The Growth of Personal Science: Implications for Statistics

Seth Roberts

Tsinghua University and University of California, Berkeley

Summary

Personal science is science done for personal reasons (to help yourself) rather than professional ones (as a job). The most common personal science is health self-measurement, which has recently become much more popular. This article describes 14 examples of personal science involving health. The topics include blood sugar, sleep, mood, body weight, resistance to infection, and brain function. Most of the examples are about new ways to improve these measures. For example, the results suggest that: 1. Skipping breakfast reduces early awakening. 2. Looking at faces in the morning improves mood. 3. Flaxseed oil improves balance. 4. Butter improves arithmetic speed. Overall, the results suggest that personal science plus expert advice can produce better health than expert advice alone. Personal science may influence statistics in two ways: 1. A new audience. Personal scientists want to learn statistics. 2. Better understanding. Learning about personal science may help statisticians understand science in general.

Introduction

Over the last few decades, new technology has made data collection much easier. Statisticians have written about the increase of large datasets (e.g., Nolan and Temple Luce, 2010, discussions of Big Data) but the increase of small datasets may eventually matter more. This article is about one type of small dataset -- health-related self-measurements, such as measurements of blood sugar, sleep or weight -- that has become far more common. Statisticians are familiar with data collected as part of a job (e.g., collected by a professor). Health self-measurement is not part of a job. People do it to improve their own health. I call science done to help yourself *personal science*. How important might it become? Will it influence statistics?

Let's consider an example. Suppose my blood sugar is too high. I can see a doctor. He may prescribe a drug. That's expert advice (probably based on professional science). Another possibility is to collect data (measure my blood sugar) and self-experiment (test possible remedies). That's personal science. I can do both: what my doctor recommends *and* personal science – for example, I can compare what my doctor recommends to other possible remedies.

It's not obvious that personal science will help me. Maybe my doctor's advice is the best I can find. Nor is it obvious that personal science *won't* help me. Both possibilities (won't/will help) are plausible. Professional scientists have big advantages over personal scientists. Billions of dollars have been spent on diabetes research, all of it going to professional scientists. They have more resources (money, labs, expensive equipment), more training, and more experience than personal scientists. On the other hand, personal scientists have big advantages over professional scientists. They have more freedom. They are under no pressure to publish or get grants. They can test any remedy, not just respectable or profitable ones. (American health care has been heavily shaped by pharmaceutical research.) Personal scientists are single-minded. They care only about improving their own health, whereas professional scientists have other goals -- prestige, salary, job security, respect of colleagues, and so on – which may interfere with finding the best possible way to improve health. Personal science also benefits from assured relevance. Whatever drug my doctor prescribes, it was developed studying other people. They may not resemble me. The research context (e.g., diet, exercise) may not resemble my life. Personal science studies exactly the person of interest.

Personal science has already been important in two areas. One is weight control. Several popular weight loss methods, such as the Atkins Diet and the South Beach Diet, were devised by persons who wanted to lose weight and were not professional researchers. The other is diabetes. In the last thirty years, home glucometers, which allow anyone to study their own blood sugar levels, have greatly changed treatment of diabetes. Their value was discovered by a person with diabetes (Richard Bernstein) rather than a professional researcher.

Personal science is growing. Many new products that measure health have recently been released. They measure sleep (Zeo, WakeMate, Somnus), activity (FitBit, BodyMedia, WalkingSpree), brainwaves (Neurosky MindWave), heart rate (myBasis), and blood composition (Nonin Plus Oximetry, OrSense), for example. Home blood tests (OptiMale, CardioCheck) and blood tests without doctor approval (Health One Labs, DirectLabs, PersonalLabs) have become available. West Coast Health Services offers on-the-spot cholesterol, osteoporosis and other tests every few months at locations throughout California. The MedHelp website, which has more than a million visitors per month, provides about 30 sets of rating scales for health problems. The anxiety/panic scales, for instance, help users track "symptoms, causes and treatments relating to anxiety and panic attacks."

In 2006, Kevin Kelly and Gary Wolf, both associated with *Wired* magazine, started a Meetup group in San Francisco called The Quantified Self (meaning self-measurement). Most of the self-measurement involved health. By 2012, there were affiliated groups in more than 50 cities (San Diego, London, Mannheim, Helsinki, Singapore, etc.). In 2011, at the first Quantified Self conference, speakers discussed heart rate, mood, sleep apnea, Crohn's disease, attention, facial expressions, blood tests, and many other sorts of health-related data. In 2012, Northwestern University started a doctoral program in Personal Health Informatics. "Personal health technologies," says a description of the program, "are those that non-health professionals interact with directly, both in and out of a clinical setting and in various life stages of illness and wellness."

Will the growth of personal science influence statistics? This article tries to answer this question by providing examples. Personal science may shape statistics in two broad ways. First, *it may provide a new audience* for statisticians. The examples of this paper suggest that doing personal science (including data analysis) can produce better health than just following expert advice (which is what almost everyone now does). The examples involve common problems such as high blood sugar, poor sleep, poor mood, common colds, obesity, and suboptimal brain function. The fraction of people with at least one of these problems (99%?) must be much larger than the fraction of people who now study statistics at some point (1%?). If people with these problems believe that personal science can help them, they will want to learn how to analyze the data they collect. Second, *it can help statisticians understand science in general.* Statistics has been shaped entirely by professional science. Personal scientists have different needs and ask different questions. For example, they are more interested in finding ideas worth testing. The Discussion makes explicit what statisticians may learn about science from personal science.

Background

Personal science is an example of DIYization, where something done by specialists as a job (in this case, health research) begins to be done by non-specialists not as a job. Photography illustrates DIYization. At first, only professional photographers took photographs. Now everyone takes them. Computers have DIYized many jobs, including word processing and graphic design. We can predict the effects of personal science by looking at other examples of DIYization. DIYization usually has two benefits: 1. *Spread*. More people benefit because the cost goes down. Word processing software costs less than a secretary. 2. *Innovation*. Innovation increases in small ways (customization – you alter something to fit your specific needs) and large ways (new processes and products are created) because more people can use X to innovate. Von Hippel (2005) described how customization by customers has helped companies improve their products.

Another indication of the future effect of personal science is the history of amateur science. Amateur science is more than personal science. It includes what might be called *hobbyist science*, science without personal benefit. Hobbyist scientists, like personal scientists, have more freedom than professional scientists. Three hobbyist scientists (Charles Darwin, Gregor Mendel, and Alfred Wegener) show what a difference this can make. Darwin spent twenty years writing *Origin of the Species*. He did not lose his job for low productivity. Nor did it matter whom he offended. Mendel too produced very little, with no bad consequences. He proposed a radical new theory without worrying what his boss or co-workers would think. The same applies to Wegener. His theory of continental drift was ridiculed for many years.

Personal science has already revolutionized treatment of diabetes (Bernstein, 2003), as I mentioned earlier. In the 1960's, Richard Bernstein was an engineer with Type 1 diabetes. The usual insulin injections, adjusted based on monthly blood sugar measurements, worked poorly. His blood sugar was often too low and too high and his health was poor. Then he learned of a new blood sugar meter that needed only a drop of blood. It was meant for doctors, but, because his wife was a doctor, he was able to get one. He measured his blood sugar several times per day. The results taught him how to control it. His health greatly improved. Home blood glucose measurement similar to what Bernstein did is now standard for diabetes.

