Self-Experimentation

Seth Roberts and Allen Neuringer

University of California at Berkeley and Reed College

Running head: Self-experimentation

Table of Contents

Self-Experimentation
<u>That you can learn how to do things by doing them has somehow always seemed mysterious to</u>
<u>me.</u>--Kermode (1995), p. 164
Introduction

By <u>self-experimentation</u> we mean experiments in which the researcher studies him- or herself. We contrast it with <u>conventional</u> research in which the experimenter studies other people or animals. This chapter tries to show that self-experiments are useful for gaining knowledge and solving problems, and that self-experimentation and conventional research have complementary strengths.

The earliest recorded self-experiment may be the work of Santorio Santorio, a seventeenth-century physician. He determined that the weight of his food and drink was usually more than twice the weight of his excretions, leading him to posit the existence of <u>insensible perspiration</u> (Castiglioni, 1931). Indeed, we sweat constantly in tiny amounts (Tokura, Shimomoto, Tsurutani, & Ohta, 1978). Since Santorio, many self-experimenters have been physicians interested in the causes and treatment of disease (Altman, 1972, 1987; Brown, 1995; Franklin & Sutherland, 1984). Early in the century, Joseph Goldberger ingested excretions of pellagra patients to show that pellagra was not contagious, and Werner Forssmann threaded a catheter to his heart through a vein in his arm to show the feasibility of the procedure (which eventually won him the 1956 Nobel Prize in Medicine). More recently, in 1984 Barry Marshall, an Australian doctor, drank a flask of water full of <u>Helicobacter pylori</u> bacteria to show that they cause ulcers. His theory of causation, now accepted as correct, was contrary to what most people believed at the time (Brown, 1995). Marshall's work began a new area of medical research–the role of bacteria in chronic disease ("Bugged by disease," 1998).

Some medical self-experiments have involved personal problems. In 1969, Richard Bernstein, an engineer with diabetes, started to measure his blood glucose several times/day. He discovered that it varied widely over a day, even though he was carefully following his doctor's recommendations. Both high and low glucose have bad effects. To reduce the variation, he began to do simple experiments. He discovered that many small doses of insulin, spread out over the day (similar to what the pancreas does for non-diabetics) maintained more stable glucose levels than one large daily dose of insulin, the usual prescription at the time. Lack of professional standing made it difficult for him to publicize his results, but he persisted and eventually his ideas spread. Glucose self-monitoring is now a $3-billion/year industry (Bernstein, 1984, 1990, 1994, personal communication, September 11, 1996), with products sold in every drugstore ("Blood-glucose meters," 1996).

In psychology, the best-known self-experiments are the memory studies of Ebbinghaus (1885/1913). Using lists of nonsense syllables as the material to be remembered, he measured speed of learning as a function of list length, retention as a function of time, and many other things. Conventional researchers have confirmed many of his conclusions (Cofer, 1979). One of Ebbinghaus's discoveries--that memory after sleep was better than expected from shorter retention intervals (Ebbinghaus, 1885/1913, pp. 76-77)--is still an active research topic (Barinaga, 1994). It is now well-accepted that sleep improves retention (Barinaga, 1994; Jenkins & Dallenbach, 1924).

The early history of psychology contains many other examples of self-experimentation (Neuringer, 1981). Thomas Young immobilized one of his eyeballs to test whether its movement was responsible for accommodation. This and other self-experiments led him to correctly conclude that accommodation is due to changes in the shape of the lens (Boring, 1942). Stratton (1897/1966) wore lenses that inverted the world both left-right and up-down and described his ability to adapt. Early behavioral psychologists, such as Thorndike (1900), joined early cognitive psychologists, such as Wundt (Bolles, 1993) and Titchener (1896), in reporting the results of experiments on their own thoughts, emotions, and behavior. More recent examples are Kristofferson's (1976, 1977, 1980) time-discrimination studies and Cabanac's (1995) work on thermoregulation. Psychophysicists often use themselves as subjects (e.g., Hecht, Schlaer, & Pirenne, 1942).

Self-experimentation has also been applied to practical psychological problems. The self-change literature contains many examples (e.g., Steinhauer & Bol, 1988; Watson & Tharp, 1993). Mahoney (1974, 1979) suggested that self-experimentation be used as a method of psychological treatment, recommending that "clients" be taught basic scientific methods, in order that the client become a "personal scientist."

Despite a long and productive history, self-experimentation is not now a major force in psychology. This is unfortunate, we argue here. Computers and other modern devices have made self-experimentation easier and more powerful than ever before, but quite apart from these advances, self-experiments can do many things more easily than conventional experiments.

In this chapter we try to show the value of self-experimentation, mainly through examples from our own research. The following "case studies" demonstrate the diversity of questions that self-experiments can help answer and some methods we have found worthwhile. Some of our examples were motivated by scientific interest, like Ebbinghaus's research, others by the desire to solve personal problems, like Bernstein's work. In the final sections of this chapter, we draw conclusions about self-experimentation in general.

Examples

Behavioral Variability (A.N.)

Variation in behavior is often useful. When a moth detects the ultrasound of a bat, the moth's flight path becomes highly unpredictable, helping the moth to elude the bat. Similarly, behavioral variation may be helpful when a person tries to solve a problem, to be creative, or to avoid an opponent in a game or battle.

Experimental psychologists have studied the limits of behavioral unpredictability by asking whether human subjects can generate random sequences of responses. Randomness implies that knowledge of prior history does not permit better prediction than no knowledge. Hundreds of studies have found that when people are asked to produce random sequences, e.g. of heads and tails, the resulting sequences can readily be distinguished by statistical tests from those produced by a random generator. Researchers have often concluded from these results that people are unable to behave randomly (Reichenbach, 1957; Tune, 1964; Waganaar, 1972).

There is a problem with this conclusion, however. Although people may commonly observe random events, they may never have needed to behave randomly. If you asked whether individuals had ever listened to a violin, most would answer "yes," but if you then provided a

violin and asked the same individuals to play, few would be able to make music. It would be wrong, of course, to conclude that people are unable to play the violin, because training is necessary. Perhaps training is also necessary for random behavior, a conjecture that led me (A. N.) to try to teach myself to behave randomly.

I entered the digits 0 through 9 on a computer keyboard as randomly as possible. At the end of each trial (consisting of 100 responses), I was shown on the computer screen a <u>random number generator score</u> (RNG), a measure of randomness used by conventional researchers (Evans, 1978). RNG assesses equality of dyads--how often "1" was followed by "1," "1" followed by "2," and so on for all possible pairs. If all dyads occurred with equal frequencies, the RNG score was 0. If responding was highly repetitive, RNG approached 1.0. The main result, shown in Figure 1, was that, over the 140 trials, RNG scores decreased. To assess the generality of the effect, a college student, M. S., received the same contingencies and feedback as I had, with similar results (Figure 1). These data showed that people can learn to improve one measure of the randomness of a response sequence.

One possible explanation for the lowered RNG scores with training involves memory for past behaviors--namely, that M. S. and I learned to avoid repetitions of response patterns. Consistent with this explanation was the finding that attention-competing tasks increased RNG scores (Evans & Graham, 1980). To test the memory explanation, I systematically varied how quickly I responded. The memory hypothesis predicted that the more slowly I responded, the less likely it was that I would remember my previous responses and the more likely, therefore, that patterns would be repeated. All contingencies were identical to the first experiment except that the time between each response (<u>interresponse time</u>, IRT) was controlled. Figure 2 shows that RNG scores <u>decreased</u> as a function of IRT between 0.3 sec and 7.5 sec, opposite to the prediction. To test this finding, I compared a 5-s IRT with a 20-s IRT in ABA fashion (3 replications of each value). RNG at 5 s IRT was .161 ± .006 (mean ± standard error) and at 20 s was .153 ± .008. Thus, slow responding increased response unpredictability, a result also found in conventional studies with people (Baddeley, 1966), as well as with animals (Neuringer, 1991). An extended discussion of these results would take us far afield from self-experimentation, but it appears that there are at least two strategies for behaving variably, one based on memory for prior events, another that appears to mimic a random generator, responses from which are relatively independent of history (Neuringer, 1986; Neuringer & Voss, 1993; Page & Neuringer, 1985).

One aspect of the above research was bothersome. When a random generator was programmed to generate 100 responses per trial, average RNG score was about .245, a value higher than many of the values in my experiments. Recall that high scores indicate <u>less</u> variability. The problem appeared to be that RNG was based on equality of dyads, and my dyads were more equal than expected from a random generator. One possible solution, and the one I followed, was to change the goal from that of minimizing a particular statistic to <u>matching</u> the output of a random generator according to many different statistics. A complete discussion is again beyond the scope of this chapter, but there is no single test of randomness (Knuth, 1969); many tests are needed to demonstrate approximations to random performance.

I therefore attempted to match a random generator on 6 different levels of the RNG metric. In the above experiments, RNG was based upon <u>contiguous</u> pairs of responses, what is

called "lag 1." Similar measures could come from responses separated by one response (lag 2), separated by two responses (lag 3), and so forth. At the end of each trial, I received feedback from 6 lags, my goal being to match the random generator at each of the lags. Table 1 compares my performance over the last 100 trials in a set of more than 300 trials with 100 trials from a random generator. At the end of training, I was generating RNG scores which did not differ significantly, across the 6 lags, from the scores produced by the random generator.

I later extended this work to as many as 30 different statistical tests, many used to evaluate the adequacy of random number generators (Knuth, 1969; Neuringer, 1986). The random generator was programmed to make 100 responses per trial (digits 0 through 9), and the data from 100 trials were then evaluated according to each of the 30 test statistics, with means and standard deviations calculated for each test statistic. Following each trial, I received feedback in terms of how I differed (in standard deviation units) from the random generator on each of these 30 statistics. Figure 3 shows an example of this feedback. At the beginning of training, most scores diverged greatly from those of the random generator, but by the end of 6 months of training, scores did not differ statistically from the random model according to these 30 statistical tests. Thus, a person can learn to approximate a random sequence of responses, as assessed by a large number of measures.

These results provide a challenge to the behaviorist's goal of prediction and control of instrumental behavior (Zuriff, 1985). Feedback can generate different levels of predictability, from highly predictable response rates in operant chambers to highly unpredictable performances. The results also suggest how behavioral variability may be engendered when creativity, problem solving, or learning of new skills is required (Holman, Goetz & Baer, 1977; Siegler, 1994; Stokes, 1995)

Commentary

An objection often raised to self-experimentation is that expectations bias results. However, as shown above, self-experiments often produce results that differ from expected. Another common objection is that the results may be relevant only to the single subject involved. However, when the last experiment was repeated with high-school and college students as subjects, and simpler feedback and procedure, the results were consistent with those from the self-experiments (Neuringer, 1986). Others have extended this line of research to animals, allowing the study of drugs, motivation, and genetic differences, including gender (Cohen, Neuringer & Rhodes, 1990; Machado, 1989; McElroy & Neuringer, 1990; Mook & Neuringer, 1994; Neuringer & Huntley, 1991). Thus, conventional studies have yielded results consistent with those from the self-experiments.

Thought, Memory, and Physical Movement (A. N.)

