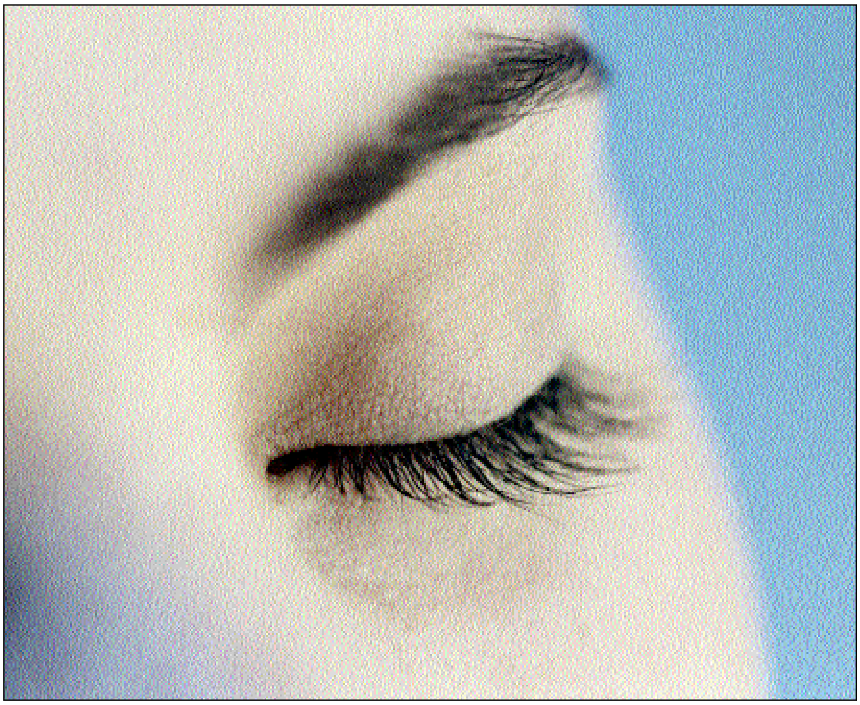
Many take it as a given that randomized experiments with many subjects are the only way to learn. Is it true?

Surprises From Self-Experimentation: Sleep, Mood, and Weight

Seth Roberts



WWW.PHOTODISC.COM

I read *Exploratory Data Analysis* by John Tukey while I was a graduate student in experimental psychology. I enjoyed reading it and absorbed at least its most basic lesson, the value of plotting data. At the time, I was doing experiments with rats to learn how they measure time. One of the book's main points is that "restricting one's self to the planned analysis — failing to accompany it with exploration — loses sight of the most interesting results too frequently to be comfortable" (p. 3). My data supported this view. Perhaps 1% of my exploratory graphs showed something surprising and interesting, and much of my research after graduate school, including some of the experiments described here, derived from ideas generated this way. It was also in graduate school that I began to do self-experiments. This article is about my slow realization that self-experimentation and exploratory data analysis have something in common. Both are ways of generating ideas. Self-experimentation seems to be better for generating ideas than more conventional ways of collecting data, just as exploratory data analysis seems to be better for generating ideas than more conventional ways of analyzing data.

My interest in self-experimentation began when I read an article about teaching mathematics by Paul Halmos, a professor at Indiana University. Halmos emphasized that "the best way to learn is to do." I was trying to learn how to do experiments; I took this advice to mean I should do as many as possible. I could do more experiments, I realized, if I not only did rat experiments but also did experiments with myself as the subject. So I started doing small self-experiments. Most of them were trivial and led nowhere (e.g., experiments about juggling).

At the time I had acne. My dermatologist had prescribed both pills (tetracycline, a wide-spectrum antibiotic) and

