna DB = 254.57 yards
2. Titleist Pro Trajectory = 252.50 yards
3. Wilson Pro Staff = 249.24 yards
4. Titleist DT = 249.16 yards
5. Spalding Top-Flite = 247.12 yards
6. Dunlop Blue Max = 244.22 yards

After presenting these results, Rawlings told the reader that "as you can see, while we can't promise you 250 yards off the tee, we can offer you a competitive edge, if only a yard or two. But an edge is an edge."

The test results provided in the advertisement (and the way Rawlings talked about these results) made it seem as if there is a cause-and-effect relationship between the brand-name of the ball used and the distance you can expect to hit the ball off the tee. Although this may in fact be the case, we submit that there is at least one alternative explanation for why the Toney Penna DB turned out to get the best average distance. **What role could participant bias play in producing these long-ball effects?**

14. Counseling Practicum: History/Maturation Effect

During their formal training in their master's program, prospective school counselors are exposed to research articles, theory, and role models. It's likely that during this period, students develop an initial set of thoughts concerning ideal characteristics of school counselors. Near the end of their graduate program, however, students are often placed in a practicum course that occurs in a field setting. The hope is that counselor-trainees can put into practice the things they have learned and, of course, learn some new things. One might ask whether, as a result of this field-based experience, there is a change in counselor-trainees' perception of the desirable characteristics associated with a competent counselor.

To answer this question, researchers at Northwestern Illinois State College conducted a study. The participants were 36 graduate students in guidance and counseling, all of whom had completed 30 credit hours of required courses and were enrolled in the practicum experience. The eight-week practicum experience involved four half-days a week in a public school, with supervision provided by local school personnel and the college faculty. In addition to conducting individual and group counseling sessions, the practicum students also performed a variety of typical guidance activities.

At the beginning and end of the eight-week practicum experience, each of the 36 participants was administered the Occupational Characteristics Index (OCI). This instrument provides 12 scores, each associated with a trait that workers might have to varying degrees (for example, organizational, realist, leader, innovator). On each administration of the OCI, the practicum students were asked to identify the ideal characteristics they believed a counselor should possess. For each of the 12 traits, means from the pre-practicum administration of the OCI and the post-practicum administration were computed and compared statistically.

The results indicated that, on average, the 36 practicum students ' ratings of ideal characteristics changed from the pre-test administration to the post-test administration on 11 of the OCI traits. At the end of the practicum experience, the students believed that five of the traits were more important than they had thought prior to their eight-week field experience, and they believed that six other traits were less important. Hence, there appeared to be substantial evidence that the practicum students had changed their perception of what an ideal counselor was like.

The researchers clearly attributed this change to the field-based practicum experience that had taken place during the eight weeks between the pre-test and post-test. For example, the investigators state that "on-the-job experience did provide a statistically significant change in their perception of the ideal counselor characteristics' (Langley & Gehrman, p. 79). **What role might "history" or "the maturation effect" play in producing this result?**

15. Student Resignations at West Point: Mortality Effect

For obvious reasons, any organization is hurt by the voluntary resignation of individuals who have demonstrated the ability to perform successfully while on the job. Organization become especially concerned when people who resign have been carefully screened, when there are far more applicants than available positions, when training is costly, and when it's impossible to substitute in midstream a new person for the one who has resigned.

Such is the case at the United States Military Academy at West Point, where students are admitted, as a class, only once per year. When a student resigns after beginning the program, not only has time and money been wasted on the lost cadet, but that slot remains vacant until graduation. The staff at West Point would like to do anything it can—without lowering its standards—to decrease the number of voluntary student resignations.

Occasionally, research studies are undertaken to find out what might help reduce attrition. One study was conducted to see whether a rather simple technique might serve to decrease the number of resignations. This technique involved mailing each prospective student a booklet that described, in a realistic way, the day-to-day conditions associated with the student's future life at West Point. The booklet included descriptions of both stressful and mundane experiences, and the researchers hypothesized that it would bring about realistic expectations and consequently decrease the number who left soon after arriving because they "really didn't know what they were in for."