My personal science began in graduate school. I was studying psychology and wanted to learn how to do experiments. I started doing self-experiments to get more practice. (Self-experiments were much faster than my usual research, which involved rats.) One topic was acne. I had acne. My dermatologist had prescribed tetracycline (an antibiotic) and benzoyl peroxide. In spite of using them, I still had plenty of pimples. I did experiments to measure their effect. I varied the treatment (e.g., the number of pills/day of tetracyline) and counted the pimples on my face each morning. At the start of my research, I thought tetracycline worked and benzoyl peroxide did not. My results showed that the opposite was true: benzoyl peroxide worked, tetracycline did not. Later studies by dermatologists agreed with me that tetracycline may not work (e.g., Eady et al., 1993). I'd had acne for six years. A few months of self-experimentation produced a big improvement.

The improvement was nice. It was also surprising. It had been curiously easy to improve on expert advice. My dermatologist had had years of training and experience. Presumably he read dermatology journals. They reported expensive research by dermatology professors. In a few months, I had learned something very useful (tetracyline may not work) that he didn't seem to know. He had never mentioned the possibility that tetracyline wouldn't work and was surprised by my tests ("Why did you do that?" he asked). I wondered if personal science could improve on expert advice in other areas.

A few years later, I wanted to sleep better (see Example 2). I started measuring my sleep and testing possible remedies. It took me ten years to discover the first useful treatment – but that treatment, in contrast to my acne research, was unknown to sleep experts. After my success with sleep (1990), I began to make useful discoveries more often. Eventually I wrote about my findings: a *Chance* article (Roberts, 2001), a scientific article (Roberts, 2004), a popular book about weight loss (Roberts, 2006), an article about why my self-experimentation was successful (Roberts, 2010), and an article about its reception (Roberts, 2012).

Statisticians may be unaware of actual examples of personal science. I know of only two articles in the statistics literature about it: my *Chance* article (Roberts, 2001) and a *Chance* article about the graphical analysis of glucometer data (Weiner and Velleman, 2008). Neither reaches the conclusions I reach here.

Examples

The fourteen examples are arranged by topic (blood sugar, sleep, etc.), roughly from simple to complex. Examples 2, 6, 8, and 9 and some of 3 are described in more detail in Roberts (2004).

The goal of the examples is to show what is possible – in particular, that personal science can improve on expert advice. No one doubts that the power of personal science has been increasing. People can better measure health. They can search the Internet for ideas and scientific papers. They can analyze their data with free software. What *isn't* clear, however, is how close personal science is to being useful. In the 1600s, it took months to cross the Atlantic. Navigation was hard. During a trip travelers knew they were getting closer to land – they could see the ship was moving in the right direction – but trip durations (how long it took to get from one side of the Atlantic to the other) varied so much that for a large fraction of a trip the travelers had little idea how close to land they were. The first signs of land were birds.

The examples resemble those birds. My success at personal science (illustrated by the examples) does not mean that many people can now do it. I had advantages that most people don't. My job (psychology professor) taught me a lot about how to do science. But others can learn on their own what I learned from my job. This is why the examples suggest we are approaching a time when many people will do useful personal science.

Example 1: Blood Sugar and Walking

In 2008, to learn more about blood sugar monitoring, I began measuring my blood sugar every morning before eating (often at 8 am). Such measurements are called "fasting" levels and are used to diagnose diabetes. (A level of 126 mg/dL may be considered diabetes.) I used an Abbott glucometer (e.g., Freestyle Lite) and drew blood by pricking my forearm, which was painless. I did not think I was at risk of diabetes.

Figure 1 shows my measurements. In 2008, I made 157 measurements, which had a median of 91. That seemed acceptable. Values of 100-125 mg/dL are considered "pre-diabetic". Perhaps 84 is ideal. I stopped measuring for a while. When I resumed, in 2009, the numbers were worse. The first 20 measurements in 2009 had a median of 102 mg/dL.

I was alarmed. The values were not just high, they were increasing. To reduce them, I ate less carbohydrate (no rice, less fruit, and so on). The median came down but it was still high. Was I drifting toward diabetes?

I wanted to further lower my fasting blood sugar but I did not know how. I already ate few carbohydrates. I tried eating even fewer. That didn't work. I tried not eating for long periods of time (e.g., one day). That was too unpleasant. Exercise is sometimes recommended (see below), but I already did aerobic exercise three times per week.

One morning in 2009 my blood sugar was about 10 points lower than usual (in the 80s rather than the 90s). Why? I remembered I'd done something unusual the previous day: walked to and from a cafe (30 minutes each way). I often went to that café but previously had always biked (5 minutes each way). The correlation suggested that an hour of walking might lower fasting blood sugar.

I tested this idea by deliberately walking 50-60 minutes every day. Figure 1 shows what happened. My fasting blood sugar was much better and over the next two years, did not drift upwards.

“Exercise” is often recommended to prevent diabetes but “exercise” usually does not include ordinary walking. To prevent Type 2 diabetes, says the Mayo Clinic website, "aim for 30 minutes of moderate physical activity a day. Take a brisk daily walk. Ride a bike. Swim laps." A recent review (Hu et al., 2007) concluded that "30 min/d of moderate- or high-level [= moderate or high intensity] physical activity is an effective and safe way to prevent type 2 diabetes in all populations." Ordinary walking is low intensity. Another review (Gill and Cooper, 2008) points in the direction of my results: “The data indicate that protection from diabetes can be conferred by a range of activities of moderate or vigorous intensity, and that regular light-intensity activity may also be sufficient, although the data for this are less consistent.”

If you want to make similar measurements, be aware of three complications: (a) Blood sugar rises briefly around dawn. (b) Blood sugar rises in anticipation of a meal (e.g., breakfast). Measure your blood sugar after the dawn rise and before the anticipatory rise. (c) The bias of the test strips apparently grows with age up to as much as 10 mg/dL (e.g., reads 90 mg/dL when the truth is 80 mg/dL) before the expiration date. Don’t change treatments at the same time you change test strip batch.

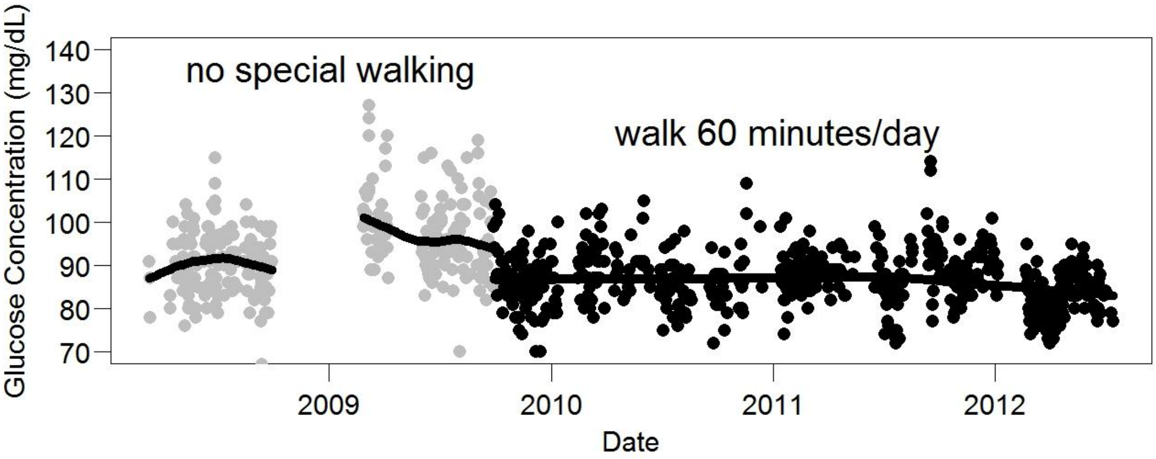

Figure 1. Blood sugar over time. Each point is one measurement (one day). The lines are loess fits.