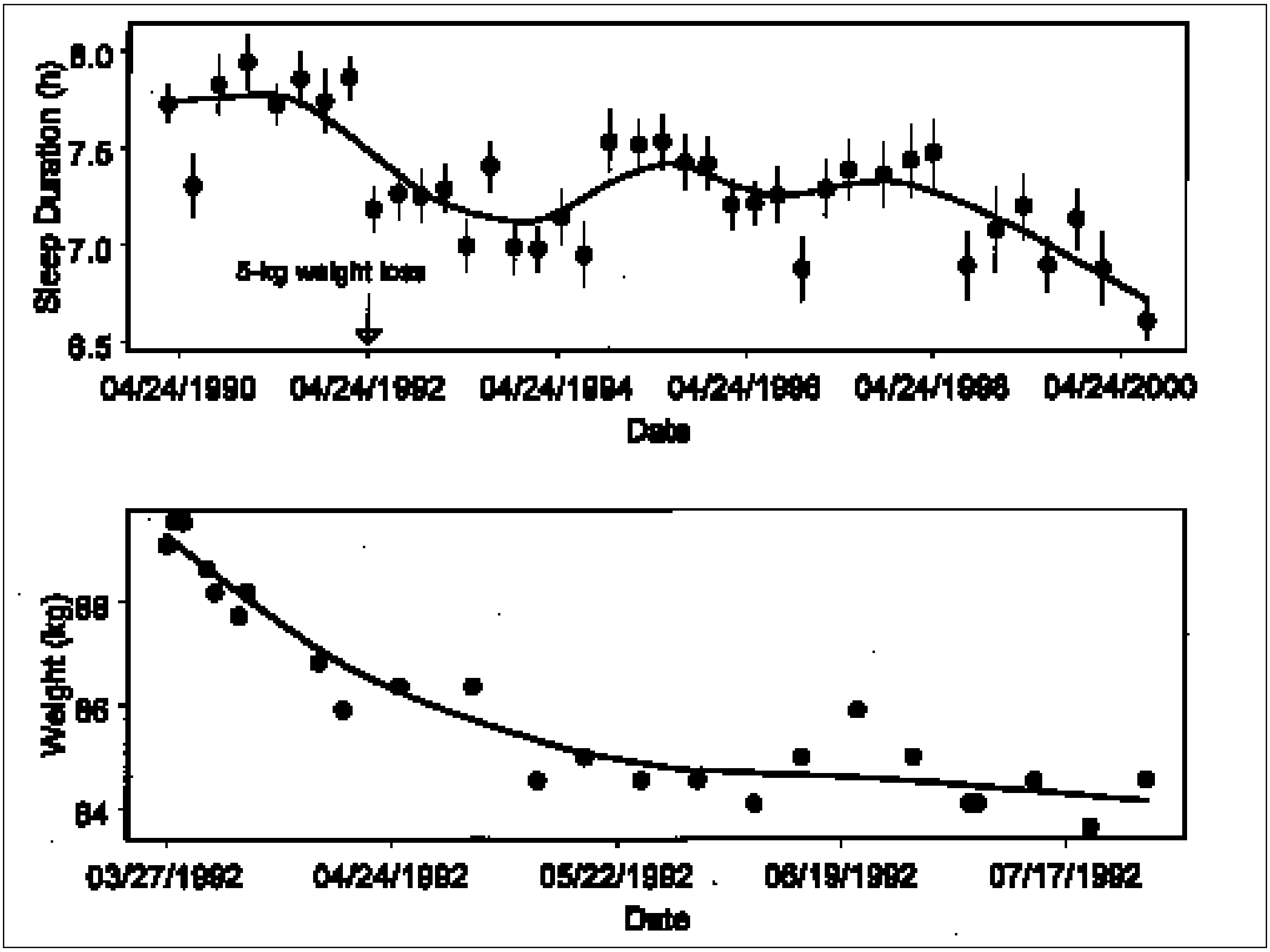


Figure 1. Upper panel: sleep duration (including naps) over time. Each point is a 10% trimmed mean. Error bars show standard errors determined by jackknifing. Lower panel: weight over time. The measurements started when the dietary change began; they stopped because the scale broke.

a cream (active ingredient benzoyl peroxide). Simply for the sake of doing experiments, any experiments, I did simple tests to measure the effectiveness of these treatments. I believed the pills were powerful and the cream had little effect. To my great surprise, the tests showed the opposite: The cream was powerful and the pills had little effect. It was very useful information. Many years later, an article in the *British Journal of Dermatology* reported that antibiotic-resistant acne is common.

Sleep and Breakfast

A few years after graduate school I began to have trouble sleeping. I would wake up early in the morning, tired but unable to fall back asleep for several hours — a type of insomnia called *early awakening*. There was no good treatment for it, and it did not go away. My experience with acne made me think self-experimentation might help.

Sleep was harder than acne. My acne studies had uncovered useful facts within weeks; my first 10 years of sleep research, however, merely showed that all my ideas about the cause of early awakening were wrong, or at least not right enough to make much difference. Among the failed treatments were dietary changes, exercise, and changes in the timing of bedroom lights that went on in the morning. What should you do when all your theories are wrong? I had no idea.

In 1990, I got a personal computer at home, making analysis of my sleep data much easier. In early 1993, while exploring the data, I looked at a graph similar to the upper panel of Fig. 1, which shows sleep duration over time. The 1993 graph had less data and was noisier than the upper panel of Fig. 1 but nevertheless revealed the same thing—that my sleep duration had decreased by about 40 minutes/day in the middle of 1992. I had not noticed the change. I never used an alarm clock, so the change implied that my need for sleep had decreased.

The change in sleep duration had happened at the same time I had lost 5 kg (lower panel of Fig. 1) by changing my diet. Before the change, I had been eating a conventional low-fat healthy diet. The dietary change was a reduction in processing (e.g., cooking, blending, adding spices). For instance, I ate raw fruit instead of fruit juice, brown rice instead of bread, and stopped eating almost all prepared foods, including delicatessen food, baked goods, and frozen entrees. I did not try to change (or keep constant) how much I ate; I always ate as much as I wanted to. The weight loss was not a surprise; based on rat experiments, I had been telling students for years that processing food usually makes it more fattening.

The next time I lectured on weight control to my introductory psychology class I showed a graph similar to Fig. 1, with its suggestion that if you lose weight you may need less sleep. A few weeks later, a student named Michael Lee came to my office to tell me that he knew another way to lose weight and sleep less: Eat a diet high in water content. In practice, this meant eating lots of fruit and salad. It had worked for him, he said. So I tried it. After a few weeks, however, it was clear the new diet had little effect. I told Michael the results. He asked how much fruit I was eating each day. Four pieces, I said. "I eat *six* pieces," he said.

So I started eating six pieces of fruit each day. This required changing my breakfast — instead of oatmeal, I had two pieces of fruit, such as a banana and an apple. After about 10 days of the new breakfast, I noticed that I was waking up too early much more often. While eating oatmeal, I had been waking up too early about a third of the time. (I defined an instance of waking up too early as a morning when I fell back asleep within six hours after getting up.) Now I was waking up too early every morning. I switched back to oatmeal and early awakening returned to its earlier level. I started eating fruit breakfasts again and early awakening again became much more common — leaving no doubt it was cause and effect. This was exciting; after 10 years of failure, I had finally found something that made a difference, albeit in the wrong direction. Tests of other breakfasts suggested that any breakfast with a substantial number of calories caused early awakening. A seven-month experiment with an ABA design (weeks of no breakfast, followed by weeks of one piece of fruit for breakfast, followed by weeks of no breakfast) showed clearly that a breakfast of one piece of fruit produced much more early awakening than no breakfast at all.

That breakfast can interfere with sleep was surprising, of course, but entirely consistent with animal research. It is well known that animals become more active a few hours before meal times. For example, if you feed a rat at noon, it will become active starting about 9 a.m. The cross-species generality of this result — birds and fish show the effect, for example — was so great it was almost certain that humans would show the effect. I had been eating breakfast at about 7 a.m. and waking up (tired but not hungry) at about 4 a.m.

The most striking feature of this work, to me, was not that it helped solve a real problem — I already believed self-experiments could do this — but that the solution was something I had not thought of. If you have a problem, and a list of possible solutions, self-experimentation (or conventional research) can clearly help *if a solution is on the list* — you test each possibility until you find one that works. If the list does not contain any actual solutions, it isn't clear that any method can help find a solution. Breakfast was not on my list of possible solutions, yet self-experimentation had found it.

Self-experimentation seems to be better for generating ideas than more conventional ways of collecting data, just as exploratory data analysis seems to be better for generating ideas than more conventional ways of analyzing data.

Was this an instance of a general rule? Could self-experimentation often discover useful cause–effect relationships that the experimenter had not thought of? Over the next several years, I did more self-experiments and repeatedly found useful cause–effect relationships that I had never thought of — indeed, that no one had thought of, as far as I know. These results suggested that, yes, self-experimentation is a good way to generate ideas. The following sections describe three examples.