I tend to pace whenever I write a paper or prepare a lecture; and during long walks I seem to generate more novel ideas than at other times. These observations led me to ask if physical activity facilitates intellectual activity. I tried to generate new and interesting ideas (on any topic) under two conditions. In the sit condition, I sat quietly at my desk. In the move condition I walked, paced, swayed, or danced around a small room. The experiment consisted of 20 trials, 14 of these 15 min in duration each, the remainder 5 min each. Sit and move trials were presented in pairs--one sit and one move--with order counterbalanced and trial durations equal. Whenever a

new and interesting idea came to mind, I would stop and record it on a pad. (I stopped the trial-duration timer while writing.) I wrote more ideas during move (1.05/min) than during sit (0.72/min), a significant difference (Wilcoxon matched-pairs signed-ranks test, p < .01, 2-tail). Days after completing the trials, I subjectively judged the overall novelty and interest (one measure) of each idea. While judging, I was unaware of the condition in which the ideas had been generated (that information was coded on the back of each sheet of paper). Estimated novelty and interest was higher in the move condition than in sit, although the difference was not significant. Other experiments compared reading speed (I read 8% faster when moving than sitting, a significant difference), performance on the Miller Analogies test of intelligence (contrary to expectations, I performed <u>worse</u> while moving than sitting), and speed of learning to associate names with faces (see Neuringer, 1981).

To do this last experiment, pictures of people who worked for a large corporation were attached to one side of index cards with the individuals' names on the other. The cards were divided into 20 sets of 20 cards each, with 10 sets arbitrarily allocated to the move condition and the other 10 to sit. The task consisted of learning the names of the individuals until I was able to go through the set of 20 cards without any errors, and do this three times in a row. In the sit condition, I sat at my desk, taking care to have not engaged in strenuous exercise for the preceding few hours. To increase the amount of activity in the move condition, I now exercised (ran or swam) for about 30 minutes before each move trial, in addition to moving during the trial. Again, move and sit sessions alternated. Figure 4 shows that learning was facilitated by exercise and movement. In the move condition, each set of cards took about 7.5 repetitions to learn to perfection; in sit, 9.7 repetitions--again, a statistically significant difference (Wilcoxon matched-pairs signed-ranks test, two-tailed p < .05).

Commentary

Self-experimentation speeds up the feedback loop involved in developing methods. In the memory experiment, for example, I first learned names of flowers, then names of trees, and then French-English equivalents, all while developing the procedure. Because the experimenter serves as his or her own subject, many additional unexpected results may emerge. In the course of this research, it became clear to me that exercise enhanced my moods. A conventional study later supported the observation (Gurley, Neuringer, & Massee, 1984).

Weight (S. R.)

The setpoint theory of weight control assumes that amount of body fat is controlled by a feedback system (Hervey, 1969), similar to the thermostatic control of room temperature. The amount of fat that the regulatory system tries to maintain is the <u>setpoint</u>. When actual body fat is below the setpoint, the system acts to increase body fat in by increasing hunger and reducing metabolic rate. According to this theory, fat people have a higher setpoint than thin people.

Most weight-control researchers believe it is hard to change one's setpoint in adulthood (e.g., Gibbs, 1996, p. 91). But some studies contradict this conclusion. Sclafani and Springer (1976) found that adult rats allowed to eat unlimited amounts of "supermarket foods" (Sclafani & Springer, 1976, p. 461), such as cookies, salami, cheese, banana, and marshmallows, gained considerable weight relative to rats given unlimited lab chow. The supermarket rats gained weight two to three times faster than one would expect from a high-fat diet (A. Sclafani, personal

communication, July 2, 1996). Similarly, Cabanac and Rabe (1976) found that adult humans who consumed Renutril (a bland liquid food, like Metrecal) in place of their regular diet lost substantial weight, even though they could consume as much as they wanted. The two studies were done independently, yet both suggested the same conclusion: The tastiness of one's food controls one's setpoint. Tasty food (supermarket food, a normal human diet) produces a higher setpoint than bland food (lab chow, Renutril). If correct, this conclusion implies there is a way to lose weight without going hungry: eat less tasty food.

In 1993, I (S. R.) wanted to lose weight and decided to test this conclusion. I am 5' 11" (1.80 m) and at the time weighed 197 pounds (89.4 kg). I reasoned that what makes food tasty is processing (including home processing). Fruit juice tastes better than whole fruit. Cooked food tastes better than raw. So I reduced the amount of processing in my food. I stopped eating deli food, bread, sweets (e.g., scones), fruit juice, and fancy frozen food (e.g. Stouffer's Lean Cuisine), ate less meat and chicken (because meat and chicken are higher on the food chain than fish), and more fruits and vegetables. I ate mostly soups, salads, fish, steamed vegetables, rice, potatoes, and fruit.

Over 3 weeks, never going hungry, I lost 11 pounds (5.0 kg), which I have kept off without effort. At first the food seemed boring, but after a few days I came to enjoy my new diet and dislike my old one. The time course of my weight change is shown in Figure 5. (These measurements ended when the scale broke.) Later, two students (J. H. and A. L.) tried the same diet. They too lost weight; and when they returned to their original diets, they gained weight (Figure 5), providing more evidence that eating less-processed food produces substantial weight loss. Experiments in which subjects changed from a "modern" diet to an "indigenous" one support the same conclusion (O'Dea, 1984; Shintani, Hughes, Beckham, & O'Connor, 1991).

I eventually resumed measuring my weight, now using three scales. In 1996, an acquaintance told me that, starting at 190 pounds (86 kg) he had lost 45 pounds (20 kg) when he switched from an ordinary American diet to a diet of fruits, vegetables, rice, lots of water, and small amounts of fish and chicken (Tray Critelli, personal communication, June 5, 1996). The weight change was remarkably large, and his diet contained an unusual element: lots of water. To learn if water intake affects weight, I started to drink 5 liters of water/day, much more than my usual intake (about 1 liter/day).

I lost weight quickly, 7 pounds (3.2 kg) in 10 days, even though I always ate as much as I wanted (Figure 6). The results were consistent with the idea that drinking more water lowered my setpoint. Because I drank almost all the water between meals, it was unlikely that the weight loss occurred because the water was "filling," i.e., influenced satiety mechanisms that control meal length. It was not easy to drink that much water every day, so eventually I reduced my water intake from 5 to 3 liters/day. Figure 6 shows that soon after the change I gained about 3 pounds (1.4 kg).

These two observations–the effect of processing and the effect of water–were not easy to explain in conventional terms (amount of calories, amount of fat). Neither involved caloric restriction (eating less than desired). Drinking water did not change the amount of fat in my diet. Eating less-processed food probably reduced my fat intake, but the resulting weight loss was much more than the weight loss produced by a low-fat diet (Kendall, Levitsky, Strupp, &

Lissner, 1991; Raben, Jensen, Marckmann, Sandström, & Astrup, 1995; Sheppard, Kristal, & Kushi, 1991).

Around this time, I developed a theory of weight control (see below) that predicted that eating what I call slow calories--food with calories that are detected relatively slowly--would lower one's setpoint. In the United States, the main sources of slow calories are legumes (beans, peas, lentils). Hearing this prediction, a friend said her boyfriend had been much thinner in high school, when he ate a lot of beans and rice (Joyce Friedlander, personal communication, August 17, 1996). At the time of this conversation, my main sources of protein were fish and rice. The next day, I started eating legumes instead of fish. Again, I lost weight quickly for a short time (Figure 6). I lost about 6 pounds (2.7 kg) in 20 days, not counting 7 days out of town when I went off of the diet.

The theory of weight control that I developed, which explains these results, was inspired by the results of Ramirez (1990) with rats. Its main assumption is that one's setpoint is controlled by the strength of the taste-calorie associations in one's diet. Tastes become associated with calories when the two co-occur–when a taste signal (generated in the mouth) happens shortly before a calorie signal (probably generated in the stomach). Many rat experiments have shown that the taste of a food becomes associated with its caloric content (e.g., Bolles, Hayward, & Crandall, 1981). My theory assumes that the brain keeps a running average of the calories associated with the tastes of one's food. The greater the value--the more strongly the tastes in the diet signal calories--the higher the setpoint (more body fat). This makes evolutionary sense: When calories are relatively abundant, we should try to stockpile more of them (via body fat) than when they are scarce.

To explain the effect of processing (Figure 5), this theory makes an additional assumption: The more strongly a food's taste is associated with calories, the better the food tastes–in behavioral terms, the more likely you will choose that food if given a choice. This assumption is supported by rat experiments that used preference tests to show the existence of taste-calorie associations (e.g., Bolles, Hayward, & Crandall, 1981). Pairing a taste with calories increased choice of the taste.

Processed food is generally preferred to the same food unprocessed (as revealed, for instance, by willingness to pay more for the processed food). For example, most people prefer roasted nuts to raw nuts, orange juice to oranges, food with spices to food without spices. Probably the main reason that processing makes food taste better is that it strengthens taste-calorie associations. The details of common processing methods support this argument. The strength of taste-calorie associations, like other examples of Pavlovian conditioning, depends on (a) conditioned-stimulus (CS) intensity, i.e., taste intensity, and (b) unconditioned-stimulus (US) intensity, i.e., calorie intensity. Many processing methods plausibly increase CS intensity or complexity: Adding spices, sauces, flavorings, or small amounts of fat (many flavoring agents are fat-soluble). Many other processing methods plausibly increase US intensity, the amount of calories detected soon after the taste: Cooking, mashing, adding large amounts of sugar or fat. Adding fat or sugar adds calories, but cooking and adding spices do not. Thus the effects of processing cannot be explained just in terms of calories. Processed food, in other words, produces stronger taste-calorie associations than the same food unprocessed, even when the

calorie content is unchanged.

Water reduces weight (Figure 6) because its taste is associated with zero calories. So it lowers the running average the brain uses to judge the abundance of calories. Legumes reduce weight (Figure 6) because they produce a relatively slow calorie signal, an assumption based on the relatively slow rate at which they raise blood glucose (Foster-Powell & Brand-Miller, 1995; the technical term for this measurement is glycemic index). Thus the taste signal and the calorie signal produced by legumes are relatively far apart in time, reducing their association.

In summary, this work suggests three ways to lose weight without going hungry: eat less-processed food; consume more water or any other source of taste and few calories (tea, pickles, low-calorie soup); and eat more legumes.

Commentary

Self-experimentation was helpful in several ways. It connected laboratory and real life. Self-experiments are often more realistic than conventional experiments. The Sclafani and Springer experiments (rat subjects, supermarket food) and the Cabanac and Rabe experiment (human subjects, liquid food) suggested that eating less-tasty food lowers the setpoint, but did so in situations remote from everyday life. My self-experiment about processing (Figure 5) showed the practical use of these discoveries. The Ramirez experiments (rat subjects, liquid food, effect of saccharine) pointed to a new mechanism of weight control, but gave no indication of the importance of that mechanism in the real-life control of human weight. My results suggest it plays a large role because it explained and predicted large effects (Figures 5 and 6). It was a good precursor to clinical trials because I tested treatments (diets) on myself before I tested them on others. Medical self-experiments have often served this purpose (Altman, 1987). Finally, it helped me lose weight. I lost–apparently for good--about 17 pounds (8 kg). I never went hungry, took no drugs, and ended up eating a healthier diet (more fruit, vegetables, fiber). Unlike many people who lose weight, I know the crucial features of what worked; maybe that helps avoid backsliding based on wishful thinking ("it won't matter if I . . . ").

<div align="center">Sleep (S. R.)</div>

I used to suffer from what is called early awakening. I awoke at 3 or 4 a.m., still tired but unable to sleep. Not until a few hours later would I be able to fall back asleep. About 15% of U. S. adults have this problem (Gallup Organization, 1991).