Example 2: Sleep and Breakfast

Soon after I moved to Berkeley, I began to wake up too early -- tired but unable to fall back asleep. To measure the problem, I started recording my sleep. I defined early awakening to be any time I fell back asleep from 15 minutes to six hours after waking up. This happened roughly one-third of the time.

Over the next twelve years, I tested all of the plausible solutions I could think of, such as more exercise and eating cheese (which made me sleepy). None made a noticeable difference.

In 1990, for unrelated reasons, I changed my breakfast from a bowl of oatmeal to two pieces of fruit (e.g., banana and apple). My sleep got worse. I started to wake up early every morning, instead of one-third of the time. I went back to oatmeal. My sleep improved. I went back to fruit. My sleep got worse. Apparently breakfast made a difference.

Why was oatmeal (33% early awakening) better than fruit (100% early awakening)? The two foods differ in many ways. One is protein content: Oatmeal contains much more protein than fruit. I tried several high-protein breakfasts, hoping that one of them would be better than oatmeal. None was.

I wasn't sure what to do next. Food is so complicated. It might be easier to understand my results if I compared something to nothing (Breakfast X to no breakfast) rather than something to something (Breakfast X to Breakfast Y).

To get a no-breakfast baseline, I stopped eating breakfast. To my great surprise, my early awakening almost disappeared (Figure 2). After a few months, I resumed eating breakfast (two pieces of fruit). My sleep got worse. Again I stopped eating breakfast. Again my sleep got better.

Now I understood why my early awakening had started when I moved to Berkeley (to be an assistant professor). It was my first long-term job. *The rest of my life is beginning*, I'd thought. *Time for better habits*, which included eating breakfast. For the first time since high school, I started eating breakfast.

My discovery that breakfast caused early awakening made sense in terms of what was already known about animal behavior. A well-established effect in animals is called *food-anticipatory activity* (Mistlberger, 1994). When fed at the same time every day, mammals, birds, and fish become active about three hours earlier. If fed at noon, for example, they become active at about 9 am. I was eating breakfast at about 7 am and waking up at about 4 am. Sleep researchers have not yet noticed this effect.

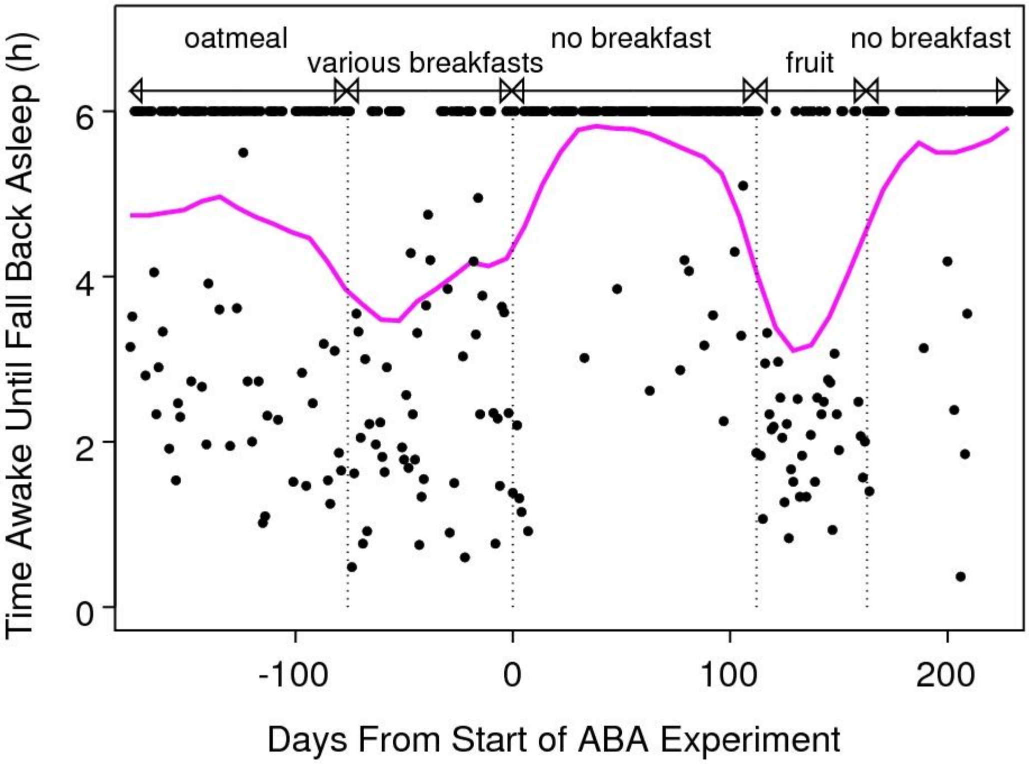

Figure 2. Early awakening over one year. Each point is one morning. The y axis indicates the time from waking up to falling back asleep. If I did not fall back asleep within 6 hours, y = 6 hours. The line is a loess fit. From Roberts (2004).

Example 3: Sleep and Standing

In 1996, a colleague (Lucia Jacobs) told me it was too bad typing didn't count as exercise. *What about standing?* I thought. Many people think exercise causes weight loss. Does standing cause weight loss?

I decided to find out. I spent much more time on my feet. For example, I walked instead of riding a bike, I read and wrote standing up, and I stood during phone calls. The first few days were exhausting but after that it wasn't hard. I measured how long I stood with a stopwatch.

After a few weeks it was clear I was not losing weight. But something else *had* changed: I was waking up early less often (upper panel of Figure 3). A few months later, in preparation for a talk, I analyzed my sleep data. This analysis showed that standing reduced early awakening only if I stood at least 8 hours (lower panel of Figure 3). After I learned this, I tried to stand 9 hours every day. Supporting the assumption of causality, the next several months produced even less early awakening (upper panel of Figure 3) and a similar dose-response function (lower panel of Figure 3).