Morning Faces and Mood

Skipping breakfast reduced early awakening but did not eliminate it. Wondering what other causes might be, I realized the breakfast results might teach a larger lesson. Our brains were shaped to work well under Stone Age conditions. During the Stone Age, it seemed safe to assume, no one ate a meal soon after waking up, at least not often, so it made some sense that eating breakfast caused trouble. Perhaps other causes of early awakening were to be found among other non-Stone Age features of my life.

When we sleep is obviously controlled by sunlight (we tend to be awake during the day), but I believe it is also controlled by social contact (we tend to be awake at the times of day that we have contact with others). During the Stone Age, people lived and slept in groups, of course, and observations of technologically primitive cultures suggest that the average Stone Age morning began with considerable face-to-face contact and conversation. In contrast, I lived alone and might work alone all morning. Perhaps lack of morning human contact caused early awakening.

To test this idea, I took advantage of results suggesting that TV affects sleep in the same way as human contact (e.g., if you stay up late watching TV, you will stay awake later the next night than if you stay up late reading). One morning in 1995, I watched about 20 minutes of TV (a tape of the Leno and Letterman monologues) soon after getting up. It was mildly amusing but seemed to have no other effect. The rest of the day was unexceptional. The next morning, however, I woke up and, to my astonishment, felt *great* — cheerful, calm, yet full of energy. I could not remember ever feeling so good early in the morning. The only unusual event in the immediate past had been the previous morning's TV viewing — I had never before watched TV early in the morning. Was that the cause? As unlikely as this connection seemed, further experience confirmed it: If I watched TV early in the morning my mood was much better than

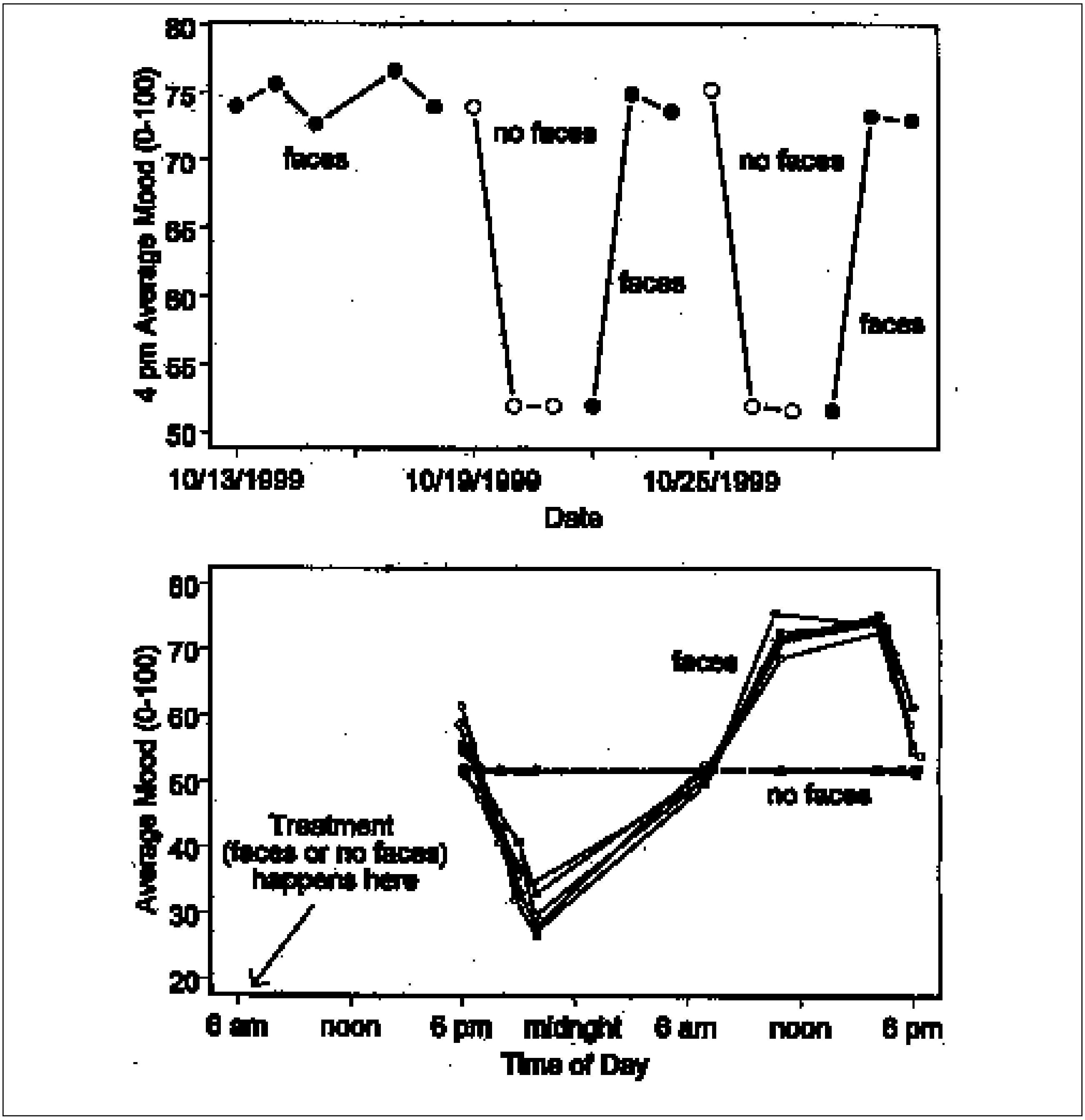


Figure 2. Upper panel: mood ratings at 4 p.m. each day. Lower panel: mood ratings throughout the day. Each line is a different day.

usual the *following* day, more than 24 hours later. It was *not* better than usual the same day.

I gradually learned several things about the effect. (1) Visual details mattered. The best stimulus was a life-sized face with both eyes visible at a distance of roughly one meter — what you would see during a conversation. Deviations from this ideal reduced or eliminated the effect. Other aspects of TV viewing, such as the sound of voices, had no detectable effect. (2) Duration mattered. Sixty minutes of faces produced a much bigger effect than ten minutes of faces. (3) Time of day mattered. The best time of day was early in the morning; faces seen an hour before or after the best time were less effective. Faces at night *lowered* my mood the next day.