In the 1980's, I began to try to solve the problem by self-experimentation. The department electronics shop made a small device that helped record when I slept. I pushed one button when I turned off the lights to try to fall asleep, another button when I got out of bed. I defined an instance of early awakening to be a morning when I fell back asleep between 10 minutes and 6 hours after getting up. The lower limit (10 minutes) was meant to eliminate trivial awakenings (e.g., getting up to urinate), the upper limit (6 hours) to eliminate afternoon naps. My life was well-suited for sleep experiments. Because the hours I worked were flexible, I never used an alarm clock and almost always could try to fall back asleep if I wanted to. It also helped that I lived alone.

In spite of such favorable conditions, for years I made little progress. I tried many treatments, most involving morning light. Some helped, but none eliminated early awakening. At best, I awoke too early on a third of all days. All my ideas about the cause(s) of early awakening

were apparently wrong, but it was not clear how to find better ideas. I was stuck.

In January 1993, during routine data analysis, I looked at a graph of my daily sleep duration (how long I slept each day, including naps) as a function of day. Figure 7 is an updated version of what I saw. The graph showed a sharp drop in sleep duration during 1992. The drop (about 40 minutes) occurred at the same time I lost weight by eating less-processed food (Figure 5; smoothing makes the sleep decrease appear to start before the weight loss).

The coincidence of these events (diet change, weight loss, less sleep) suggests that either weight or diet controls how much sleep we want or need. Other data suggest that it is weight, not diet, that is crucial because diverse ways of changing weight have similar effects. Loss (or gain) of weight has been associated with less (or more) total sleep time (Crisp, Stonehill, & Fenton, 1971; Crisp, Stonehill, Fenton, & Fenwick, 1973; Lacey, Crisp, Kalucy, Hartmann, & Chen, 1975; Neuringer, 1981). In surveys, less weight has been associated with less total sleep time (Paxton, Trinder, Montgomery, Oswald, Adam, & Shapiro, 1984; Shephard, Jones, Ishii, Kaneko, & Olbrecht, 1969; Walsh, Goetz, Roose, Fingeroth, & Glassman, 1985). A connection between body fat and sleep makes functional sense. Sleep is a luxury, a kind of vacation: While you sleep, you use more calories than you take in. The greater your wealth (in terms of body fat), the more sleep you can afford.

I showed my students a graph similar to Figure 7. It inspired one of them to tell me, a few weeks later, that eating a diet high in water content (e.g., fruit) had reduced how much sleep he needed (Michael Lee, personal communication, March, 1993). This was fascinating. I tried the diet he suggested, at first eating four pieces of fruit per day. My sleep duration did not change. When I told him my results, he said, "I eat six pieces of fruit per day." So I started eating six pieces of fruit each day. To do this, I had to change my breakfast–there was nowhere else to put the extra fruit. Instead of eating oatmeal for breakfast, I started eating two pieces of fruit, often a banana and an apple.

My sleep duration remained about the same. However, after about a week of fruit breakfasts I noticed that my early awakening had gotten worse–I woke up too early every morning instead of a third of all mornings. Was this a coincidence? I alternated between fruit and oatmeal breakfasts and established it was cause and effect. A fruit breakfast made early awakening the next morning much more likely than an oatmeal breakfast.

The connection was a total surprise; when I awoke at 3 or 4 a.m., I did not feel hungry. Yet I had known for years about laboratory observations of food-anticipatory activity, a well-established effect in animals (Bolles & Stokes, 1965; Boulos & Terman, 1980). Mammals, birds, and fish become more active shortly before the time of day that they are fed (Boulos & Terman, 1980). In spite of its cross-species generality, the effect had never been related to human behavior, as far as I know. The animal results made a cause-and-effect relation between breakfast and early awakening much more credible. However they did not make clear how to reduce early awakening.

I had been eating oatmeal for breakfast for reasons unrelated to sleep. Because a randomly-chosen breakfast was unlikely to be optimal, some other breakfast would probably produce less early awakening. Oatmeal produced less sleep disturbance than fruit, and the obvious nutritional difference is that oatmeal has more protein. This led me to try a variety of

high-protein breakfasts, but always with the same result: I continued to wake up too early quite often.

   Mistlberger, Houpt, and Moore-Ede (1990), studying rats, found that carbohydrate, protein, and fat can each produce anticipatory activity. I learned of these results during my study of different breakfasts, and they made me realize that I needed to consider more than just protein in my search for a better breakfast. That was not easy. Food can be described on many dimensions (calories, fat, cholesterol, sugar, etc.) and I had no idea which were important.

   My experimental comparisons had contrasted one breakfast (e.g., oatmeal) with another (e.g., fruit). To make interpretation easier, I decided to compare something (one breakfast) with nothing (no breakfast). I began with a baseline of no breakfast (nothing to eat or drink before 11 a. m.). The result: My early awakening disappeared! It went away gradually, during the first week of no breakfast. To check that the absence of early awakening was due to the absence of breakfast, I started eating breakfast again (one piece of fruit between 7 and 8 a.m.). Early awakening returned. When I stopped eating breakfast again, early awakening disappeared again. These changes also affected how I felt: When I ate no breakfast, I woke up feeling more rested.

   Figure 8 shows the results of this ABA experiment in detail. It plots the probability that I had fallen back asleep as a function of time since getting up. The oatmeal function is from the 300 days before I started varying my breakfast. The none–1st function is from the first period when I ate no breakfast, omitting the first eight days (when the treatment was gradually taking effect). The fruit function is from the days when I ate one piece of fruit for breakfast. The none--2nd function is from the beginning of the second block of days when I ate no breakfast, omitting the first two days. As Figure 8 shows, skipping breakfast greatly reduced early awakening.

   Details of my results resembled food anticipation in rats. In rats, anticipatory activity begins a few hours before food when large amounts of food are given (e.g., Bolles & Stokes, 1965). During the fruit phase of my experiment, I ate between 7 and 8 a. m., and I awoke at the average time of 5:35 ± 0:11 a. m. (10% trimmed mean ± standard error). Rat experiments have found that when food stops, the anticipatory activity gradually disappears over the next 5-10 days (Boulos, Rosenwasser, & Terman, 1980). When I stopped eating breakfast, waking up too early took about 8 days to nearly disappear. Calling the first morning with no breakfast Morning 1, I woke up too early Mornings 2, 4, 5, 6, 9, and much less often after that. These similarities between rat and human results make the human results more credible.

   Although skipping breakfast reduced early awakening, it did not eliminate it. During the two no-breakfast periods of Figure 8, I awoke early on 12% of the mornings. This suggested that early awakening had more than one cause. More evidence for a second cause was that the absence of breakfast changed the "latency" of falling back asleep (the time between when I got up and when I fell back asleep). It was less during the fruit phase (2.1 ± 0.1 hours) than during the two no-breakfast phases (3.3 ± 0.2 hours). When one cause (breakfast) was removed, the existence of another cause--which produced falling back asleep with a different latency– became apparent.

   The data of Figure 8 come from September 1992-May 1994. During the following months, early awakening became much more frequent, eventually happening on about 60% of mornings, even though I never ate before 10 a.m. I had no idea what caused this increase, but it

was more evidence that early awakening had more than one environmental cause, if my conclusions about breakfast were correct.

In February 1995, searching for other causes, I started watching TV every morning. Although this had a large effect on my mood (see below), it had little effect on the probability of early awakening, in spite of great expectations and considerable trial and error. The lack of change was an especially clear indication that expectations and hopes did not substantially influence the results.

A second solution to the problem turned up by accident, like the first solution (no breakfast). Asking friends about weight control, I had heard two anecdotes in which a person who walked much more than usual (many hours/day) lost significant weight. Perhaps it was cause and effect: Walking a lot caused weight loss. Walking combines movement with standing (placing all your weight on your feet). Either might have caused the weight loss.

I could not walk many hours/day but could stand many hours/day if I stood while working at my desk. This lifestyle change interested me partly because I assumed that our hunter-gatherer ancestors usually stood much more than I did (about 3-4 hours/day), and my breakfast and mood results had suggested that solutions to psychological problems can often be found in aspects of Stone Age life. On August 27, 1996 I began to stand much more. I raised my computer screen and keyboard so that I wrote standing up, stood during phone calls, walked more, bicycled less. At first this was exhausting but after a few days I got used to it. I used the stopwatch on my wristwatch and a small notecard to keep track of how long I stood each day. I included any time that all my weight was on my feet: standing still, walking, playing racquetball. My weight did not change. Within a few days, however, it became obvious that I was waking too early much less often. During the three months before August 27, I had woken up too early on about 60% of days; during the first few months after August 27, on about 20% of days.

How long I stood each day varied, mostly because of variation in events during which I had to sit (meals, meetings, etc.). After I noticed the effect of standing, I initially assumed that any substantial amount of standing (e.g., 6 hours) would be enough to reduce early awakening. However, in October 1996 I analyzed my data in preparation for a talk. The analysis suggested that the amount of standing necessary to get the maximum benefit was much more than I had thought (Table 2). Standing from 5 to 8 hours had little effect on early awakening, judging from the pre-treatment baseline (before August 27); standing from 8.0 to 8.8 hours reduced early awakening, but did not eliminate it; and standing 8.8 hours or more eliminated early awakening . After noticing this, I tried to stand at least 9 hours every day. This completely solved the problem (Table 2). After a day during which I stood at least 8.8 hours, I almost never awoke too early. Although the data of Table 2 are correlational, the correlation implies causation because the correlation is strong, long-lasting, unexpected, and seems to have no other plausible explanation.

That standing affects sleep makes functional sense. The muscles we use to stand no doubt did more work in an average Stone-Age day than any other muscles. Because we sleep lying down, sleep time can be used to do routine maintenance on these muscles. And if these muscles were shaped by evolution to take advantage of sleep for maintenance, then they will need sleep for maintenance. The more use during the day, the more maintenance needed at night. So we will need a system that makes us sleep more than usual after we have stood more than usual. At the

time our sleep-controlling system evolved into the form it now has, people probably stood many hours/day. Without the pressure to sleep provided by considerable standing, sleep is not deep enough.

Figure 9 summarizes seven years of research . The wiggly line in the upper panel is a smooth of the data, a moving average based on 31 points (the 15 days before the target day, the target day, the 15 days after the target day). The wiggly line in the lower panel indicates the probability of early awakening (i.e., the density of points in the upper panel); each point on the wiggly line is based on the 31 neighboring days. Smoothing should be part of the analysis of any long time series (Tukey, 1977), which self-experimentation sometimes generates. It often reveals previously-unnoticed structure. Figure 9 provides three examples. (a) In the upper panel, the smoothed latency rises from about 1.5 hours to 3 hours starting when morning TV began. It is only a correlation, but it raises the possibility that morning TV affected the mechanism(s) that cause early awakening, even though morning TV had no clear effect on the probability of early awakening (lower panel). (b) The function in the lower panel suggests there was a yearly rhythm in early awakening--more frequent during the summer. (c) The lower panel also shows that standing many hours was associated with a more sustained reduction in early awakening than ever before--more evidence for the power of standing. Early awakening remained very rare from the end of the period covered by Figure 9 until the time this is written (March 1998).

Commentary

These self-experimental results challenge some widely-held beliefs. That skipping breakfast was helpful contradicts the popular idea that breakfast is "the most important meal of the day" (Bender, 1993, p. 488). (Bender [1993] found no clear support for the popular view.) Most sleep researchers believe that light and time awake are the main environmental events that control when we sleep (e.g., Borbély, 1982; Moore-Ede, Sulzman, & Fuller, 1982). This work found that breakfast and standing can also have powerful effects. The effect of breakfast is just an extension of animal research on food-anticipatory activity (e.g., Boulos & Terman, 1980) but the effect of standing seems to be without precedent.