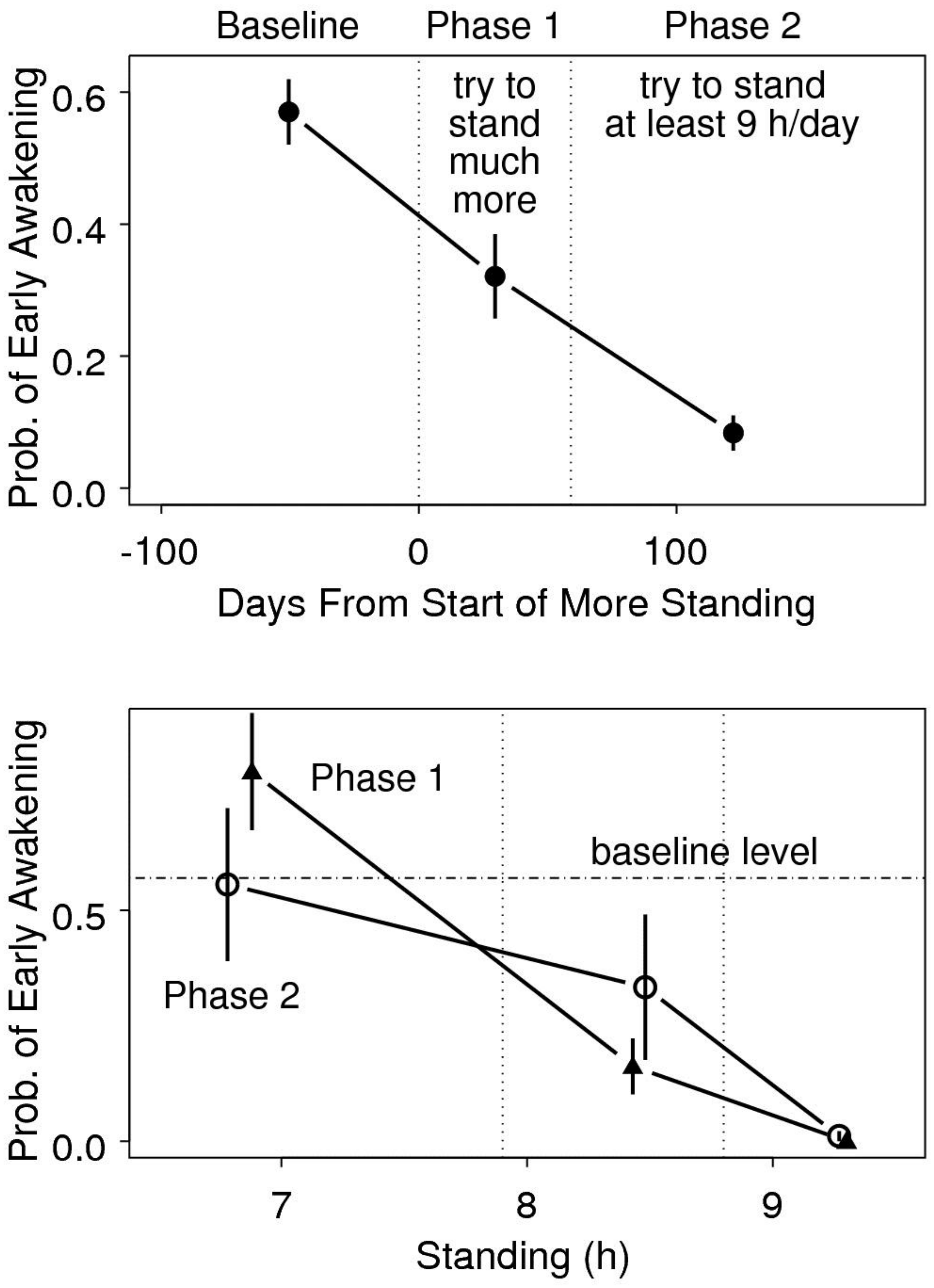

Figure 3. Effect of standing on early awakening. Early awakening = Fell back asleep within 6 hours after getting up. Vertical segments show standard errors assuming a binomial distribution.

Unfortunately standing 9 hours/day was too hard. After a few years, I stopped doing it.

One day in 2008, I woke up feeling much more rested than usual. I had not slept more than usual, so why did I feel unusually rested? I made a list of nine ways the previous day had been unusual (e.g., usually wear contact lenses while sleeping but didn't). Two days later I remembered something else. In the morning, I had stretched the hamstrings of both legs while standing up – that is, I stood on one leg and pulled the foot of the other leg behind me. The total time spent stretching had been about 4 minutes (each leg 2 minutes). That seemed too brief to make a difference, but my earlier standing results (Figure 3) made me consider it.

I tested each of the ten ways the previous day had been unusual. The only one that increased how rested I felt when I awoke was one-leg standing, more precisely standing on one leg to exhaustion (= standing on one leg until it is too painful to continue). I usually did it twice (left leg once, right leg once) or four times (left leg twice, right leg twice) per day. Time of day didn't seem to matter so long as at least four hours separated stretching bouts. (For example, if the first bout is at 9 am, the next bout is at 1 pm or later.) As my legs got stronger, it took longer to reach exhaustion. To save time, I started standing on one *bent* leg to exhaustion, which I could do for 3-5 minutes.

In 2011, I compared different amounts of one-leg standing (two, three, or four per day) in a randomized experiment. Every morning I did it twice (left once, right once). In the evening I randomly chose between zero, one, and two additional one-leg stands. Sometimes I forgot. The next morning, when I woke up, I rated how rested I felt on a scale where 0 = not rested at all (as tired as when I went to sleep), and 100 = completely rested, not tired at all.

Figure 4 shows the results for three sets of days: (a) "baseline" days (before the experiment and during the experiment when I forgot, (b) "random" days (days when I randomly chose) and (c) a later set of days ("baseline 4″) when I did four one-leg stands every day. The results suggest that three was better than two and four better than three.

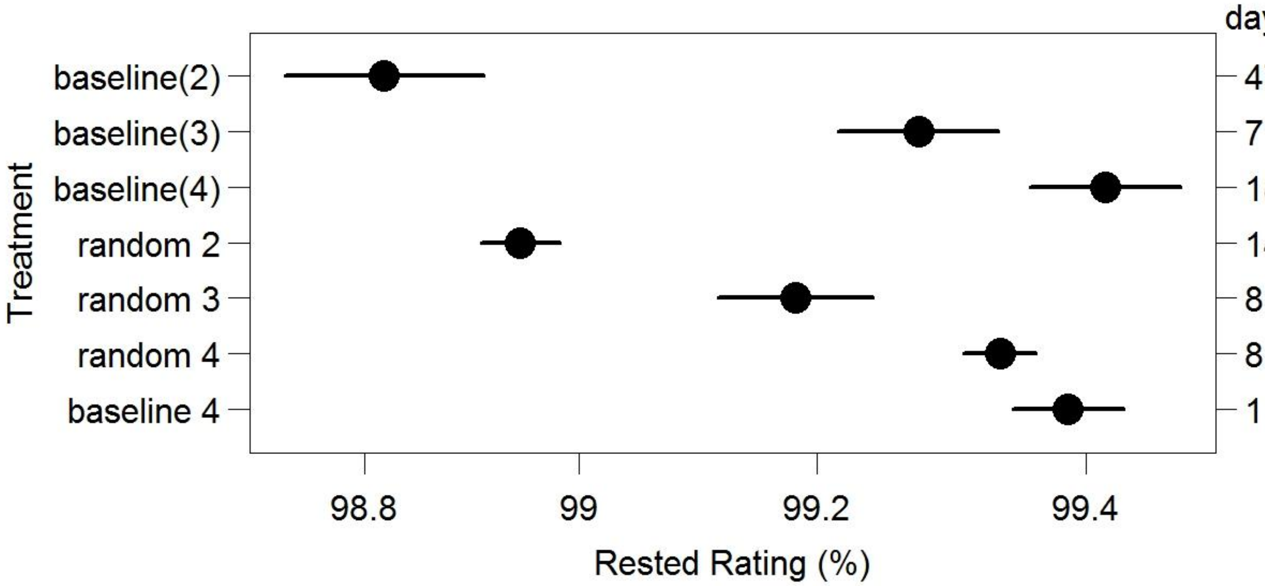

Figure 4. Effect of amount of one-leg standing on rested ratings. When I awoke I rated how rested I felt on a 0-100 percentage scale where 100% = completely rested, not tired at all, 99% = 99% of tiredness gone, and so on.

The similarity between random 4 (when the days preceding each target day were a mix of 2, 3, and 4) and baseline 4 (when the preceding days were all 4) implies that the amount of one-leg standing on previous days didn't matter much (e.g., my sleep Monday night did not depend on what had happened Sunday).

The differences in how rested I felt when awoke (Figure 4) were not reflected in how long I slept. Figure 5 shows “first” sleep durations, meaning the time from when I went to sleep to when I woke up for the first time, which is when I judged how rested I was (Figure 4). On a small fraction of days, I fell back asleep a few hours later. Because I felt more rested after roughly the same amount of sleep, these results suggest that one-leg standing made me sleep more deeply.