Figure 2 shows results from an experiment done to show the basic effect. During each morning of the experiment, I watched a 27-inch TV starting about 6:00 a.m. I watched until I had seen 60 minutes of life-size faces; the median stopping time was 7:30 a.m. Mostly I watched *Washington Journal* (C-SPAN) and videotapes of *Booknotes* (C-SPAN), *The Newshour with Jim Lehrer* (PBS), *Charlie Rose* (PBS), *Larry King Live* (CNN), and *The O'Reilly Factor* (Fox News). When watching tapes, I usually

Table 1 — Correlation Between Standing and Early Awakening

When	Standing (hr.)	Days	Days w/ early awakening the next morning	Percent
May 18, 1996- August 26, 1996	not measured	100	57	57
August 27, 1996-October 24, 1996	5.0–8.0	20	12	60
	8.0–8.8	34	5	15
	8.8–11.0	5	0	0
October 25, 1996-February 28, 1997	5.0–8.0	10	6	60
	8.0–8.8	8	2	25
	8.8–11.0	90	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.

skipped portions that did not consist of one face filling the screen and looking at the camera. During some days ("no faces") the upper two-thirds of the TV screen was covered; during other days ("faces") it was uncovered — that is, normal. In other ways, the two sets of days were the same.

I rated my mood on three scales — happy/sad, calm/irritable, and eager/reluctant — several times each day. Each scale went from 10 to 90 (higher ratings = more positive), with 10 = very negative (very sad, very irritable, very reluctant), 20 = quite negative (quite sad, quite irritable, quite reluctant), 25 = negative, unmodified (sad, irritable, reluctant), 30 = somewhat negative (e.g., somewhat sad), 40 = slightly negative (e.g., slightly sad), 50 = neutral (e.g., neither happy nor sad), 60 = slightly positive (slightly happy, slightly calm, slightly eager), 70 = somewhat positive, 75 = positive, unmodified (happy, calm, eager), 80 = quite positive (e.g., quite happy), and 90 = very positive.

Ratings on the three scales were similar, so Fig. 2 shows averages of the three scores. The upper panel shows the average rating at 4:00 p.m. Faces increased mood the next day but not the same day. The lower panel shows how the effect varied throughout the day. Starting at about 6 p.m. — about 12 hours after the treatment — an oscillation in mood (down, then up) began that lasted about 24 hours. Before 6 p.m. the treatment had no detectable effect (data not shown).

These results are interesting partly because the main symptoms of depression are the opposite of the effects of seeing morning faces — a depressed person is unhappy, does not want to do anything, and is often irritable. Moreover, depression is strongly correlated with sleep difficulties, especially staying up late. Researchers found a *forty-fold* increase in the risk of developing major depression in persons who reported insomnia in two interviews a year apart compared to persons who reported no insomnia at both interviews. Maybe depression is often due to seeing faces for too little time in the morning and/or too much time at night.

Standing and Sleep

Because morning faces had a powerful effect on mood, I assumed that they would also have a powerful effect on sleep and that the right "dose" would eliminate early awakening. I tried many different variations on the theme of morning faces over the next year but never achieved this result — at least showing that expectations had little effect.

Around this time, my interest in weight control led me to wonder about the connection between walking and weight. It is well known that a large amount of walking often causes weight loss. When you walk more than usual you probably stand (place all your weight on your feet) more than usual. It might be standing, not movement, that causes weight loss. I decided to test this possibility by standing (but not walking) much more than usual.

In August 1996, I began. I raised my computer to work standing up, stood during phone calls, and, when possible, walked instead of riding a bike. It was hard at first but after a few days became much easier.

I did not lose weight. After about a week of extra standing, however, I noticed I was waking up early much less often. At first I assumed that any large amount of standing would provide this benefit. After a few months, however, I examined the connection between duration of standing and early awakening. The upper half of Table 1 shows the results. Standing obviously helped, but about eight hours seemed to be needed to see a big improvement. After I saw these results I began to stand much more, and my early awakening nearly vanished (lower half of Table 1).

I later discovered that early-morning exposure to an hour of bright light with the spectrum of sunlight had the same effect as about two hours of standing. The combination of eight hours of standing and an hour of bright light in the morning (another Stone Age-like solution) eliminated early awakening completely.

Sugar Water and Weight

The weight loss shown in the lower panel of Fig. 1 was not a surprise, but the ease with which it could be detected was. The idea that processing makes food fattening, derived from animal research, had not been applied to humans, so there had been no reason to think the effect would be so clear. The fact that I had no trouble sustaining the weight loss for years was also encouraging. In contrast, the most studied method of producing weight loss — reducing caloric intake without changing what is eaten — causes weight loss that is rarely sustained.

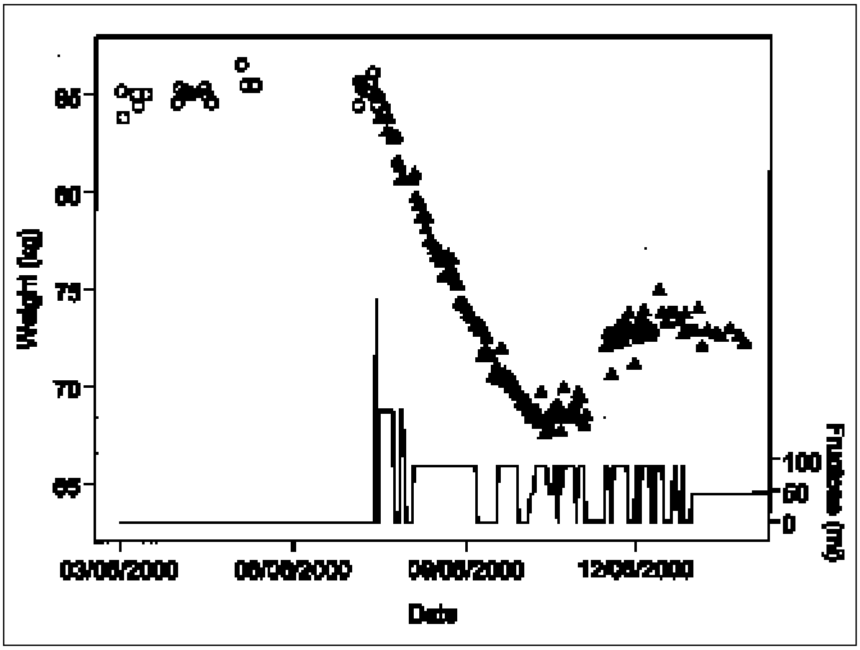


Figure 3. Effect of drinking sugar water on weight. Each point gives the average of the readings of three scales. Fructose amounts are volumes of crystalline fructose; the fructose was dissolved in .5–2 liters of water.