A common criticism of self-experimentation, mentioned earlier, is that the experimenter's expectations and desires may shape the results. When I first began self-experiments to reduce early awakening, I worried about this possibility. I could consciously control whether or not I fell back asleep after getting up– mainly, by deciding whether or not to lie down. But the longer the research continued, the less I worried. Because I failed for years to find a solution (in spite of wanting to), because a solution I strongly expected to work (morning TV) did not, and because the solutions I eventually found (no breakfast , standing a lot) were unexpected, it became clear that desires and expectations had little effect.

Serendipity, often important in conventional research (Skinner, 1956; Siegel & Zeigler, 1976), played a large role. I noticed the connection between breakfast and early awakening while trying to sleep less; I noticed the connection between standing and early awakening while trying to lose weight. Yet these discoveries were not accidental, because the lifestyle change that made the difference was no accident (e.g., I stood more on purpose). Self-experiments lend themselves to this sort of discovery because they implicitly measure many things at once. Even when focused on one measure (e.g., weight), you can easily notice if other measures change. The next section

describes another example of serendipity.

<div align="center">Mood (S. R.)</div>

Because early awakening persisted after I stopped eating breakfast, breakfast was not its only cause. Trying to think of other possible causes, I realized there might be a general lesson to be learned from the effect of breakfast. Our brains were shaped by evolution to work well during long-ago living conditions. Our Pleistocene ancestors, I believed, could not regularly eat a rich breakfast soon after waking up, but I could–and a rich breakfast caused early awakening. Maybe other "unnatural" aspects of my life also caused early awakening.

Based on studies of people living with no time-of-day information (e.g., in caves), Wever (1979) concluded that human contact affects the phase of an internal clock that controls when we sleep. Wever's evidence for this conclusion could be interpreted in other ways, and later commentators have been skeptical (e.g., Moore-Ede, Sulzman, & Fuller, 1982). However, Wever's conclusion makes evolutionary sense. An internal clock "set" by human contact would tend to make us awake when other people are awake, just as the internal clock that causes food anticipation tends to makes us awake when food is available. Our Stone-Age ancestors lived in groups, of course, and field studies of technologically primitive cultures (e.g., Chagnon, 1983) suggest that our ancestors had a great deal of contact with other people every morning. In contrast, I lived alone and often worked more or less alone all morning. Maybe lack of morning human contact caused early awakening.

In 1964-66, an international survey of time use found that adults in the United States stayed awake an hour later than adults in each of eleven other countries (Szalai, 1973). In every country except the United States, adults fell asleep about 11 p.m.; in the United States, they fell asleep about midnight. Only one other activity measured in the survey differed so dramatically in its timing between the United States and other countries–television watching. Americans watched TV an hour later than everyone else. Watching TV resembles human contact in many ways. The survey results raised the possibility that the aspects of human contact that control when we sleep could be supplied by TV.

In February 1995, this reasoning led me one morning to watch about 20 minutes of TV soon after I awoke--specifically, a tape of the Leno and Letterman monologues (resembling what Americans were watching at 11 p.m.). There was no obvious effect, and I fell back asleep about an hour later. The next morning, however, soon after I awoke I felt exceptionally good--cheerful, refreshed, relaxed, energetic. I could not remember ever feeling so good early in the morning. We do not usually attribute how we feel in the morning to whether we watched TV the previous morning. Yet the prior days had been ordinary in every way, except for the morning TV. Over the next few weeks, I watched morning TV some days and not others and became convinced that the morning TV/next-day mood correlation I had observed reflected causation. Whether I watched TV one morning (e.g., Monday) clearly affected how I felt the next morning (Tuesday), despite having no noticeable effect on my mood while watching or soon afterwards and being drastically different from what we usually think controls happiness.

As mentioned above, my first detailed study of TV effects tried to find an arrangement that eliminated early awakening. Although I was not carefully measuring my mood, I noticed that several shows I wanted to watch did not improve my mood the next day--in particular, The

Simpsons, The Real World, documentaries, and the O. J. Simpson criminal trial. In contrast, stand-up comedy seemed to work fine. These observations suggested that the crucial stimulus was a reasonably large face looking more or less straight at the camera (i.e., both eyes visible). The Simpsons had no real faces. Documentaries and The Real World rarely showed a face looking at the camera. The Simpson trial had plenty of faces but almost always in profile. Stand-up comedy is usually a person looking at the camera. I also found that playing racquetball (i.e., one form of actual human contact) was not effective. During racquetball, you rarely see your opponent's face. Apparently the visual aspect of human contact (seeing a face) was far more important than the auditory aspect (hearing a voice), the cognitive aspect (decoding language, thinking about people), or the emotional aspect (feeling happy, sad, etc.), all of which the ineffective stimuli (The Simpsons, and so on) provided.

The conclusion that the effective stimulus was a frontal view of a face led me to measure face time, defined as the duration of faces at least 2 inches (5 cm) wide with two eyes visible. (I watched TV with my eyes about 40 inches [1 meter] from a 20-inch [51 cm] TV.) I used 2 inches (5 cm) as the minimum width because it was slightly less than a stimulus that I knew was effective--the average width of Jay Leno's face during his monologue. I kept track of face time using a stopwatch. After I realized the importance of faces, I usually watched a mix of shows with a high ratio of face time to total time (Charlie Rose, Charles Grodin, Rivera Live, much of The News Hour with Jim Lehrer) and shows that I liked more but which showed fewer large faces (Friends, 60 Minutes). I watched everything on tape, of course.

Eventually I stopped trying to reduce early awakening with morning TV and instead studied the mood effect more carefully. To quantify it, I measured my mood at about 11 a.m. each day on three 0.5-9.5 scales (Table 3). The three scales--happiness, serenity, and eagerness--reflected the three most obvious ways that watching TV in the morning seemed to make a difference: I felt happier, more serene (less irritable, less easily upset), and more eager to do things. The three measures were almost always close (happiness and eagerness about equal, serenity about 1 point higher) so I give only their average. (Later work found that each measure could change independently from the other two, so it was a good idea to measure all three.)

Figure 10 shows the results of an experiment done to show the basic effect. During the first phase, a baseline, I watched TV every morning (40 minutes of face time, starting at 7 a.m.). When I was sick for a few days, I did not record my mood. During the second phase, I did not watch TV. My mood got worse, but the change happened a day after the change in treatment. The final phase was a return to baseline: I watched TV every morning, just as in the first phase. My mood improved but again it took at least a day for the change to take place.

How much TV was needed to get the maximum effect? Figure 11 shows the results of conditions done to answer this question. (It includes the data shown in more detail in Figure 10.) I tested 0, 20, 30, and 40 minutes of face time in the order shown in Figure 11 (30, 40, 0, etc.), each for several days. Twenty minutes of face time produced a better mood the next day than 0 minutes ($t[8] = 3.8$, one-tailed $p < .005$). (I did statistical tests by averaging the three mood ratings for each day to get a single number, then treated that number as an independent random sample.) Thirty minutes was more effective than 20 minutes ($t[13] = 2.8$, $p < .01$); 30 and 40 minutes produced similar effects.

Figure 12 shows the results of conditions run to learn more about what controls the effect. I interspersed blocks of baseline days with blocks of days during which I changed the baseline treatment in various ways. The baseline treatment was watching enough TV to get 30 minutes of face time, starting at 7 a.m., with my eyes about 40 inches [1 m] from the TV screen. The four baseline conditions of Figure 12 did not differ reliably, $F(3, 21) = 1.7$. Watching TV one hour later lowered mood, $t(29) = 6.4$, $p < .001$, comparing the "one hour later" results to the combined baseline results. To my surprise, watching TV one hour earlier also lowered mood, $t(28) = 3.7$, $p < .001$. Watching TV twice as far from the screen (80 inches [2 m] rather than 40 inches [1 m]) lowered mood, $t(29) = 8.3$, $p < .001$.

In order to obtained the desired amount of face time (e.g., 30 minutes), the total time I watched TV each morning varied with what I watched. Watching 10 minutes of Charlie Rose yielded about 8 minutes of face time; watching 10 minutes of Friends yielded about 2 minutes of face time. Was the difference between Charlie Rose and Friends important? Soon after the results of Figure 12, I looked at a scatterplot of some of the data. I computed face density, defined as

$$\text{face density} = \text{face duration/total duration},$$

for each TV session. For instance, if it had taken 50 minutes of TV to accumulate 30 minutes of face time, face density = 0.6. I plotted mood the next day versus face density (Figure 13). I used all of the "baseline" data I had collected at that time--all days with 30 or 40 minutes of face time, 7 a.m. starting time, and 1 m viewing distance. Figure 13 suggests that density or something correlated with density made a difference. Except for two outliers, greater face densities were associated with higher next-day mood.

Did the correlation shown in Figure 13 reflect causation? Table 4 gives the results of an ABA experiment done to find out. During Phases 1 and 3, face density was relatively high; during Phase 2, it was lower. I manipulated density by varying what I watched (always on tape); to increase density, I watched more Charlie Rose, Charles Grodin, etc. I kept face time constant across conditions. The results showed that density (or some correlate) made a difference. My mood ratings were higher during Phases 1 and 3 than during Phase 2, $t(15) = 6.7$, $p < .001$.

A possible explanation of the density effect of Table 4 is that the potency of faces declines with time, i.e., that faces soon after 7 a.m. have more effect than later faces. As this explanation predicts, watching one hour later lowers mood (Figure 12). If this explanation is correct, then the results of Figure 11 could be misleading. Because 30 and 40 minutes of face time had the same effect, the results of Figure 11 seem to imply that 30 minutes of face "saturates" the system. However, the difference between 30 and 40 minutes--the additional 10 minutes--may have had no effect because it happened too late. Maybe it would have made a difference if it happened earlier. The last row of Table 4 shows the results of a test of this idea. Indeed, 50 minutes of dense face produced higher mood ratings than 30 minutes of dense face, $t(18) = 4.7$, $p < .001$, in contrast to the results of Figure 11. (Later work found that 70 minutes of dense face produced higher ratings than 50 minutes.) The density by duration interaction supports the idea that faces more than a few hours after 7 a.m. have little effect. Another possible explanation of the density effect of Table 4, not yet tested, is that density was confounded with

face size (e.g., faces on <u>Charlie Rose</u> are larger than faces on <u>Friends</u>) and that face size makes a difference.

The parametric results (Figures 11-13, Table 4) helped show how to get a large effect and pointed toward an explanation of the basic effect (Figure 10). Faces (or part of them) were surely the crucial feature. This conclusion is supported not only by my informal observations (e.g., that <u>The Simpsons</u> was ineffective) but also by other results. That distance from the screen mattered (Figure 12) implies that the crucial stimulus is visual, a conclusion supported by two additional findings: (a) Size of TV mattered. A 27-inch TV produced a higher mood than a 20- inch TV or a 32-inch TV. Full-screen faces on a 27-inch TV are closer to life-size than on the smaller and bigger TVs. (b) Angle of view mattered. Looking down at the TV (at a 35° angle) produced a lower mood than looking straight ahead. Any theory that assumes faces are <u>not</u> the crucial stimulus will have difficulty explaining the finding that under some conditions <u>decreasing</u> the total duration of TV increased the effect (Table 4). The best stimulus seems to be close to what you see during an ordinary conversation.

Also impressive is the importance of time of day. One hour is a small fraction of a day, yet a one-hour change in time of exposure to TV, in either direction, made a clear and consistent difference (Figure 12). This suggests that the TV acts on an internal mechanism that changes considerably and consistently from one hour of the day to the next (e.g., from 7 a.m. to 8 a.m.). The obvious candidate for such a mechanism is an internal circadian clock.