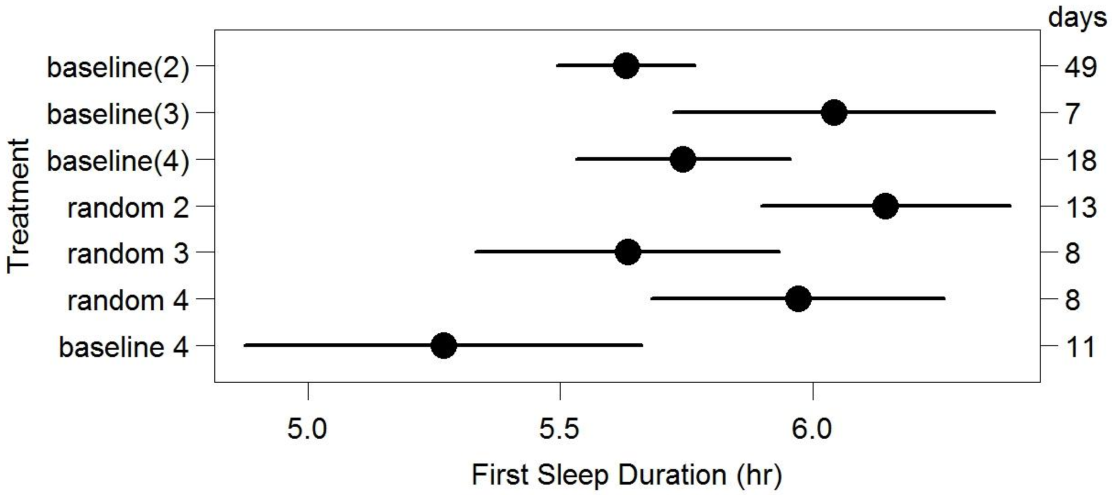

Figure 5. Effect of amount of one-leg standing on sleep duration. Averages are means. Error bars show standard errors.

Many professional studies have measured the effect of exercise on sleep (Youngstedt et al., 1997), but the exercise involved has been aerobic exercise (Reid et al., 2010) or weight lifting (Singh et al., 1997). One-leg standing to exhaustion is much more convenient.

Example 4: Sleep and Pork Fat

In 2008, I bought a box of pork from a farmer. It contained a variety of cuts. I ate the familiar ones. That left a cut I hadn't seen before, which was maybe 80% fat. I eventually learned it was pork belly. In America, pork belly is used for bacon and rarely sold unprocessed. At the time, like many Americans, I believed animal fat was bad. The pork belly repulsed me but I couldn't bear to throw it out. It sat in my freezer for months.

Finally I ate it (in soup). That night I slept longer than usual (8.3 hours). Figure 6 shows how that compares to preceding days. The previous 130 nights of sleep were under similar conditions, except that on those days I'd eaten little or no animal fat. I'd slept more than 8.3 hours on only two of them (2%). The next day I felt much more energetic than usual. I hadn't felt so energetic in years.

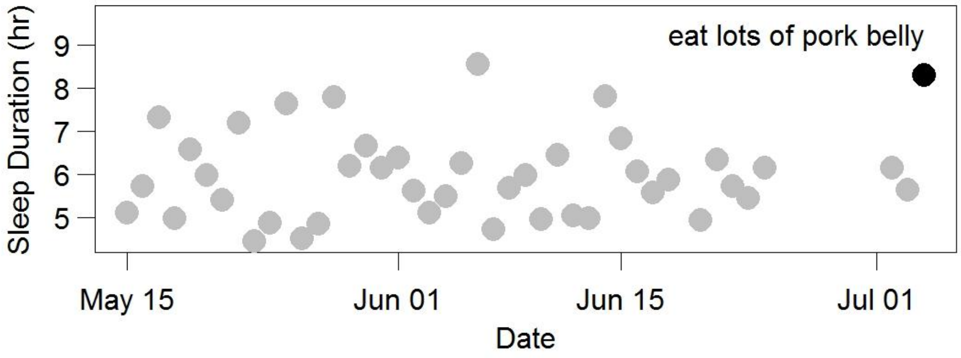

Figure 6. Pork belly outlier. Sleep duration as a function of day. The rightmost point is from a night that followed a day on which I ate a lot of pork belly. I had eaten no pork belly (and little animal fat) on all previous days.

Eventually I did an experiment. I ate pork belly (about 250 g) for lunch some days but not others. It was perhaps two-thirds fat by weight. On baseline days I ate my usual lunch, which had little animal fat. I tried to alternate baseline and pork-belly days, but sometimes failed. When I woke up, I rated my sleep on the 0-100 scale used in the one-leg standing experiment (Figure 4). I did the experiment in two phases. During the first, I kept one-leg standing constant at four times/day; during the second, at two times/day.

Figure 7 shows the results. The lines were fit separately to each set of points. The difference was very clear. I woke up more rested after eating large amounts of pork fat.

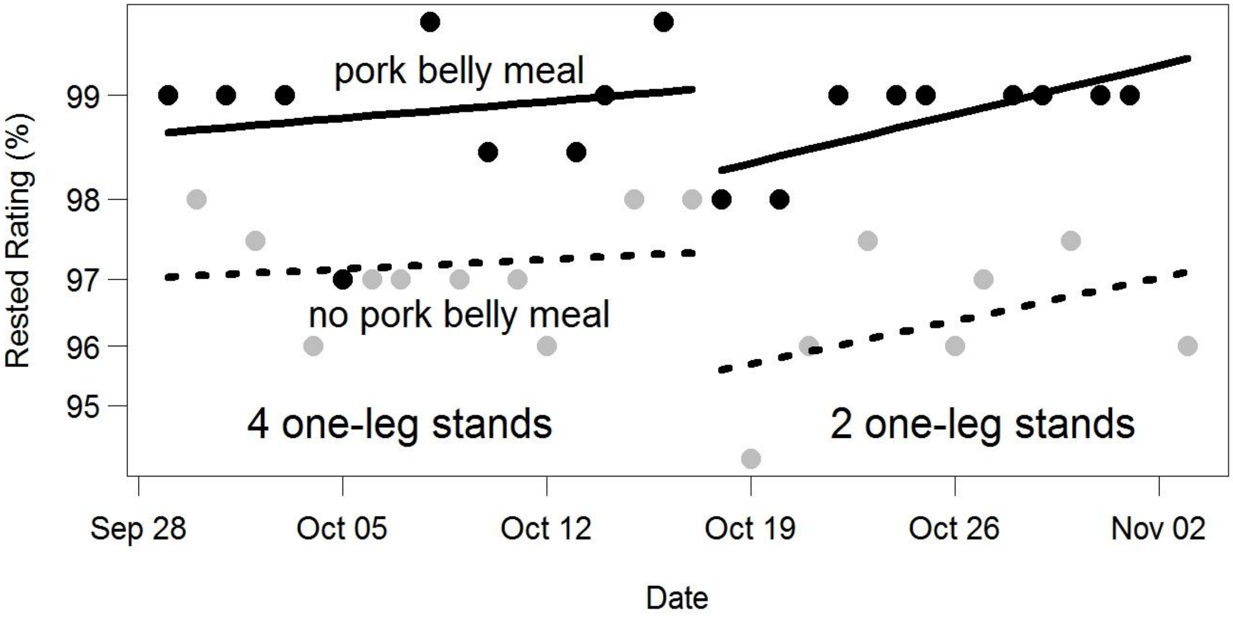

Figure 7. Effect of pork belly meal on rested rating.

Most nutrition experts say that large amounts of animal fat are unhealthy. For example, the 2010 Dietary Guidelines for Americans says "reduce the intake of calories from solid fat" – as if *any* solid fat was too much.