The clarity encouraged me to test other ideas about weight control on myself. I found that drinking five liters of water per day caused me to lose 3 kg. Apart from the difficulty of drinking so much water, it was easy to maintain the lower weight. When I stopped drinking extra water, I regained the lost weight. I found that eating a low-glycemic-index diet reduced my weight by 3 kg compared to the low-processing diet. (The glycemic index of a food is a measure of how quickly its carbohydrate is digested. Foods with a high glycemic index include bread, potatoes, and sweets; foods with a low glycemic index include beans and lentils.) I never regained the lost weight. I found that eating a diet of mostly sushi without wasabi plus fruits and vegetables caused me to lose about 6 kg compared to the low-glycemic-index diet; once again, it was easy to maintain the lower weight apart from the difficulty of eating lots of sushi. When I stopped eating lots of sushi, I regained the lost weight.

In June 2000, I visited Paris. The food was excellent. To my dismay, I had little appetite, for no obvious reason. Wondering why, I realized that some weight-control ideas of mine suggested an answer. An old and well-established idea about weight control is that your body tries to maintain a certain amount of fat the same way a thermostat-controlled heating system tries to maintain a certain temperature: When the actual amount of body fat goes below the "set point" level, changes (especially more hunger) occur that tend to push the amount of body fat back up to the set point. My addition to this theory is two assumptions. The first is that your set point is determined by the tastes of what you eat. The more strongly a taste is associated with calories, the more the taste raises the set point. Tastes become associated with calories in much the way Pavlov's dogs learned to associate a ringing bell with food — by repeated pairings. The first time you drink a Coke sweetened with sugar, its taste will *not* be associated with calories. By the tenth time you drink it, its taste *will* be associated with calories. The second assumption is that when you are not eating, your set point falls. It was hot in Paris, and I had been drinking several sugar-sweetened soft drinks each day. I drank foreign brands with unfamiliar tastes. Because the tastes were unfamiliar, they had not yet become associated with calories, I reasoned, and therefore did not raise the set point. The drinks tasted sweet, of course, but maybe the sweet taste was relatively mild. I knew that eating food with weak flavors (e.g., sushi without wasabi) can cause weight loss. Weak flavors form weaker taste–calorie associations than strong flavors.

According to this explanation, drinking unflavored sugar water — water with a substantial calorie content but no taste besides sweetness — should cause weight loss. When I returned home, I tested this prediction. After 10 baseline days with my usual diet, I started drinking fructose-sweetened water every day. I used fructose instead of sucrose (the sugar in the Parisian soft drinks) because it is digested more slowly. On the first day, I consumed 12 ounces (24 tablespoons) of crystalline fructose, about 1,100 kcal, dissolved in two liters of water. The loss of appetite was immediate and so profound that I drank no fructose water the next day. I reduced my fructose intake repeatedly, finally settling on three ounces (275 calories) per day. Figure 3 shows how my weight changed.

While losing weight not only did I almost never *go* hungry — that is, stop eating while I was hungry — I almost never *felt* hungry, at least between meals. I skipped over half of my usual meals without discomfort. After beginning a meal, I usually wanted to continue (the appetizer effect) but never felt strongly about it. The near-total absence of a familiar sensation (hunger between meals) reminded me of stories about becoming color blind.

After I lost about 18 kg, and stayed at the lower weight for a few weeks, I noticed that the negative comments ("are you healthy?" "don't lose any more") clearly outnumbered the positive. The unanticipated had happened: I was too thin. I took advantage of a trip to New York to gain 5 kg. It was easy to stay at the new weight — 13 kg (30 lbs) below the weight I started at — even when I cut my daily intake of fructose in half.

My explanation of what happened in Paris was both right and wrong. Yes, it was the sugar-sweetened unfamiliar sodas that caused the loss of appetite. But it was not because the familiar portion of the taste (sweetness) was weak. Fructose water caused *too much* weight loss, far more than the same number of calories per day of bland food. For instance, if I had added 275 calories/day of mild-tasting sushi to my diet, I might have lost 1–2 kg, but not 17 kg. The results are understandable, however, if one of my weight-control assumptions is amended: although the set point is, in general, raised by calorie-associated tastes, sweetness is an exception. Whether calorie-associated or not, it does not raise the set point. If you ingested all your calories without *any* taste, you would become very thin, according to my theory; ingesting a fraction of your calories in a way that doesn't raise the set point is a step in that direction.

As surprising as these results may be — the conventional idea is that sugar *causes* obesity, of course — they have some precedent. Many surveys have found a reliable negative correlation between sucrose intake and a measure of obesity; none has found a reliable positive correlation.

Self-Experimentation: Pro and Con

Some strengths are obvious. Self-experimentation makes it much easier to test new treatments, new ways of doing things. However hard it was to stand eight hours per day for many days, it would have been much harder to have others do so. Medicine and nutrition have long histories of self-experimentation for this reason. Self-experimentation also allows the experimenter to notice change on dimensions not the focus of interest. I watched morning TV hoping that my sleep would improve; my mood improved. I stood a lot thinking I might lose weight; I slept better. Conventional experiments, which rarely measure more than a few dimensions, could easily have missed these unexpected effects.

After a new idea, a new hypothesis, has been "conceived," self-experimentation makes testing it relatively easy. My sugar-water experiment, for example, asked if a correlation reflected causality. Many medical self-experiments fall in this category. These tests usually require, of course, that expectations have little effect on the results. For the examples described here, there was plenty of support for this assumption. In the case of sleep, for instance, many treatments I had expected to solve the problem failed to do so. Millions (or is it billions?) of other people's failed weight-loss attempts had shown that expectations do not cause sustained weight loss.

Although self-experimentation is a good way to learn (generate a new idea worth testing, and test it), it is a poor way to teach — that is, communicate what has been learned to others. Self-experimentation, like any *n*-of-1 study, reveals nothing about between-person variation. There are several reasons to think this is a minor problem. One is that history supports self-experimentation. In medicine and nutrition, it has a good track record; I know of many cases in which the results pointed in the right direction and none where they misled. Moreover, examination of individual differences in psychology experiments usually shows that most or all subjects changed in the same direction, albeit by different amounts. Experiments with strong effects, such as the examples described here, are especially likely to have this feature. However, neither body of evidence (the history of self-experimentation, examination of individual differences) is easily conveyed.