Thus the results point to the existence of an internal circadian oscillator that is (a) sensitive to faces and (b) controls mood. This theory predicts there will be large-amplitude circadian rhythms in happiness, serenity, and eagerness. (There will be a circadian rhythm of some size in almost any biological measure--what is interesting is finding a <u>large</u> circadian rhythm.) Many conventional studies have observed circadian variation in mood (Boivin et al., 1997), albeit relatively small. Hundreds of times I measured my mood throughout the day and found what the theory predicts. My average mood was low in the early morning (roughly 4 soon after waking up, at 4-6 a.m.), rose to a maximum (about 8.5 under optimal conditions) from about noon to 4 p.m., and declined sharply after that to below 5 around 9 p.m. (I fell asleep 10-11 p.m.).

The existence and properties of this clock make evolutionary sense. As mentioned earlier, people who live together should be active at the same time. This clock tends to produce such synchronization. It obviously controls the timing of sleep, and its sensitivity to conversational faces makes us tend to be awake at the same times of day we have conversations. I did not study the effect of TV on when I slept because the effect was so clear. When I stopped watching the 11:00 p.m. news, for example, I started falling asleep an hour earlier.

That this clock controls several dimensions of mood also makes evolutionary sense. Eagerness--a desire to do things--is helpful during the day but harmful at night, because activity at night will prevent you from falling asleep and may keep your neighbors awake. A daily rhythm in serenity helps protect sleep. If you are irritable when awakened, others will try to avoid awakening you (good). But if you are irritable during the day, others will avoid you during the day (bad). The function of a daily rhythm in happiness is less obvious. Perhaps a clock-controlled lowering of happiness makes problems more urgent. Suppose that problems (e.g., hunger) reduce

happiness but only when your level of happiness goes below neutral–you become unhappy--do you take strenuous action to solve the problems (e.g., get food). (In agreement with this idea, the term happy is sometimes used to mean satisfied: "Are you happy?"). A clock may reduce happiness as bedtime approaches because existing problems become more urgent at that time--if not solved, they will interfere with sleep.

My results raise the possibility that we have an internal clock (or clocks) that, to work properly, needs (a) daily exposure to faces (30 minutes or more) in the morning and (b) non-exposure to faces at night. If so, many people have malfunctioning clocks, because many people get too little exposures to faces in the morning and/or too much exposure to faces at night. If this clock controls both sleep and mood, we should see many cases where the two are disrupted simultaneously. In fact, simultaneous disruption of sleep, happiness, serenity, and eagerness is a good description of depression, the mental disorder. An oscillator can malfunction with either the wrong phase or low amplitude. Wrong phase will cause trouble falling asleep; low amplitude will cause difficulty staying asleep. Both problems are common features of depression (American Psychiatric Association, 1994, Diagnostic and statistical manual of mental disorders, 4th ed.; Brown & Harris, 1978), and insomnia is correlated with depression (Soldatos, 1994). Wrong phase and lowered amplitude of the happiness, serenity, and eagerness rhythms would cause someone to be less happy, more irritable, and less eager to do things than usual (at least, while awake). Lack of eagerness to do things is the core symptom of depression (DSM-IV), and lack of happiness and irritability are common concomitants (DSM-IV; Brown, & Harris, 1978). (They are not inevitable concomitants, however, suggesting that depression has other causes, too. For example, this theory does not explain Seasonal Affective Disorder.) Many other facts about depression also link it to circadian rhythms (Wehr & Goodwin, 1983; Van den Hoofdakker, 1994). Several theories have assumed that a circadian-rhythm disturbance is the source of depression (e.g., Ehlers, Frank, & Kupfer, 1988; Kripke, 1984; Van Cauter & Turek, 1986; reviewed by Van den Hoofdakker, 1994) but this is the first to suggest the crucial role of exposure to faces.

Commentary

This is a good illustration of how self-experimentation and conventional research can work together. My self-experiments were stimulated by the conventional research of Wever (1969) and Szalai (1973). In the area of circadian-rhythm research, the study of social zeitgeibers (a zeitgeber is an environmental event, such as light, that entrains an internal oscillator) is moribund; The Journal of Biological Rhythms, which began publication in 1986, has, as of early 1998, never carried an article on the topic. Several years ago, psychiatrists who were not circadian-rhythm researchers proposed that depression was due to disruption of "social rhythms" (Ehlers, Frank, & Kupfer, 1988, p. 948). They had a good idea, but could not effectively follow it up; working with a clinical population (depressives), it would have been extraordinarily hard to do experiments to isolate the crucial ingredient of social rhythms–experiments that presumably would have found that the sight of life-size faces in the morning, at a conversational distance, for 30 minutes or more, is the potent ingredient. That the effect of faces is apparent only a day later would have made the research even more difficult; most depressed patients are not hospitalized, so they are almost never tested two days in a row. Self-experiments not only made it much easier

to discover the effect of morning faces on next-day mood, they also made it much easier to explore the "parameter space" of the effect, the many procedural dimensions that might influence its size.

On the other hand, self-experimentation can only <u>raise</u> the question of whether lack of morning faces (and/or exposure to faces at night) is an important cause of depression; it can do little to answer it. For that, conventional research with depressed subjects is needed.

<u>Summary</u>

_____ The examples show that self-experimentation can take many forms. The goal may be mainly scientific (A. N.'s randomness work) or highly practical (S. R.'s sleep research). The setting can be a laboratory (A. N.'s randomness work) or the real world (S. R.'s sleep research). It may start with an apparent solution to a problem (S.R.'s weight research) or may find solutions only after much trial and error (S. R.'s sleep research). A series of experiments may emphasize the various effects of an interesting treatment (A. N.'s movement work), study in detail one effect of one treatment (S. R.'s mood research), or test a wide range of treatments (S. R.'s weight and sleep work). Treatments may be familiar (feedback) or novel (standing 9 hours/day), may involve small changes in lifestyle (one of our students found that her acne disappeared when she stopped using soap to clean her face [Neuringer, 1981]) or large ones (standing 9 hours/day). The measures may be objective (A. N.'s randomness work) or subjective (S. R.'s mood work). Data collection may last weeks (many of our students do short-term self-experiments) or many years (S. R.'s sleep research). Self-experiments can explore a completely new area or test pre-existing beliefs about a familiar one. Some of the examples (behavioral variability, cognitive effects of movement) helped confirm the experimenter's hypotheses, more or less. In other cases (sleep, mood), the main results were a complete surprise.

Within psychology, Ebbinghaus is the best-known example of self-experimentation. However, our examples show that some of the less attractive features of Ebbinghaus's work--the tedious nature of the measurements (learning and relearning lists), the artificial nature of the memorized items (nonsense syllables)--do not describe all self-experimentation. Most experiments contain artificial aspects, usually to reduce variation (i.e. noise). Yet many of our examples involved no artificial elements--the experiments studied exactly the measure of interest under natural conditions (e.g., S. R.'s studies of sleep and mood). Self-experimentation can be sufficiently powerful to handle the variability of everyday life.

Above all, the examples show that self-experimentation can generate valuable data and theory that would be hard to get in other ways. Most of the topics we studied had been the subject of a great deal of conventional research, yet the self-experiments uncovered new, strong, and useful effects.

Methodological Lessons Learned

We began self-experimentation with the belief that it was done the same way as conventional experimentation. Our experiences mostly supported that belief. However, we also learned some lessons, not necessarily specific to self-experimentation:

1. <u>Measure something you care about</u>. Self-experimentation is often exploratory, and exploratory research is often difficult because of its uncertainty and unfamiliarity. The more you care about a topic, the more likely you will persist in spite of difficulties. S. R.'s extensive self-

experiments began with bothersome personal problems--first, acne, later, early awakening. The acne research made progress within months, but the sleep research lasted eight years before substantial progress was made. A. N. was intrigued by the many implications of the questions "can behavioral variability be trained?" and "if so, what are the limits of such training?" He has done self-experiments on these questions for years.

2. Make data collection and analysis as easy as possible. Progress on serious problems will probably be slow. If you want to do something every day for many days, ease and convenience are important (Skinner, 1956). S. R.'s collection and analysis of sleep data improved considerably when he obtained a custom-made recording device and a home computer, making data analysis much easier.

3. Taking more than one measure is usually worth the trouble (in slight contradiction to Lesson 2). A few measures of behavior are almost always better than one, if adding the extra measures is not hard. This obviously makes sense if the additional measures reflect different dimensions. A. N. evaluated a number of different effects of activity and used multiple measures of response variability. What may be less obvious is that use of multiple measures also makes sense if the additional measures reflect the same dimension as the first. S. R. began measuring his weight using only one scale. When that scale was damaged, it became hard to compare new weights with old weights. When he resumed, he weighed himself each day on three different scales. Use of three scales made it possible to be much surer that any weight change was not due to a scale change, and made it possible to measure weight more precisely. However, some measurement procedures are too time-consuming. Early in his TV research, S. R. measured mood with a well-known questionnaire, the Profile of Mood States (the short form, Curran, Andrykowski, & Studts, 1995), in which the subject rates how much he feels each of 30 emotions (tense, angry, lively, etc.) on a 5-point scale (from 0 = not at all to 5 = extremely). This took several minutes–too hard to do each day for many days.

4. Make graphs. Graphs help find surprises (Tukey, 1977). S. R. did not notice his decrease in sleep duration coincident with weight loss (Figure 7) until he plotted sleep duration over time. A. N. and his students have found that graph-keeping helps maintain research activities, what the self-control literature calls a "reactive"effect, similar to the usefulness of monitoring in self-control projects (e.g., Taylor, 1985; Nelson, Boykin, & Hayes, 1982).

5. Communicate. To have new ideas, it is said, tell others the ideas you have now. A newsletter, Self-Experimentation/Self-Control Communication, distributed by Irene Grote, of the Department of Human Development, University of Kansas, has facilitated communication among self-experimenters.

6. A flawed experiment is better than none. In our experience, obvious flaws and weaknesses are inevitable, but rarely fatal. S. R.'s acne research involved counting the number of new pimples each day, a measurement with so much room for error that S. R. wondered if it could be worthwhile. It was. Measurements of mood (Table 3) are inevitably subjective and vague yet produced useful results. Moreover, doing research on a question today may increase the likelihood that you will do research on that question in the future (Neuringer, 1988). Paul Halmos, a renowned math teacher, taught his students that "the best way to learn is to do" (Halmos, 1975, p. 466). The best way to learn how to do a better experiment is often to do an

imperfect one.

        7. Value simplicity. The smallest step forward–the simplest, easiest way to increase what you know–is often the best. Complex or difficult experiments usually make more assumptions than simpler or easier ones–for example, assumptions about how quickly a treatment will act if it has any effect. Unless you know that those assumptions are true–because you have done simpler experiments that test the assumption rather than require it–some of them are likely to be wrong, making the results of complex experiments hard to interpret. The overconfidence that causes us to overvalue complex experiments also causes us to undervalue simple ones. As several of our examples show (e.g., S. R.'s discovery of the mood-altering effect of morning TV), simple manipulations often turn out to be more revealing or helpful than expected.

Statistical Issues

        Some writers have claimed that experiments with a single subject require different statistics than other experiments (e.g., Edgington, 1987, Chapter 10). This is wrong. Every experiment is n = 1 in many ways--one school, one stimulus set, one place, one time of day, etc. (Tukey, 1969). If an experiment uses only one school, it cannot generalize across schools; if it uses only one subject, it cannot generalize across subjects. If an experiment has 10 subjects, a t-test with $\underline{n} = 10$ will indicate what to expect if the experiment is done again with a similar group of subjects (different subjects chosen in a similar way). If an experiment has one subject, and data are collected from 10 days, a t-test with $\underline{n} = 10$ will indicate what to expect if the experiment is done again with the same subject and a similar group of days (different days chosen in a similar way). R. A. Fisher, who originated the statistical concept of significant difference, used a single-subject experiment (a woman tasting tea) to explain the concept (Fisher, 1951). The number of subjects in an experiment affects what can be concluded but not how those conclusions are reached.