Example 5: Sleep and Vitamin D3

In 2011, I met a California journalist named Tara Grant. Her sleep was terrible, she told me. She woke up 20-30 times every night and in the morning felt like she had not slept at all. I said that sleep is affected by sunlight and that timing matters. Morning exposure to sunlight improves sleep, evening exposure to sunlight-like light makes sleep worse. Because sunlight causes Vitamin D3 synthesis, Grant wondered if the timing of Vitamin D3 mattered. She often took Vitamin D3 in the evening. Maybe D3 in the evening resembled sunlight in the evening.

She started taking her usual dose of Vitamin D3 in the morning. "That night I slept like a rock. And the next night. And the next night," she blogged. "My sleep issues completely resolved" (Grant, 2011). Two months later, the improvement persisted. "My sleep has continued to be solid," she wrote me. "I have not had ONE night of bad sleep since I started paying attention to when I was taking my Vitamin D." When she was young and into her thirties, she never had trouble sleeping. Her sleep problems began around the time that she changed her diet to be more "paleo" (e.g., high fat, low dairy). At the same time, she started taking supplements, one of which was D3.

In the preceding few years, I taken Vitamin D3 for brief periods two or three times, but never in the morning. I had not noticed any effect. After learning about

Grant's discovery, I tried taking it at 8 am every day. I gradually increased the dose from 2000 IU to 8000 IU. When I woke up in the morning I rated how rested I felt on the same scale I used in the one-leg standing experiment (Example 3) and the pork-fat experiment (Example 4).

Figure 8 shows the results. Apparently Vitamin D3 at 8 am made me wake up more rested, and the necessary dose was more than 2000 IU. Grant made her discovery taking 10000 IU/day and found that 5000 IU/day was slightly less effective, which is consistent with my results.

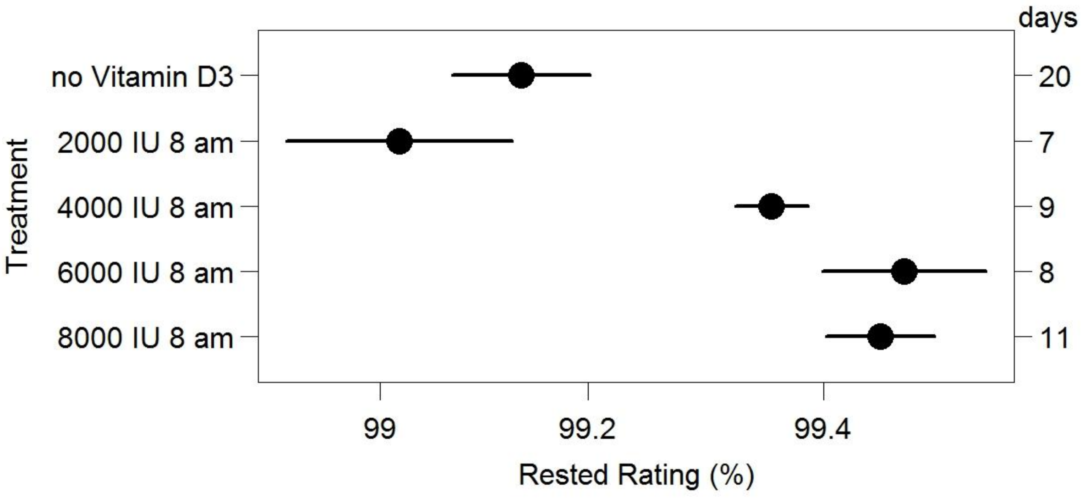

Figure 8. Effect of Vitamin D3 dosage on rested rating. The right-hand column of integers ("days") are the numbers of days each condition was in effect. Averages are means. Error bars show standard errors.

"Everyone suggests taking supplements," wrote Grant, "but I've never heard anyone mention the optimal time [of day] to take them" (Grant, 2011). Nor have I. No Vitamin D3 study with humans has controlled the time of day it is taken, as far as I know. This may be why Vitamin D prevention research with nonskeletal measures (e.g., sleep) has found little evidence of benefit (Maxmen, 2012).

Example 6: Resistance to Infection

In the spring of 1997, I noticed I had not had any colds that winter. I was surprised, but there was an obvious explanation: better sleep. In January 1997, I had started to stand about 10 hours/day (see Example 3 for explanation) and had started to sleep much better. In January 1998, I started to get at least one hour of sunlight exposure every morning, which also improved my sleep (Roberts, 2004).

My freedom from ordinary colds continued. (Except when I travelled, when my sleep was worse.) I had records of when I was had a cold that went back to 1989. In 2002, while writing Roberts (2004), I compared my rate of colds before and after I started to stand many hours almost every day (January 18, 1997). Figure 9 shows my colds, standing and exposure to morning sunlight over the years. (For information about morning TV, see Example 8.) The reduction is clear.