When wondering how far the results from one person will generalize, it is important to distinguish between *effects* and *solutions*. It is likely that the various treatments considered here will have the same direction of effect in different people. For instance, standing much more will probably make anyone who stands only a few hours per day sleep more deeply. However, the medical literature is full of reports in which a treatment cured some patients but not others. Sleep is controlled by many environmental events, so it is likely that sleep problems, such as early awakening, have several possible environmental causes. My results suggest that sitting too much is one of those causes, but they do not suggest that it is the *only* cause. (Indeed, my results suggest that breakfast size and exposure to morning light also play a role.) It would be surprising if sitting much less cured no one's early awakening but mine, but it would also be surprising if it cured every case of early awakening.

As my self-experimentation continued, and the surprises continued (the sugar-water example occurred between submission and completion of this article), I came to realize that self-experimentation had an unappreciated strength: It was good for discovering

Table 2 — A Gap in Scientific Methodology

	Time period	
Goal	Before and during data collection	After data collection
Generate ideas	?	Exploratory data analysis
Test ideas	Experimental design	Statistics (e.g., t test)
	Clinical trials	Model fitting

When wondering how far the results from one person will generalize, it is important to distinguish between *effects* and *solutions*. It is likely that the various treatments considered here will have the same direction of effect in different people ... However, the medical literature is full of reports in which a treatment cured some patients but not others.

new cause–effect relationships. In medicine and nutrition, almost all cases of self-experimentation had involved testing or demonstrating ideas derived from other sources. The examples described here and a few others showed me that self-experimentation can *generate* ideas, not just test them. This is important because well-known scientific methods have a noticeable gap, shown in Table 2 (page 13). The basic statistical methods that most scientists learn, such as how to do a *t* test, are of course meant to *test* ideas. They are used after data collection. Other bodies of knowledge, such as principles of experimental design, are also about how to test ideas but are used before or during data collection. Techniques of exploratory data analysis, which helps users find the unexpected, are a different sort of method, better suited for generating plausible ideas than for testing them. They are used after data collection, of course. Missing are bodies of knowledge about generating plausible ideas that help you decide what data to collect — what to vary, what to measure. For example, suppose you want to generate new ideas about the cause of asthma. There is no body of knowledge (apart from asthma research) to help you decide what data to gather. Yet some sorts of data will surely be more helpful than others.

Self-experimentation, the examples described here suggest, is an instance of the missing sort of method. Another instance is combinatorial chemistry, which is designed to find new drugs. It consists of techniques that help chemists generate a kind of factorial design of new chemicals, each of which is tested for biological activity. Of course, self-experimentation and combinatorial chemistry can only be used with a small range of problems. What about other problems? Self-experimentation and combinatorial chemistry, I believe, illustrate a general rule of idea generation — namely, *ease of search*. A method will be good for generating plausible new ideas if it makes it easy to search a big space of possible ideas for the plausible ones. Self-experimentation makes it easy — or, at least, much easier — to search a big space of cause–effect relationships because (a) it is easy to try new causes (i.e., new ways of doing things) and (b) many possible effects are "searched" at once — that is, any experiment can detect change on many dimensions. For example, to search a space of two causes by 20 effects (40 cause–effect pairs) might require only two self-experiments; to search the same space using conventional experiments would be much harder. Combinatorial chemistry makes it easier to search a big space of possible new drugs. Plotting data — the main technique of exploratory data analysis — makes it easy to search a big space of possible summaries of the data.

References and Further Reading

Altman, L. K. (1987), *Who Goes First? The Story of Self-experimentation in Medicine*, New York: Random House.

Ford, D. E., and Kamerow, D. B. (1989), "Epidemiologic Study of Sleep Disturbances and Psychiatric Disorders: An Opportunity for Prevention?" *Journal of the American Medical Association*, 262, 1479–1484.

Halmos, P. R. (1975), "The Problem of Learning to Teach: I. The Teaching of Problem Solving," *American Mathematical Monthly*, 82, 466–470.

Hill, J. O., and Prentice, A. M. (1995), "Sugar and Body Weight Regulation," *American Journal of Clinical Nutrition,* 62 (supplement), 264S–274S.

Roberts, S., and Neuringer, A. (1998), "Self-experimentation," in *Handbook of Research Methods in Human Operant Behavior,* eds. K. A. Lattal and M. Perrone, New York: Plenum, pp. 619–655.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Comment:

Lessons from Self-Experimentation: Counting for Content, Inspecting for Insight

Robert Rosenthal

Seth Roberts's report has provided us with two interesting continuities applied to systematic observation of the self. The first of these continuities is the continuity of *counting for content*; the second is the continuity of *inspecting for insight*.

Counting for Content

Roberts follows and further illuminates a rich tradition in the history of counting for content with such forebears as, for example, Gustav Fechner (1801–1887), who helped found the field of psychophysics by intensively judging the perceived heaviness of varying weights. A generation later, Francis Galton (1822–1911), who must have ranked among the greatest counters of all time, learned how many brush strokes it took to have his portrait painted, as well as founding the field of fidgetometry in which the fidgets per minute of attentive audiences were compared to those of bored audiences at the lectures he attended. A generation after that, Hermann Ebbinghaus (1850–1909) helped found the scientific study of memory by inventing nonsense syllables (for example, fid, lub, yar), and memorizing lists of them to study the effects, for example, (a) of amount of material on learning speed, (b) of repetition on retention, and (c) of elapsed time on forgetting.

Inspecting for Insight

In addition to Roberts's continuing in the tradition of counting for content, he has continued in a much more recent tradition of inspecting for insight. Working very much in the spirit of John Tukey's classic 1977 book, *Exploratory Data Analysis*, Roberts illustrates a far newer look in the analysis of psychological data than we are used to seeing in the social and behavioral sciences. It is that newer look that encourages approaching data in an exploratory spirit more than, or at least in addition to, a confirmatory spirit. It is a spirit that makes friends with the data, holds it up to the light in different ways, and thinks of data analysis, at its best, as the opportunity to confront a surprise.