        Repetition of self-experiments is often relatively easy. The purpose of inferential statistics (t tests, etc.) is to predict what would happen if the experiment were done again. The more you use actual repetition to answer this question, the less you need to rely on statistics (e.g., Tukey, 1969; Comstock, Bush, & Helzlsouer, 1992). That is, the less you need to assume that the requirements of your statistical test (independent samples, etc.) are close enough to the truth. If you are worried about the validity of a statistical inference (a test says that A and B are different-- is this correct?), the best check is to repeat the experiment.

Strengths of Self-Experimentation

        In a situation where self-experimentation and conventional research are both possible, self-experiments have several attractive features:

        1. Easy to measure treatment effects. "Does X influence Y?" is the most basic question of experimental science. All of our self-experiments answered this question more easily than a conventional experiment would have. The difference in required effort can be large. S. R. has done similar sleep experiments on himself and others. It required about 50-100 times more of the experimenter's time to do similar experiments with others than with himself. The conventional experiments required more time for recruitment, instructions, travel, and data collection. The self-experiments, unlike the conventional experiments, could be done without help, did not have subjects stop in the middle, did not require subject payments, and had fewer concerns about

treatment and measurement accuracy. Likewise, A. N. found it easier to develop methods to train behavioral variability when he used himself as subject than when he studied others.

Not only are self-experiments usually less time-consuming and expensive than conventional research, they often have more resolving power, i.e., better distinguish signal from noise. They can use more potent treatments, increasing signal. S.R. would not ask anyone else to drink 5 liters of water/day for many days but did not mind doing it himself. Other features decrease noise. The subject is usually experienced and always well-motivated, and the experiment can last a long time. A.N. could not have easily convinced others to spend hundreds of hours entering "random digits" (as he did) to develop methods to study operant variability.

The ease and power of self-experiments allow a researcher to answer many questions using self-experiments with the same effort it would take to answer one question using conventional methods. Self-experiments are well-suited for exploring a parameter space and finding optimal values. A. N.'s randomness work (Figure 2) and S. R.'s mood experiments (Figures 11 and 12) illustrate this point. When it comes time to do a conventional experiment to assess generality, these parametric results will help the researcher choose procedural details. The high power/effort ratio of self-experimentation also allows a researcher to do studies that seem to have a low or uncertain probability of success, as shown by many of our examples.

2. Can detect many changes and correlations without formal measurement. A self-experimenter's deep involvement helps him or her to notice when uncontrolled factors make a difference. Early in his work on variability, A. N. noticed that his behavior was less variable in the evening (when he was tired) than early in the day (when he was rested). After that, he did the research before noon. A later self-experiment showed that, indeed, time of day affected the randomness of his sequences. Likewise, a self-experimenter can easily notice when the experimental treatment changes something that is not being measured. After serving as his own subject in an experiment on thermoregulation, which involved prolonged immersion in hot water, Cabanac accidentally noticed that very cold water felt pleasant. This observation led to a great deal of research (Cabanac, 1995). S. R.'s mood and sleep research also illustrate the value of incidental observations. A self-experimenter knows in detail what a subject thinks and feels.

3. Encourage action. Because self-experiments are often easy, they encourage the researcher to do something. Our examples show how actions based on wrong ideas can be helpful.

4. Personally helpful. S. R. became enthusiastic about self-experimentation after it reduced his acne by 80% in one year. To S. R., the mood boost produced by morning TV is well worth 1-2 hours/day, quite apart from its research value. A. N.'s exercise research caused him to swim each day at about 4 o'clock, thereby improving his ability to work on academic and scientific tasks during the late afternoon and early evening.

5. Long-term records allow instructive correlations to be noticed. Self-experimentation often leads to months or years of measurements. Such records allow uncontrolled environmental changes ("experiments of nature") to yield useful information. When the data base is long, changes that take months can be noticed and interpreted (e.g., Figure 7). Day-to-day variation can also help show the "dosage" of a treatment needed to get the full effect (Table 2).

These strengths can be summed up by saying that self-experimentation often facilitates

three basic activities of science:

First, <u>hypothesis elimination</u>. Self-experiments make it relatively easy to test an idea, and hard to dismiss the results when they contradict a favored theory. Because you know so clearly what happened (Strength 2), there may be little room for post-hoc explanations ("well, that was because . . . ").

Second, <u>hypothesis generation</u>. Because self-experiments produce many results (Strength 1), they tend to produce many surprises–and surprises often lead to new ideas. The surprise may be an unexpected change (Strength 2) or correlation (Strengths 2 and 4). Or it may be the convincing elimination of a favorite hypothesis. Because the surprises of psychological self-experiments involve the experimenter's own behavior, they are especially thought-provoking.

Finally, <u>trial and error</u>. Edison's description of genius ("99% perspiration") could have been "99% making mistakes." Because mistakes in a self-experiment have a relatively low cost (Strength 1), more can be made.

<div align="center">Weaknesses of Self-Experimentation</div>

Self-experiments have important limitations, of course. In rough order of importance (most important first):

1. <u>Expectations may influence results</u>. In most cases, self-experiments cannot be run "blind," which allows expectations to vary with treatment and cause differences between treatments (Rosenthal, 1966). However, the results of our and our students' self-experiments have often been surprising, implying that expectations often have small or short-lived effects. Examples come from studies of random sequence generation, acne, study efficiency, sleep, mood, and weight, among others. A simple test of the power of expectations is to ask if the results have always turned out as expected; if so, expectations may be powerful in that situation.

A researcher may notice that two rare events happened at about the same time and do an experiment to ask if one caused the other. For instance, S. R.'s TV/mood research began when he noticed that he woke up feeling refreshed and cheerful (rare) the day after he watched TV early in the morning (rare). These correlations are interesting <u>because</u> they are unexpected, and thus could not be due to expectations. If an experiment confirms the correlation--if intended variation of X produces the effect that the correlation suggested--this implies that the experimental effect was not due to expectations.

A few self-experiments <u>can</u> be done completely blind. One of our students studied the effect of caffeine by having a friend place caffeinated instant coffee in a jar marked "A" and identical-looking decaffeinated instant coffee in a jar marked "B." Following an ABA design, she alternated between jars. Contrary to what she expected, she found her behavior and mood to be strongly affected by caffeine (Neuringer, 1981).

2. <u>Generality across subjects unclear</u>. As discussed earlier, every experiment is n = 1 in many ways. Yet because people plainly differ in important ways the fact that a self-experiment involves only one person is an obvious concern. The best way to assess generality across subjects is of course to do the same experiment with other subjects. However, other sources of data also help answer questions about generality.

One is history. In dozens of cases, the conclusions from self-experiments have turned out to be widely true (e.g., Altman, 1987); we know of no case where the result of a self-experiment

was misleading because the subject was unusual. Psychologists care about individual differences partly because any sensitive experiment will reveal them--different subjects are affected differently by the same treatment. However, experimental reports with data for individual subjects show that between-subject differences are generally differences in the size of the effect, not its direction.

And few self-experiments stand alone. In our experience, other evidence usually sheds light on the generality of the results. S. R.'s breakfast results, for instance, closely resembled food-anticipatory activity in animals. S. R.'s discovery that morning TV improved his mood suggested a theory about depression supported by data from other subjects (e. g., data linking depression and circadian rhythms).

3. Limited subject matter. Self-experiments can study only a few of the topics that experimental psychologists investigate. Many experiments compare different groups of people, e.g., studies of sex differences. The experimenter may not belong to the group of interest, e.g., persons with an illness or disability.

4. Easy to lie. Self-interest should minimize this problem. It is hardly in the experimenter's interest to do a long series of self-experiments based on false or misleading results, to do conventional experiments based on false or misleading results (based on earlier self-experiments), or to publish false or misleading results. Moreover, attempts by other experimenters to replicate important findings provide protection, just as with conventional research.

5. Interferes with daily life. Self-experimentation is sometimes a burden, but not always. Self-experimentation led S. R. to stop eating breakfast, maybe saving a half-hour each day. Weight loss due to self-experimentation caused him to sleep about a half-hour less each night, more time saved. Standing rather than sitting took no additional time, nor did exposure to morning light. S. R.'s TV/mood research was time-consuming (1-2 hours/day) and intrusive, but, as mentioned earlier, the benefits made the cost seem a bargain.

6. Real life noisier than lab. A common objection to self-experimentation is "my life varies too much." In some cases, this is probably true. A student who works the graveyard shift (11 p.m. to 7 a.m.) two nights/week should probably not try to do an experiment on circadian rhythms. On the other hand, our examples show that a lot can be learned from real-life experiments and that real-life noise is not always a problem. Some self-experiments, like Ebbinghaus's memory studies and A. N.'s randomness work, can be done in labs or lab-like isolation. In some cases, treatment effects are much larger than real-life noise, e.g., S. R.'s body-weight results (Figures 5 and 6). Often enough data can be collected to greatly reduce noise by averaging (e.g., Figure 7).

<div align="center">Conclusions</div>

Biological diversity (in terms of number of species) is especially high along boundaries between two different habitats (e.g., forest and meadow), a phenomenon that ecologists call the edge effect (Harris, 1988, p. 330). Much the same principle holds for economies: Cities prosper and diversify their economic activity where a number of economic conditions co-exist (Jacobs, 1984). Something similar should be true for psychology: The combination of self-experimentation and conventional research will be more fruitful than either alone.

Self-experimentation and conventional research have complementary strengths. The essential strength of self-experiments is how easy they are (compared to conventional research on the same topic). They can try many treatments, measure many things at once, generate and test many ideas, allow considerable trial and error. However, some topics cannot be studied, the subject pool is limited, and generality across subjects is unclear. Conventional experiments are usually more difficult but are also more versatile–they can study a wider range of topics and subjects, including animals–and more convincing. The use of complementary methods is central to public health research, where epidemiology, laboratory research, and clinical trials work well together. Epidemiology (i.e., survey research) is better than laboratory work and clinical trials for generating ideas about cause and effect, but worse for testing them.

Self-experiments and conventional research can help each other in several ways. Self-experiments can suggest conventional ones. For instance, self-experiments can filter anecdotal evidence. In spite of plenty of anecdotes about water and weight loss, no conventional experiments have measured the effect of water intake on weight. S. R.'s self-experiment that showed a clear effect of water on weight provides a better basis for conventional experimentation than an anecdote would. In addition, self-experiments can help decide the details of conventional experiments–e. g., how long treatments should last. Help can flow the other way, too. Self-experiments can take a scientific question raised by conventional research and try to answer it, as in A. N.'s randomness work. Conventional research may suggest solutions to practical problems; self-experiments can ask if the solutions work. S. R.'s weight experiments, for instance, were partly inspired by animal studies. Conventional research may also help show that a surprising observation in a self-experiment is correct, e.g., S. R.'s sleep and mood research.