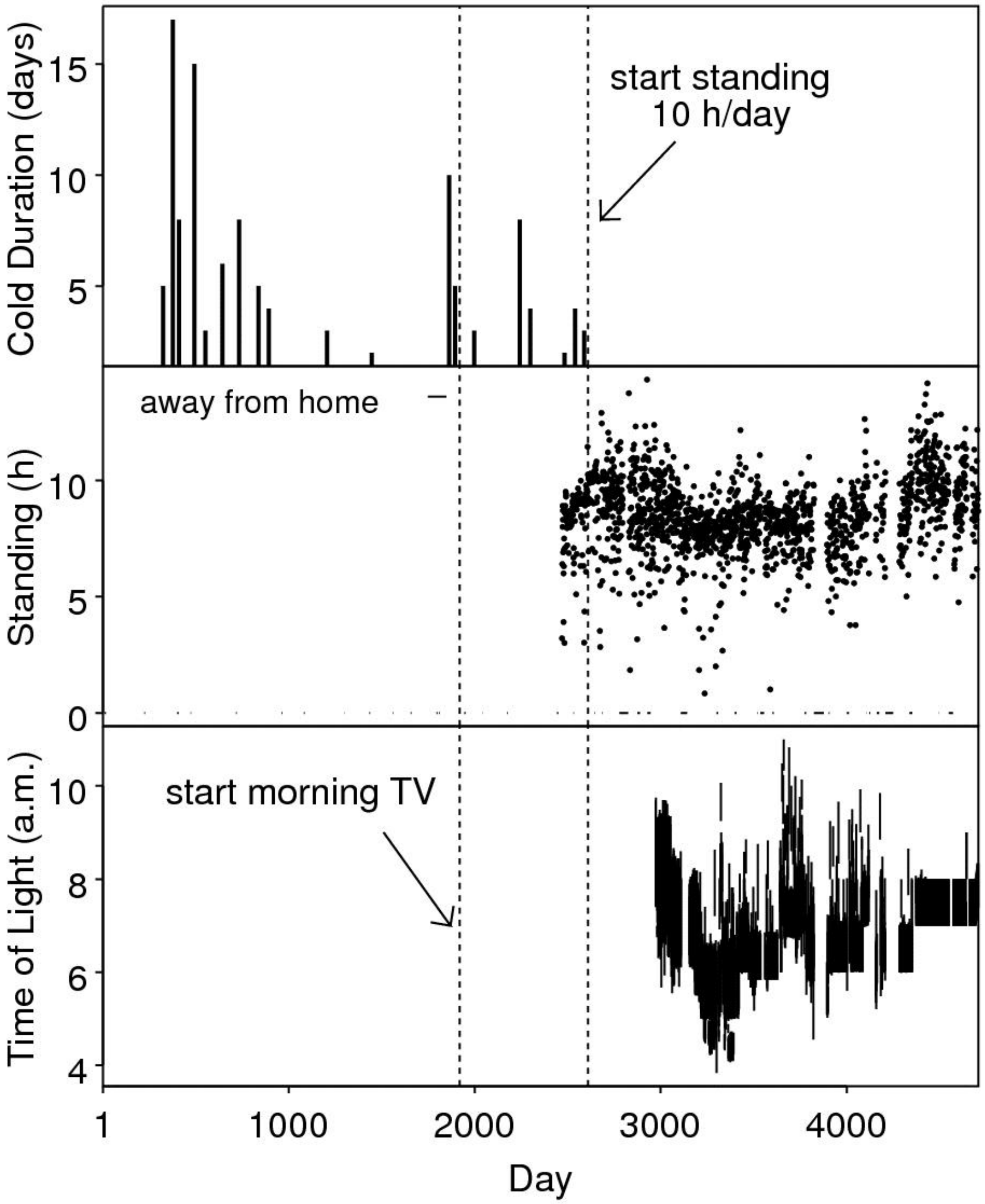

Figure 9. Health, standing, and morning light.

After I stopped getting full-blown colds, I started noticing less obvious signs of infection. Now and then my throat would tingle for a few hours. Now and then I would be more tired than usual and sleep longer than usual for a day or two. These were probably cases where my immune system mobilized fast and well enough to prevent a mild infection from getting worse. These observations support the idea that I was still getting exposed to ordinary amounts of cold viruses and the reduction in colds was because my immune system was doing a better job of fighting them off.

This example makes more plausible the view that if you get full-blown colds, your immune system has room for improvement. Americans average several full-blown colds per year. Current health science pays little attention to how to improve the immune function of people who are not yet sick.

Example 7: Mood and Mood Sharing

For many years, Jon Cousins, a British entrepreneur now in his fifties, suffered from severe mood swings (Morris, 2011). He tried antidepressants and psychotherapy. Neither worked well. In 2006, he realized his mood was getting worse and sought help. A psychiatrist told him he might have cyclothymia, a mild form of bipolar disorder. To confirm the diagnosis, she asked him to track his mood and bring the data to their next meeting. She did not say how to do this.

He decided to measure his mood using his own adaptation of a research tool called the Positive and Negative Affect Schedule. For each of twenty adjectives, such as *alert*, *proud*, and *excited*, he scored himself on a 0-4 scale. The overall score comes from separately summing the positive cards and negative cards and using a table to get an overall rating (0-100, 100 best).

He measured his mood daily (Figure 10). After he'd been recording his mood for three months, a friend asked to see the scores. As soon as he started sharing his scores (by email), they improved. He continued to share his scores with his friend and added other friends to the mailing.

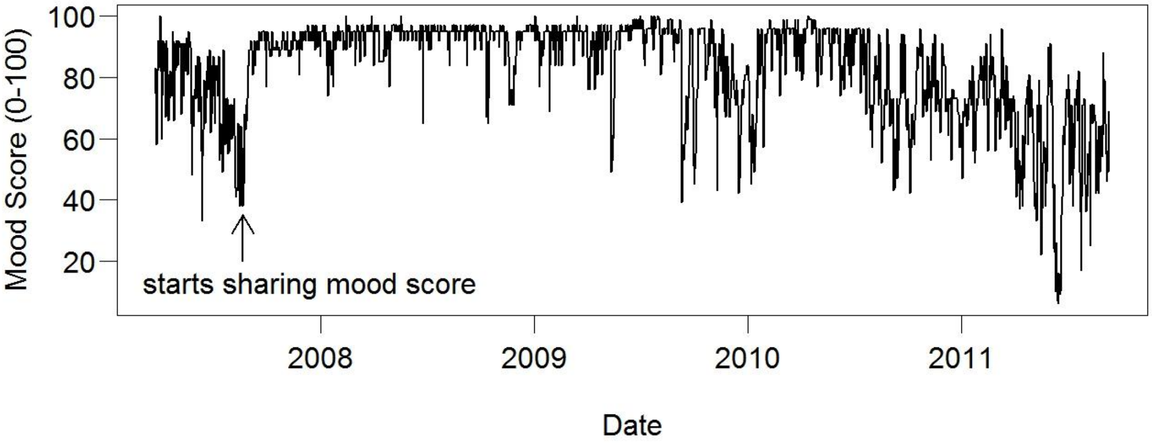

Figure 10. Jon Cousins's mood over five years. He measured his mood daily.

It is very likely that sharing caused the sudden and sustained improvement. Nothing else important happened at the same time. During the five years before he began measuring his moods, Cousins estimates that he suffered from low moods (= score below 40) about 15 percent of the time. Tracking his mood surely made the sharing effect easier to discover.

Unlike other treatments for low mood (psychiatric drugs, psychotherapy, street drugs, alcohol, and so on) mood sharing can be provided to many people at almost no cost and is perfectly safe. In 2009, Cousins founded a website called Moodscope that makes it easy for users to rate and share their mood. It is a big success. By 2012, it was attracting more than 6000 daily users. A survey done in 2011 by the United Kingdom Department of Health to rate new health ideas ranked it #1 (best) of about 600 submitted ideas.

In 2012, I asked Cousins about the generality of the sharing effect he had discovered. He surveyed Moodscope users. Of the 125 people who responded, 50% (62) said that they had tried mood sharing. (For the rest, it was unclear if they had.) Of these, 64% (40) said the effect of sharing was positive, 23% (14) said there was no clear effect, and 13% (8) said the effect was negative. Apparently many Moodscope users find it helpful to measure their mood even without sharing it.

Example 8: Mood and Morning Faces

Skipping breakfast reduced my early awakening (Example 2) but even after that discovery I still sometimes woke up too early. Stone Age people probably did not eat breakfast, I thought. Maybe breakfast was harmful because it was not part of the ancient environment that shaped our genes. Maybe my residual early awakening was due to another difference between my life and Stone Age life.

Because of my research (about internal clocks), I knew that the timing of sleep is controlled by social contact. We tend to be awake at the times of day that we have contact with others. For example, social contact Monday night will make you more awake Tuesday night. In the Stone Age, I imagined, people chatted with their neighbors in the morning. In contrast, I lived alone and might spend the morning alone. Maybe absence of morning social contact made my sleep worse.

I wondered how to test this idea. A time-use survey (Szalai, 1972) suggested that TV has the same effect on sleep as human contact. Maybe I could test my idea by watching TV early in the morning. So one Monday morning in 1995, I watched about 20 minutes of taped Jay Leno monologues soon after I woke up. It had no clear effect. It seemed like ordinary TV. The rest of the day was normal. The next morning (Tuesday), when I woke up, I was stunned how good I felt: cheerful, calm, yet energetic.

The correlation was very strong. It was the first time I had watched TV early in the morning (Monday) and the first time I had felt so good early in the morning (Tuesday). Did it reflect cause and effect? I had heard many ideas about how to be happy but never anything like this. On the other hand, many studies linked depression and poor sleep, so it made some sense that something done to improve sleep had improved mood.

I did small experiments to see if the effect could be repeated. It could. It *was* cause and effect. The crucial event was faces looking at me, I found. Faces in profile and TV without faces had no effect.