This spirit, so well reflected in Seth Roberts's report, is also remarkably consistent with the views of the American Psychological Association's Board of Scientific Affairs' Task Force on Statistical Inference. That task force, for which statisticians Fred Mosteller and John Tukey and psychologists Lee Cronbach and Paul Meehl served as senior advisors, concluded that the psychological sciences would be further ahead if data analysis were seen less as a process of sanctification and more as a process of detective work (borrowing terms from Tukey's 1969 article in *American Psychologist*).

Supplementing the Sample Size of "Selves"

Roberts's report is rich in propositions to be examined further within other sampling units, either as more traditional larger sample research or as a series of $N = 1$ researches with additional "selves."

Since there are so many selves in the world, if more of us became self-experimenters perhaps more would be learned more quickly in the social, behavioral, and biomedical sciences. One could imagine a time in the future when "self-experimenter" became a new part-time (or full-time) profession. One could imagine insurance companies' prevention programs, HMO's, and other organizations employing promising self-experimenters on modest retainers but with the possibility of large serendipity bonuses for promising discoveries then to be replicated in more conventional studies. Whatever the future may hold for the wider practice of self-experimentation, one can only applaud the experimental designs and procedures employed by Seth Roberts and the open-eyed and open-minded approach to his analysis of his results.

Note: Since the American Psychological Association task force also pushed hard for effect size estimation as a standard product of research, and because no two columns of numbers should ever go uncorrelated, I provide the Pearson correlation, based on Roberts's Table 1, between the six levels of hours of standing (using midpoints of the ranges shown) and the percent of days with early awakening (r = -985). Taken separately for the early and late periods of research, these correlations were –.977 and –.999, respectively; for the three levels of hours of standing averaged over the early and late periods of research, the correlation was –.992. These calculations are presented in the spirit of demonstrating that there can be secondary analyses of data from self-experimentation.

Comment:

Self-Experimentation for Causal Effects

Donald B. Rubin

Self-experimentation — what does that have to do with the academic field of statistics? It has a lot to do with it because thinking carefully about the issues raised by the type of self-experimentation Seth Roberts discusses leads to a better understanding of the foundations of causal inference.

Let's begin by considering what the causal effect is for one "unit" — that is, one object — let's say Seth Roberts at one point in time. He wants to reduce his acne and is considering whether to use a pill or a cream that evening. "What will my acne be like tomorrow morning if I take the pill?" "What will my acne be like if I use the cream?" If he could have answers to both of these questions, he would use the product that would lead to less acne in the morning. But he cannot get answers to both questions; the best he can do is to choose one product and observe the result. The causal effect, however, is the comparison of the observed result under the chosen treatment with the unobserved result under the unchosen treatment. How does Seth learn about the causal effect, which involves the comparisons of two "potential outcomes" from the observation of only one?

The answer — replication, more units. Now in statistics replication usually means more objects, as in a big randomized experiment with half the people assigned to one treatment and half to another. Two problems face Seth. First, he cares most about what works on him and not on others, although he'd certainly like to know what appears to work on others because it would suggest an answer for him. Second, he can't conduct such a trial without a major effort, whereas he can conduct his own self-experiment with very little effort. Self-experimentation also involves replication but of the same object (Seth) repeatedly in time.

Suppose Seth contemplates using two units — that is, contemplates repeating his self- experiment using two periods. Then there are two potential outcomes at the end of the first period, (1) the state of his acne with the pill [P] and (2) the state of his acne with the cream [C], and four potential outcomes at the end of the second period, (1) the state of his acne if he had taken the pill the first and second times [PP], (2) the state of his acne if he had used the cream both times [CC], (3) the state of his acne if he had taken the pill first and then used the cream [PC], and (4) the state if he had used the cream first and then taken the pill [CP]. Yet Seth gets to observe only one potential outcome after the first period and only one potential outcome at the end of the second period. How does the replication help him, especially since there may be variation in the effectiveness of different applications of the pill or the cream? We need two available doses of each to be able to contemplate all potential outcomes: What if pill-1 works well but pill-2 does not?

He must make an assumption; typically the assumption to be made is called "stability," or the "stable-unit-treatment value assumption" (SUTVA). Under stability, the potential outcomes for each unit just depend on the treatment assigned to that unit. That is, first, there is no variability in the efficacy of the treatments. Furthermore, the units do not interfere with each other so that in the second period the potential outcome associated with the "pill" is the same no matter what happened in the first period, and similarly for cream. In other words, the change in the state of Seth's acne during the second time period does not depend on the treatment received in the first period. Under this assumption, all we need to contemplate are the two potential outcomes for unit 1 and the two potential outcomes for unit 2. In fact, under stability, no matter how many units we have, we can represent their potential outcomes using only two columns, one for P and one for C, rather than having to consider a bewildering array of possible outcome values.

Now, this stability assumption could be (almost certainly is) wrong in the present context; for example, there could be real or imagined carryover effects from one time period to the next. Moreover, some doses may be more effective than other doses. But the stability assumption is commonly used in many settings — for example, in clinical trial designs for medical research. The stability assumption may be more plausible there because the units are distinct people who don't know each other and presumably cannot interfere with each other. The key point is that some such "exclusion restriction," which excludes variation in certain potential outcomes, is needed for causal inference.

This general perspective for causal inference is sometimes called "Rubin Causal Model" for work of mine that applied the potential-outcomes perspective to both randomized and nonrandomized studies and allowed forms of inference beyond those that were randomization based.

Let's now suppose that we make the stability assumption in Seth's case, and let's also suppose he is contemplating a large number of units off into the future, say for a few years. The potential outcomes will be represented in two columns of values, the first column for "pill" and the second column for "cream." What Seth wants to do is to choose the column that will be best for his acne, in the sense of having the better typical response. Now further suppose that Seth effectively randomizes his choice of which treatment to apply to the units. (Seth was silent on this — did he choose treatments by tossing a coin? Let's assume that he did.) Then he is essentially in the classical setting of an experiment. Of course, he cannot learn about other people (except by assumption), but he can learn about himself under the stability assumption.

Of course, thus far we have not considered the fact that the treatments, whose effects we're trying to estimate through self-experimentation, include the psychological effects of knowledge of the treatment given — that is, they include "expectancy effects." This is also true in any experiment in which the people are not "blinded." For Seth, this is probability fine — if pills worked better than cream because he thought they would, that works for Seth. (In fact, Seth says he discovered the opposite of his expectancy.) This, however, is not the same thing as knowing which of the *blinded* treatments would work better, which is usually the objective in a clinical trial of drugs.