When self-experimentation and conventional experiments are both possible–when a researcher has a choice--the difference between them often resembles the difference between learning and showing: <u>Self-experiments are better for discovery</u> (solving an everyday problem, answering a scientific question) <u>but worse for convincing others</u> that the solution is helpful or the answer is correct. Of course, most scientists want to do both--discover something <u>and</u> convince others of their discovery. Thus psychologists should consider doing both self-experiments and conventional ones, if self-experiments would help answer the question they are asking. The best use of resources may often be self-experiments followed by conventional ones. The researcher begins with self-experiments that, if all goes well, find large effects and/or generate and eliminate many hypotheses. This exploratory and theory-building phase lasts until a convenient solution or large effect is found. Then the researcher uses self-experiments to find the procedural parameters (duration, time of day, intensity, etc.) that optimize the solution or maximize the effect. Only then would the researcher begin conventional experiments, using the optimized parameters.

Science involves both hypothesis creation and hypothesis testing; the tools that are best for one are unlikely to be best for the other. Unfortunately, education in scientific methods emphasizes testing far more than creation. Statistics textbooks, for example, are full of ideas about how to test hypotheses but often ignore methods of hypothesis creation. Our examples show that self-experimentation is especially good for hypothesis creation.

New techniques and equipment often lead to bursts of progress shortly after they become available. Self-experimentation is an old technique but also, our examples suggest, an unwisely

neglected one. Unlike most new tools--unlike, say, a magnetic resonance imaging machine–self-experimentation is available to everyone at a price everyone can afford. It will never be the last word, but it may often be a good place to start.

References

Altman, L. K. (1972). Auto-experimentation: An unappreciated tradition in medical science. New England Journal of Medicine, 286, 346-352.

Altman, L. K. (1987). Who goes first? The story of self-experimentation in medicine. New York: Random House.

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.

Baddeley, A. D. (1966) The capacity for generating information by randomization. Quarterly Journal of Experimental Psychology, 18, 119-129.

Barinaga, M. (1994). To sleep, perchance to... learn? New studies say yes. Science, 265, 603-604.

Bender, A. E. (1993). Breakfast–Role in the diet. In R. Macrae, R. K. Robinson, & M. J. Sadler (Eds.), Encyclopaedia of food science, food technology, and nutrition (Vol. 1, pp. 488-490). London: Academic Press.

Bernstein, R. K. (1984). Diabetes: The glucograF method for normalizing blood sugar. Los Angeles: J. P. Tarcher.

Bernstein, R. K. (1990). Diabetes type II: Living a long, healthy life through blood sugar normalization. New York : Prentice Hall Press.

Bernstein, R. K. (1994, March 12). To control diabetes, cut down carbohydrates. New York Times, 143, p. 14 (National Edition).

Blood-glucose meters: They're small, fast, and reliable. (1996, October). Consumer Reports, 61, 53-55.

Boivin, D. B., Czeisler, C. A., Dijk, D.-J., Duffy, J. F., Folkard, S., Minors, D. S., Totterdell, P., & Waterhouse, J. M. (1997). Complex interaction of the sleep-wake cycle and circadian phase modulates mood in healthy subjects. Archives of General Psychiatry, 54, 145-152.

Bolles, R. C. (1993). The story of psychology: A thematic history. Pacific Grove, CA: Brooks/Cole.

Bolles, R. C., Hayward, L., Crandall, C. (1981). Conditioned taste preferences based on caloric density. Journal of Experimental Psychology: Animal Behavior Processes, 7, 59-69.

Bolles, R. C., & Stokes, L. W. (1965). Rat's anticipation of diurnal and adiurnal feeding. Journal of Comparative and Physiological Psychology, 60, 290-294.

Borbély, A. A. (1982). A two process model of sleep regulation. Human Neurobiology, 1, 195-204.

Boring, E. G. (1942). Sensation and perception in the history of experimental psychology. New York Appleton-Century-Crofts.

Boulos, Z., & Terman, M. (1980). Food availability and daily biological rhythms. Neuroscience and Biobehavioral Reviews, 4, 119-131.

Brown, G. W., & Harris, T. (1978). Social origins of depression. New York: Free Press.

Brown, K. S. (1995, December 11). Testing the most curious subject--oneself. The Scientist, 9 (no. 24), pp. 1, 10.

Boulos, Z., Rosenwasser, A. M., & Terman, M. (1980). Feeding schedules and the

circadian organization of behavior in the rat. Behavioural Brain Research, 1, 39-65.

Bugged by disease. (1998, March 21). The Economist, 346, 93-94.

Cabanac, M. (1995). La quête du plaisir. Montreal: Liber.

Cabanac, M., & Rabe, E. F. (1976). Influence of a monotonous food on body weight regulation in humans. Physiology & Behavior, 17, 675-678.

Castiglioni, A. (1931). Life and work of Sanctorius. Medical Life, 38, 729-785.

Chagnon, N. A. (1983). Yanomamo: The fierce people. New York: Holt, Rinehart and Winston.

Cofer, C. N. (1979). Human learning and memory. In E. Hearst (Ed.), The first century of experimental psychology (pp. 323-370). Hillsdale, NJ: Erlbaum.

Cohen, L., Neuringer, A., & Rhodes, D. (1990). Effects of ethanol on reinforced variations and repetitions by rats under a multiple schedule. Journal of the Experimental Analysis of Behavior, 54, 1-12.

Comstock, G. W., Bush, T. L., & Helzlsouer, K. (1992). Serum retinol, beta-carotene, vitamin E, and selenium as related to subsequent cancer of specific sites. American Journal of Epidemiology, 135, 115-121.

Crisp, A. H., Stonehill, E., & Fenton, G. W. (1971). The relationship between sleep, nutrition, and mood: A study of patients with anorexia nervosa. Postgraduate Medical Journal, 47, 207-213.

Crisp, A. H., Stonehill, E., Fenton, G. W., & Fenwick, P. B. C. (1973). Sleep patterns in obese patients during weight reduction. Psychotherapy and Psychosomatics, 22, 159-165.

Curran, S. L., Andrykowski, M. A., & Studts, J. L. (1995). Short Form of the Profile of Mood States (POMS-SF): Psychometric information. Psychological Assessment, 7, 80-83.

Ebbinghaus, H. (1913) Memory: A contribution to experimental psychology. New York: Columbia University. (Original work published 1885)

Edgington, E. S. (1987). Randomization tests (2nd ed.). New York: Marcel Dekker.

Ehlers, C. L., Frank, E., Kupfer, D. J. (1988). Social zeitgebers and biological rhythms. A unified approach to understanding the etiology of depression. Archives of General Psychiatry, 45, 948-52.

Evans, F. J. (1978). Monitoring attention deployment by random number generation: An index to measure subjective randomness. Bulletin of the Psychonomic Society, 12, 35-38.

Evans, F. J. & Graham, C. (1980). Subjective random number generation and attention deployment during acquisition and overlearning of a motor skill. Bulletin of the Psychonomic Society, 15, 391-394.

Fisher, R. A. (1951). The design of experiments (6th ed.). London: Hafner.

Foster-Powell, K., & Brand-Miller, J. (1995). International tables of glycemic index. American Journal of Clinical Nutrition, 62, 871S-893S.

Franklin, J., & Sutherland, J. (1984). Guinea pig doctors: The drama of medical research through self-experimentation. New York: Morrow.

Gallup Organization (1991). Sleep in America. Princeton, NJ: The Gallup Organization.

Gibbs, W. W. (1996, August). Gaining on fat. Scientific American, 275, 88-94.

Gurley, V., Neuringer, A., & Massee, J. (1984). Dance and sports compared: Effects on psychological well-being. The Journal of Sports Medicine and Physical Fitness, 24, 58-68.

Halmos, P. R. (1975). The problem of learning to teach: I. The teaching of problem solving. American Mathematical Monthly, 82, 466-470.

Harris, L. D. (1988). Edge effects and conservation of biotic diversity. Conservation Biology, 2, 330-332.

Hecht, S., Schlaer, S., & Pirenne, M. H. (1942). Energy, quanta, and vision. Journal of General Physiology, 25, 819-840.

Hervey, G. R. (1969). Regulation of energy balance. Nature, 222, 629-631.

Holman, J., Goetz, E. M., & Baer, D. M. (1977). The training of creativity as an operant and an examination of its generalization characteristics. In B. Etzel, J. Le Bland, & D. Baer (Eds.), New developments in behavioral research: Theory, method and application (pp. 441-471). Hillsdale, NJ: Erlbaum.

Jacobs, J. (1984). Cities and the wealth of nations. New York: Random House.

Jenkins, J. G., & Dallenbach, K. M. (1924). Oblivescence during sleep and waking. American Journal of Psychology, 35, 605−612.

Kendall, A., Levitsky, D. A., Strupp, B. J., & Lissner, L. (1991). Weight loss on a low-fat diet: consequence of the imprecision of the control of food intake in humans. American Journal of Clinical Nutrition, 53, 1124-1129. Kermode, F. (1995). Not entitled. New York: Farrar, Straus and Giroux.

Knuth, D. E. (1969). The art of computer programming. Vol. 2. Semi-numerical algorithms. Reading, MA: Addison-Wesley.

Kripke, D. F. (1984). Critical interval hypotheses for depression. Chronobiology International, 1, 73-80.

Kristofferson, A. B. (1976). Low-variance stimulus-response latencies: Deterministic internal delays? Perception & Psychophysics, 20, 89-100.

Kristofferson, A. B. (1977). A real-time criterion theory of duration discrimination. Perception & Psychophysics, 21, 105-117.

Kristofferson, A. B. (1980). A quantal step function in duration discrimination. Perception & Psychophysics, 27, 300-306.

Lacey, J. H., Crisp, A. H., Kalucy, K. S., Hartmann, M. K., & Chen, C. N. (1975). Weight gain and the sleeping electroencephalogram: Study of 10 patients with anorexia nervosa. British Medical Journal, 4, 556-558.

Machado, A. (1989). Operant conditioning of behavioral variability using a percentile reinforcement schedule. Journal of the Experimental Analysis of Behavior, 52, 155-166.

Mahoney, M. J. (1974). Cognition and behavior modification. Cambridge, MA: Ballinger.

Mahoney, M. J. (1979). Self-change: Strategies for solving personal problems. New York: Norton.

McElroy, E., & Neuringer, A. (1990). Effects of alcohol on reinforced repetitions and reinforced variations in rats. Psychopharmacology, 102, 49-55.

Mistlberger, R. E., Houpt, T. A., & Moore-Ede, M. C. (1990). Food-anticipatory rhythms

under 24-hour schedules of limited access to single macronutrients. Journal of Biological Rhythms, 5, 35-46.

Mook, D. M., & Neuringer, A. (1994). Different effects of amphetamine on reinforced variations versus repetitions in Spontaneously Hypertensive Rats (SHR). Physiology & Behavior, 56, 939-944.

Moore-Ede, M. C., Sulzman, F. M., & Fuller, C. A. (1982). The clocks that time us: Physiology of the circadian timing system. Cambridge, MA: Harvard University Press.

Mosteller, F., & Tukey, J. W. (1977). Data analysis and regression: A second course in statistics. Reading, MA: Addison-Wesley.

Nelson, R. O., Boykin, R. A., & Hayes, S. C. (1982). Long-term effects of self-monitoring on reactivity and on accuracy. Behaviour Research and Therapy, 20, 357-363.

Neuringer, A. (1981). Self-experimentation: A call for change. Behaviorism, 9, 79-94.

Neuringer, A. (1986). Can people behave "randomly"? The role of feedback. Journal of Experimental Psychology: General, 115, 62-75.

Neuringer, A. (1988). Personal paths to peace. Behavior Analysis and Social Action, 6, 51-56.

Neuringer, A. (1991). Operant variability and repetition as functions of interresponse time. Journal of Experimental Psychology: Animal Behavior Processes, 17, 3-12.

Neuringer, A., & Huntley, R. W. (1991). Reinforced variability in rats: Effects of gender, age and contingency. Physiology & Behavior, 51, 145-149.