This particular study was conducted in the summer of 1971, during which time the new cadets were involved in an intensive training program designed to familiarize them with fundamental military discipline. As you might imagine, this two-month program required some quick adjustments on the part of the new students. The new cadets learned in a speedy fashion that their life at West Point was frequently going to be more stressful—in both physical and psychological terms—than what they had previously experienced.

About a month before the start of the rigorous summer program, a sample of 246 was randomly selected from the 1230 individuals who had indicated that they were going to accept their appointments. These 246 individuals constituted the experimental group, and each person was mailed a copy of the descriptive booklet. Presumably, the individuals receiving this information would arrive at West Point a month later better prepared for the training program.

One of the 246 booklets wasn't delivered to a cadet in the experimental group due to a change of address, and 11 other members of the experimental group did not even report for the training program (despite previously accepting the appointment at West Point). Therefore, the experimental group comprised 234 cadets, rather than 246.

A control group was randomly selected from 234 new cadets who were not sent the experimental booklet and who showed up on the first day of the training program. Cadets in the experimental and the control groups were not told about the experiment, and it can be assumed that cadets in the experimental group thought that all the new cadets had been sent the informative booklet, while cadets in the control group were simply unaware that something had been withheld from them that other cadets had received.

Records were kept of all the voluntary resignations that took place after the beginning of the summer training program (July 1) and before the first day of classes in the fall. During this period, only 14 of the 234 experimental cadets resigned, while 27 of the 234 cadets in the control group dropped out. When tested statistically, this difference —14 versus 27 resignations— was significant.

In the words of the researchers, these results "supported the hypothesis that candid information presented after the decision to participate but before entering the organization reduced the probability of voluntary resignation" (Ilgen & Seely, p. 453). **But how could "mortality" explain the differences between the experimental and control groups rate of resignations?**

16. Cigarette Smoking and Physical Fitness: Self-Report Bias

High school and college coaches advise their athletes to refrain from smoking. There are probably many reasons for this training rule, but we're convinced that it exists primarily because of a presumed relationship between smoking and physical fitness. A researcher has put this hypothesis to the test. The study was quite simple, with the subjects being required to perform a series of tasks. The researcher's hunch was also quite straightforward, individuals who smoked heavily would perform less well on tasks involving the circulatory and respiratory systems, while smokers and non-smokers would perform about the same on tasks involving minimal physical activity.

The participants in this investigation were 88 military personnel who ranged in age from 19 to 39, and they had been performing moderate calisthenics two hours a week for six weeks prior to the actual experiment. On the day of the study, each participant was required to perform five physical-fitness tests. The three tests that involved the circulo-respiratory system were crawling-feet on the ground as fast as possible, running through a 50-yard obstacle course as fast as possible, and running a mile on a grass track. The other tests—those that involved minimal stress—involved swinging from a rung on a horizontal ladder as far as possible without touching the ground and throwing five hand-weights (from a kneeling position) toward a target location 90 feet away.

All 88 participants performed the crawling test first and the one-mile run last. The other three tests were administered in different orders to different subgroups of the total participant pool. After completing the five tests, each participant was required to write down the number of they cigarettes smoked per day.

Analyzing the data from the five fitness tests and the smoking frequency question, the researcher discovered that his hypothesis had been supported. On the three tests involving physical stress (crawling, the obstacle course, and the 1-mile run), the participants who reported smoking more were found to perform significantly worse than the participants who reported smoked infrequently; on the two tests that involved minimal stress on the circulorespiratory system (climbing and throwing), no significant differences were observed between participants who smoked heavily and participants who smoked infrequently or not at all.

The implication of this study is clear: Smoking impairs your ability perform well on fitness tests involving stress to the circulatory and respiratory systems. **Do you accept this conclusion? Or can the self-report bias serve as a suitable alternative explanation for the results that were obtained?**

17. Humor, Curiosity, and Verbal Absurdities: Regression to the Mean

What is humorous to some people is not humorous to others. But why does this variability exist? Several different reasons have been suggested, and surprisingly (at least to us), one of these explanations is based upon Freud's psychoanalytic theory. From this point of view, the person who frequently sees humor is "more able to think like a child, to escape the restraints of rationality and logic, and to feel secure enough to explore further into his/her environment." In the eyes of two researchers, this description seemed somewhat synonymous with the notion of being curious. And to test the validity of their thinking, a simple research study was conducted.