Now let's consider the ability of self-experimentation to *discover* cause-and-effect relationships. If a new treatment generates a large enough response — large with respect to the natural variation seen in units in the past — then it is a rare event, *unless* we modify our model to allow for the possibility that the new treatment's effect is different from what we've seen in the past. This is the same logic that underlies a traditional test of significance in a randomized trial: Assume the null hypothesis of no effect (and no trends in time) and calculate the probability of observing something this extreme. If the event appears to be too extreme, we'd rather believe it is not extreme under a new model that allows for the new treatment to be more effective than the previous treatment (but still maintains the no-trend-in-time hypothesis).

The conclusion is that I find myself in agreement with Seth Roberts that self-experimentation can be useful for estimating cause-and-effect relationships — it better be because that's how we learn most lessons in life. I am less convinced that discoveries of self-experimentation are as infallible as indicated ("and none where they have misled") or that traditional methods aren't helpful to generate ideas. For example, observational studies, which have seen an explosion of formal statistical activity in recent years, are designed to be an inexpensive alternative to randomized experiments. Databases with thousands of distinct people, such as SEER, NMES, NHANES, and so forth, are often used to search for evidence supporting a new theory. The limitations of these databases are with respect to the number of possible treatments and possible outcome, as well as the well-known limitation with respect to lack of randomization. But a primary motivation for the existence of such databases, called "population-laboratories" years ago by W. G. Cochran, is to allow future researchers to study them to try to find evidence of the type of causal and effect relationships that Seth Roberts seeks to find through self-experimentation.

I enjoyed this article, and I think part of a lecture on the role of self-experimentation will become part of my course "Causal Inference."

References and Further Reading Added By Discussants

Boring, E. G. (1950), *A History of Experimental Psychology*, New York: Appleton-Century-Crofts.

Cochran, W. G. (1952), "An Appraisal of the Repeated Population Censuses in the Eastern Health District, Baltimore," in *Research in Public Health*, New York: Milbank Memorial Fund, pp. 255–265.

Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960.

Rosenthal, R., and Jacobson, L. (1968), *Pygmalion in the Classroom*, New York: Holt, Reinhart and Winston.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

——— (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 7, 34–58.

——— (1980), Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test" by A. Basu, *The Journal of the American Statistical Association*, 75, 591–593.

——— (1990), "Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, 5, 472–480.

Stigler, S. M. (1986), *The History of Statistics*, Cambridge, MA: Harvard University Press (Belknap).

Tankard, J. W., Jr. (1984), *The Statistical Pioneers*, Cambridge, MA: Schenkman.

Tukey, J. W. (1969), "Analyzing Data: Sanctification or Detective Work?" *American Psychologist*, 24, 83–91.

—— (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley.

Wilkinson, L., and the Task Force on Statistical Inference, APA Board of Scientific Affairs (1999), "Statistical Methods in Psychology Journals: Guidelines and Explanations," *American Psychologist*, 54, 594–604.

Rejoinder

Seth Roberts

I thank Professor Rosenthal for his kind words. The possibility he imagines at the end of his piece — "insurance companies' prevention programs, HMOs, and other organizations employing promising self-experimenters on modest retainers" — has in part come to pass, in the sense that self-experimentation has already substantially reduced medical costs. In 1969, Richard Bernstein, an engineer, purchased a new device that could measure blood glucose level with just one drop of blood. Its intended use was to allow emergency-room personnel to determine if an unconscious person was diabetic. Bernstein, who had diabetes, realized he could use it to study his own glucose levels. He discovered that his glucose level varied far too much, even though he was carefully following the conventional recommendations. By trial and error he found a new schedule of insulin injections and a new diet that together kept his blood glucose level much closer to optimum. His work eventually led to the widespread practice of blood-glucose self-monitoring, with products (meters, needles, and test strips) found in every drugstore. Glucose self-monitoring has helped millions of diabetics stay healthy — and probably saved health-care providers billions of dollars. This makes Rosenthal's suggestion all the more reasonable.

Professor Rubin does a nice job of showing how self-experimentation can provide a context for discussion of basic assumptions. My self-experimentation took me into research areas I knew little about, and as it progressed I came to appreciate the value of making underlying assumptions explicit and testing them before putting weight on them. The overriding lesson when deciding what experiment to do next was *take the smallest possible step forward*. That is, do the simplest, easiest experiment that will provide new information. The more "progress" from what had already been done, the more assumptions being made. Assumptions not already verified, I found, had a good chance of being wrong.

I learned this rule many times. As it sunk in, I came to see that everything I had been taught about experimental design had been misleading, at least in this context, because the benefits of various choices had been made clear and the costs, often large, had not. Typical textbooks do not even discuss costs. Randomization, which Rubin mentions, is a good example. In the early days of my self-experimentation, while I still took seriously what I had been taught — before I took care to use the simplest possible designs — I did an experiment that included randomization. Each day I randomly chose one of two treatments. I stopped after a week or so. The biggest problem was that randomization made it much harder to look at the results and see anything interesting — agreement *or* disagreement with prediction — because it became impossible to see at a glance what the prediction was. I switched to a design in which treatment days and baseline days slowly alternated (several days per block) several times (e.g., the mood experiment of Fig. 2 in my article). The results from such a design could be understood at a glance. For instance, it was easy to judge the assumption of treatment stability, which Rubin discusses. Discordant results could be noticed immediately rather than at the end of the experiment, which helped identify important factors I had not known about. (If, say, my acne was much worse than expected, it was no help to learn this two weeks later when I would have forgotten what I had eaten at the time.) Not only was something important gained from the nonrandomized design, little was lost. The alternative explanation ruled out by randomization — something else alternating at the same phase and frequency as the treatment — was too implausible to worry about; moreover, it could easily be tested later. I had failed to appreciate that randomization is worthwhile only if the alternative explanations it makes less likely are plausible and not easily tested. What Rubin has noticed, I think, is that self-experimentation offers (or requires) more choices than most research because it is more novel and more flexible and thus provides a fresh look at questions (such as whether to randomize) whose answers are often taken for granted.

Reference and Further Reading

Bernstein, R. K. (1997), *Dr. Bernstein's Diabetes Solution*, Boston: Little, Brown.