Neuringer, A., & Voss, C. (1993). Approximating chaotic behavior. Psychological Science, 4, 113-119.

O'Dea, K. (1984). Marked improvement in carbohydrate and lipid metabolism in diabetic Australian aborigines after temporary reversion to traditional lifestyle. Diabetes, 33, 596-603.

Page, S., & Neuringer, A. (1985). Variability is an operant. Journal of Experimental Psychology: Animal Behavior Processes, 11, 429-452.

Paxton, S. J., Trinder, J., Montgomery, I., Oswald, I., Adam, K., & Shapiro, C. (1984). Body composition and human sleep. Australian Journal of Psychology, 36, 181-189.

Raben, A., Due Jensen, N., Marckmann, P., Sandström, B., & Astrup, A. (1995). Spontaneous weight loss during 11 weeks' ad libitum intake of a low fat/high fiber diet in young normal subjects. International Journal of Obesity, 19, 916-923.

Ramirez, I. (1990). Stimulation of energy intake and growth by saccharin in rats. Journal of Nutrition, 120, 123-133.

Reichenbach, H. (1957). The rise of scientific philosophy. Berkeley: University of California Press.

Rosenthal, R. (1966). Experimenter effects in behavioral research. New York: Appleton-Century-Crofts.

Sclafani, A., & Springer, D. (1976). Dietary obesity in adult rats: Similarities to hypothalamic and human obesity syndromes. Physiology & Behavior, 17, 461-471.

Shephard, R. J., Jones, G., Ishii, K., Kaneko, M., & Olbrecht, A. J. (1969). Factors affecting body density and thickness of subcutaneous fat. Data on 518 Canadian city dwellers. American Journal of Clinical Nutrition, 22, 1175-1189.

Skinner, B. F. (1956). A case history in scientific method. American Psychologist, 11, 221-233.

Sheppard, L., Kristal, A. R., & Kushi, L. H. (1991). Weight loss in women participating in a randomized trial of low-fat diets. American Journal of Clinical Nutrition, 54, 821-828.

Shintani, T. T., Hughes, C. K., Beckham, S., & O'Connor, H. K. (1991). Obesity and cardiovascular risk intervention through the ab libitum feeding of traditional Hawaiian diet. American Journal of Clinical Nutrition, 53, 1647S-1651S.

Siegel, M. H., & Zeigler, H. P. (1976). Psychological research: The inside story. New York: Harper & Row.

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. Current Direction in Psychological Science, 3, 1-5.

Soldatos, C. R. (1994). Insomnia in relation to depression and anxiety: Epidemiological considerations. Journal of Psychosomatic Research, 38, Suppl. 1, 3-8.

Steinhauer, G. D. & Bol, L. (1988) Computer assisted self observation. Berkeley, CA: Artificial Behavior, Inc.

Stokes, P. (1995) Learned variability. Animal Learning and Behavior, 23, 164-176.

Stratton, G. M. (1966). Vision without inversion of the retinal image. In Herrnstein, R. J., & Boring, E. G. (Eds.), A source book in the history of psychology (pp. 103-112). Cambridge, MA: Harvard University Press. (Original work published 1897)

Szalai, A. (1973). The use of time. The Hague: Mouton.

Taylor, I. L. (1985). The reactive effect of self-monitoring of target activities in agoraphobics: A pilot study. Scandinavian Journal of Behaviour Therapy, 14, 17-22.

Thorndike, E. (1900). Mental fatigue. Psychological Review, 7, 466-482.

Titchener, E. B. (1896). An outline of psychology. New York: Macmillan.

 Tokura, H., Shimomoto, M., Tsurutani, T., & Ohta, T. (1978). Circadian variation of insensible perspiration in man. International Journal of Biometeorology, 22, 271-278.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work. American Psychologist, 24, 83-91.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Tune, G. S. (1964). A brief survey of variables that influence random generation. Perceptual & Motor Skills, 18, 705-710.

Van Cauter, E., & Turek, F. W. (1986). Depression: a disorder of timekeeping? Perspectives in Biology and Medicine, 29, 510-519.

Van den Hoofdakker, R. H. (1994). Chronobiological theories of nonseasonal affective disorders and their implications for treatment. Journal of Biological Rhythms, 9, 157-183.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of literature. Psychological Bulletin, 77, 65-72.

Walsh, B. T., Goetz, R., Roose, S. P., Fingeroth, S., & Glassman, A. H. (1985). EEG-monitored sleep in anorexia nervosa and bulimia. Biological Psychiatry, 20, 947-956.

Watson, D. L., & Tharp, R. G. (1993). Self-directed behavior: Self-modification for personal adjustment. 6th ed. Pacific Grove, CA: Brooks Cole.

Wever, R. A. (1979). The circadian system of man: Results of experiments under

temporal isolation. New York: Springer-Verlag.

Zuriff, G. E. (1985). Behaviorism: A conceptual reconstruction. New York: Columbia
University Press.

Author Note

Table 1

Comparison of 100 Trials by A. N. With 100 Trials by Berkeley-Pascal Random Number
Generator

Lag	Mean		Standard Deviation		t	p
	A. N.	R. N. G.	A. N.	R. N. G		
1	.245	.242	.028	.024	0.85	>.4
2	.239	.243	.037	.023	1.01	>.2
3	.243	.245	.027	.025	0.60	>.5
4	.240	.246	.025	.023	1.71	>.05
5	.243	.243	.028	.026	0.01	>.8
6	.241	.244	.027	.026	0.56	>.5

Note. R. N. G. = Berkeley-Pascal random number generator.

Table 2
Correlation Between Standing and Early Awakening

When	Standing (hr)	Days	Median standing (hr)	Days with early awakening the next morning	Proportion (percent)
May 18, 1996- August 26, 1996	not measured	100	not measured	57	57
August 27, 1996- October 24, 1996	5.0-8.0	20	7.0	12	60
	8.0-8.8	34	8.5	5	15
	8.8-11.0	5	9.3	0	0
October 25, 1996- February 28, 1997	5.0-8.0	10	6.8	6	60
	8.0-8.8	8	8.5	2	25
	8.8-11.0	90	9.3	1	1

Note. Early awakening = Fell back asleep between 10 minutes and 6 hours after getting up.
Because of travel and illness, some days were not included. Median standing gives the median
duration of standing for the days in that category; e.g., 7.0 is the median of 20 days.

Table 3
Mood Scales

Rating	Dimension		
	unhappy --> happy	irritable --> serene	reluctant --> eager
9.5	extremely happy	extremely serene	extremely eager
9	very happy	very serene	very eager
8	quite happy	quite serene	quite eager
7.5	happy	serene	eager
7	somewhat happy	somewhat serene	somewhat eager
6	slightly happy	slightly serene	slightly eager
5	neither happy nor unhappy	neither serene nor irritable	neither eager nor reluctant
4	slightly unhappy	slightly irritable	slightly reluctant
3	somewhat unhappy	somewhat irritable	somewhat reluctant
2.5	unhappy	irritable	reluctant
2	quite unhappy	quite irritable	quite reluctant
1	very unhappy	very irritable	very reluctant
0.5	extremely unhappy	extremely irritable	extremely reluctant

Table 4
Face Density and Duration Effects on Mood

Phase	Days	Face duration (min)	Total duration (min)	Face density	Mood
before experiment	33	30	67	46%	7.4 ± 0.1
1	7	30	36	83%	7.9 ± 0.1
2	4	30	76	40%	6.7 ± 0.1
3	6	30	39	78%	7.8 ± 0.1
after experiment	10	50	64	79%	8.5 ± 0.1

Note. Values in the columns Face duration, Total duration, Face density, and Mood are 10% trimmed means of the values for individual days. Standard errors were computed with the jackknife (Mosteller & Tukey, 1977, Chapter 8).

Figure Captions

Figure 1. RNG, a measure of sequence uncertainty, as a function of training trials (averaged across blocks of 10) for the author (AN) and an experimentally naive subject (MS). (High RNG values indicate repetitious sequences, lower values indicate more variable sequences.)

Figure 2. RNG as a function of interresponse time. (Each point is an average of 10 trials; the error bars show standard errors.)

Figure 3. Some of the feedback given the author (A. N.) after each trial. (Each trial consisted of 100 responses. Shown are 8 of the 30 statistics provided at the end of a trial. The dashed lines show the means for the performance of a random number generator; the letters indicate performance by the subject relative to the random generator in standard deviation units, with A = +0.5, B = +1.0, a = -0.5, b =-1.0, etc.)

Figure 4. Learning curves for the Move and Sit conditions showing the average number of faces correctly identified (in a list of 20 faces) as a function of the number of times the list had been studied.

Figure 5. Body weight as a function of day. (Weights were normalized by subtracting the pre-diet weight for each subject. Pre-diet weights were, in pounds, 197 for SR, 167 for JH, and 146 for AL.)

Figure 6. S. R.'s body weight as a function of day. (Each weight is the average of three scales. Each scale could be read to the nearest pound. The wiggly line is a moving average [mean] of three days.)

Figure 7. Long-term record of S. R.'s sleep duration. (Each data point is the 10% trimmed mean of 21 days. The wiggly line gives the moving averages [means] of 93 days. The tick marks for each year indicate the first day of that year.)

Figure 8. Probability of having fallen back asleep as a function of breakfast and time since first getting up. (Probability of having fallen back asleep after 2 hours = number of days on which S. R. fell back asleep within 2 hours after getting up/number of days included in the condition. The functions are based on days both with and without early awakening, i.e., both the days on which S. R. fell back asleep after getting up and the days on which he did not fall back asleep after getting up. The results are from the mornings after eating the indicated breakfasts. For instance, if fruit was the breakfast Monday through Friday, the fruit results would be based on Tuesday through Saturday morning. The "oatmeal" function is based on the 300 days before the breakfast variations began, during which oatmeal was almost always the breakfast. The "none--1st" function is based on the first 102 days when no breakfast was eaten, omitting the first eight days. The "fruit" function is based on the following 47 days, when breakfast was one piece of fruit. The "none--2nd" function comes from the 62 days after that, omitting the first two, when again no breakfast was eaten.)

Figure 9. S. R.'s early awakening 1990-1997. (Each point is a different day that early awakening occurred. The height of the points indicates the time between getting up and falling back asleep; the density of the points indicates the probability of early awakening. In the upper panel, the wiggly line shows the moving average [mean] of 31 points. In the lower panel--based on exactly the same data as the upper panel--the wiggly line gives the probability of early

awakening based on the 31 days in the neighborhood of the day at which the point is plotted--the 15 days before, the day itself, and the 15 days after. The tick mark for each year indicates the first day of that year. The data are from April 10, 1990 through February 28, 1997.)

      Figure 10. Mood as a function of morning TV and day. (Each mood rating is a mean of the ratings on the three scales of Table 3. No ratings were made during a block of days during the first phase because of sickness.)

      Figure 11. Mood next day as a function of face time. (The number of days indicated by each data point is the number of days that contributed to that datum. The mood for each day was an average over the three scales of Table 3. The results for each duration were taken from the days after the treatment was used. For instance, if a treatment was used Monday through Thursday, the results for that treatment came from the mood ratings on Tuesday through Friday. Averages over days were 10% trimmed means, with standard errors computed using the jackknife [Mosteller & Tukey, 1977, Chapter 8].)

      Figure 12. Mood next day as a function of TV starting time and distance. (See the caption of Figure 11 for more information.)

      Figure 13. Mood next day as a function of face density. (Each point is a different day. The data come only from days where face time was 30 or 40 minutes, the viewing distance was 1 m, and the starting time was 7 a.m.)