The researchers theorized that people with high levels of curiosity would be more likely to see things as humorous. However, in their study, they didn't want to deal with the messy problem of trying to assess whether or not someone thinks something is humorous. Therefore, the researchers

decided to measure people's ability to detect verbal absurdities, such as the sentence, “The woman picked up the melted ice cubes and dropped them into the pail.” The researchers reasoned that humorous events or statements often contain some sort of absurdity in them, and to be able to see absurdity one often has to think in a humorous manner.

The participants in this study came from a pool of 191 fifth-grade children. Curiosity scores on these subjects were derived from teacher ratings, peer ratings, and self-ratings (equally weighted to arrive at the final score). All ratings were based on data-gathering techniques possessing high reliability, and students receiving the very highest composite scores were those who, according to the ratings, showed interest in learning more about themselves, scanned surroundings for new experiences, and explored stimuli to find out more about them. Once the curiosity scores were tallied, the total participant pool was divided into upper and lower thirds. The students were not given information, however, about which subgroup they were in.

All 191 students were then administered a 51-item test of verbal absurdities, made up of 27 items (like the one presented earlier) wherein an absurdity was present, plus 24 items that were "normal" (like, "The woman hurried to school so she would not be late for her first class"). Participants read each sentence and placed a check next to any sentence that "seemed foolish." When the verbal absurdity scores for the two groups of participants were compared, it was found that the average score for the 51 pupils in the high-curiosity group (23.06) was significantly higher than the average score in the low-curiosity group (18.77).

Before concluding that their hypothesis had been supported, the researchers decided to try to rule out an alternative hypothesis. Perhaps the high-curiosity group were simply better readers than the low-curiosity group. That might explain why the students in the top third of the curiosity ratings were better able to identify the absurd sentences than were the students in the bottom third.

Therefore, the researchers administered an additional test to the high- and low-curiosity groups. The additional test, a 45-minute standardized reading test, showed some a slight difference between the two groups chosen previously on their rated curiosity, but not as big of a difference as the sentence absurdity test had shown. So, the researchers concluded that it was curiosity not reading ability that caused the previous, and larger, difference on the sentence absurdity test and that more curious people were more likely to see things as humorous.

Do you agree with this conclusion? If not, how might regression to the mean account for the three sets of results (the initially large difference in curiosity ratings, the subsequent large but not as large difference on the sentence absurdity test, and the much later and smaller difference on the later reading test)?

18. Growing Old: Cross-Sectional vs Longitudinal Effect

The theory of child rearing to which we subscribe is somewhat to the right of Mark Twain's. He believed that children should be put in a barrel and fed through the bunghole for the first 16 years—at which point he recommended plugging up the hole. Needless to say, this theory is not too popular with our families, and our children are starting to worry about the next ten years. Having just encountered the following study, we are starting to worry about the next thirty or so years ourselves.

The study is described in Wechsler's “The Measurement and Appraisal of Adult Intelligence” along with a number of similar studies on the relationship of age to intelligence in people 16 years of age and older. In this study, Wechsler Adult Intelligence Scale (WAIS) IQ scores are reported on more than 2000 participants. These participants were categorized by their age at the time of the IQ testing (16-17 years, 18-19 years, 20-24 years, and then in five-year blocks up to age 75 and over).

An inspection of the data and an accompanying graph leads to the conclusion that intelligence increases slightly from adolescence (where IQs average 103) up into the late twenties (where the

average IQ is 113), at which point a steady decline begins that brings IQ into the 90s between age 50 and 60 and then even lower after age 60.

After discounting alternative explanations such as a difference between the participants in the speed that they took the test (the study's data suggest there aren't such differences), a difference in the number of years out of school (and therefore less recent experience with test taking), and differences in the appropriateness of the tests for older persons, Wechsler reached the obvious conclusion that individuals are likely to lose a considerable amount of intelligence during their life span. Based on the findings, should you no longer trust those over 30? **Could the study's cross-sectional design be an alternative hypothesis?**