To measure the effect better, I made rating scales. Morning faces caused three distinct changes: I became more cheerful, more serene (less irritable), and more eager to do things. So I constructed three scales. Each went from 0 to 100 with 50 = neutral and 100 = best. One scale was happy/unhappy. On this scale, 50 = neither happy nor unhappy, 60 = slightly happy, 70 = somewhat happy, 75 = happy, 80 = quite happy, 90 = very happy, 40 = slightly unhappy, 30 = somewhat unhappy, and so on. Another scale was serene/irritable, with 60 = slightly serene, 70 = somewhat serene, and so on. The third scale was eager/reluctant, with 60 = slightly eager, and so on. To get a single measure, I averaged the three scores, which almost always changed in the same direction by similar amounts.

I varied time of day of face exposure, duration of face exposure, face size, and TV distance. The more closely what I saw resembled what I would see during an ordinary conversation, the stronger the mood-raising effect.

Figure 11 shows the results of an experiment in which I measured my mood several times per day. Morning faces had no effect for about twelve hours. After that, they lowered my mood (in the evening) and raised my mood (the next day).

The results of Figure 11 suggest that I have a circadian oscillator that controls my mood. The oscillator needs exposure to faces to oscillate, just as a swing needs to be pushed. Why should such an oscillator exist? The need for face-to-face exposure synchronizes the moods of people living together. It pushes everyone to sleep at the same time (irritability at night is beneficial –if someone wakes you up you will snap at them) and cheerful and eager to do things at the same time, thus promoting cooperation.

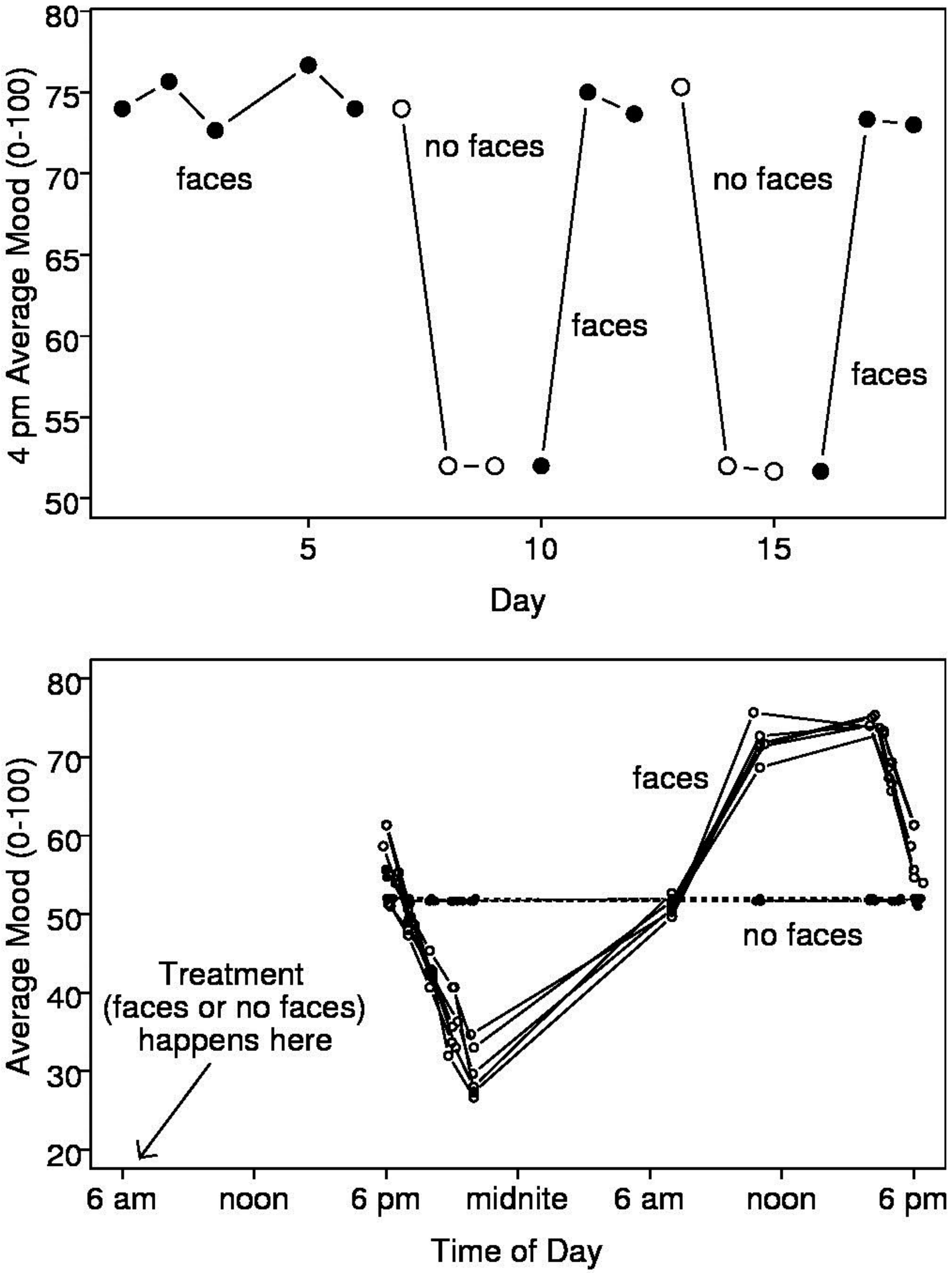

Figure 11. Effect of morning faces on mood. Mood ratings over 17 days in 1999.

This example and the next (Example 9, the Shangri-La Diet) involved plenty of subject-matter knowledge (e.g., the link between depression and bad sleep), which I knew because of my job. They suggest the power of *combining* professional and personal science. In contrast to the rest of the examples, which could conceivably have been done (or were done) by non-professional scientists, these two examples strike me as requiring professional levels of subject-matter knowledge. They suggest that the spread of science will spread to non-professionals – the point made by the rest of the examples -- because they support the ideas that (a) there is a lot we don't know (in these two cases, about such important subjects as mood and weight control) and (b) personal science will be needed to find it.

Example 9: The Shangri-La Diet

In college I weighed about 165 pounds (I am 5' 10"). In 1990 I reached 200 pounds and wanted to lose weight. At the time, I taught introductory psychology. One of my lectures was about weight control. To lose weight, I tried a method suggested by that lecture, which was to eat food closer to its original form (e.g., oranges instead of orange juice). Over the next two months, I easily lost 13 pounds. I was surprised that I had managed to learn something new and useful on such a well-studied topic.

I became more interested in weight control. In 1995, a researcher named Israel Ramirez, at the Monell Chemical Senses Center in Philadelphia, sent me copies of many of his papers. One of them (Ramirez, 1990) led me to think of a new theory of weight control. The theory suggested two new ways to lose weight: (a) eat foods with a weak smell (e.g., sushi) and (b) eat foods with a low glycemic index (e.g., beans). I tried both. Both worked. This increased my belief in my theory.

During a trip to Paris in 2000, I lost my appetite for many days. I must have lost weight. My theory suggested a surprising cause: the unfamiliar sugar-sweetened soft drinks I had been drinking because of a heat wave. This was counter-intuitive. Almost all weight-control experts said (and still say) that sugar causes weight *gain*.

When I returned home, I tested my theory's explanation. It predicted that unflavored sugar water would work better than flavored sugar water (i.e., soft drinks) so I drank unflavored sugar water. I lost my appetite so completely that I lost 30 pounds in about 3 months (Figure 12).

As far as I know, no one had ever lost so much weight so quickly without hunger while eating ordinary food. In a famous experiment done in 1944-5, subjects ate a "semi-starvation" (low-calorie) diet and lost roughly the same amount of weight in a similar amount of time (Keys, Brožek, Henschel, Mickelsen & Taylor, 1950). They suffered greatly from hunger.

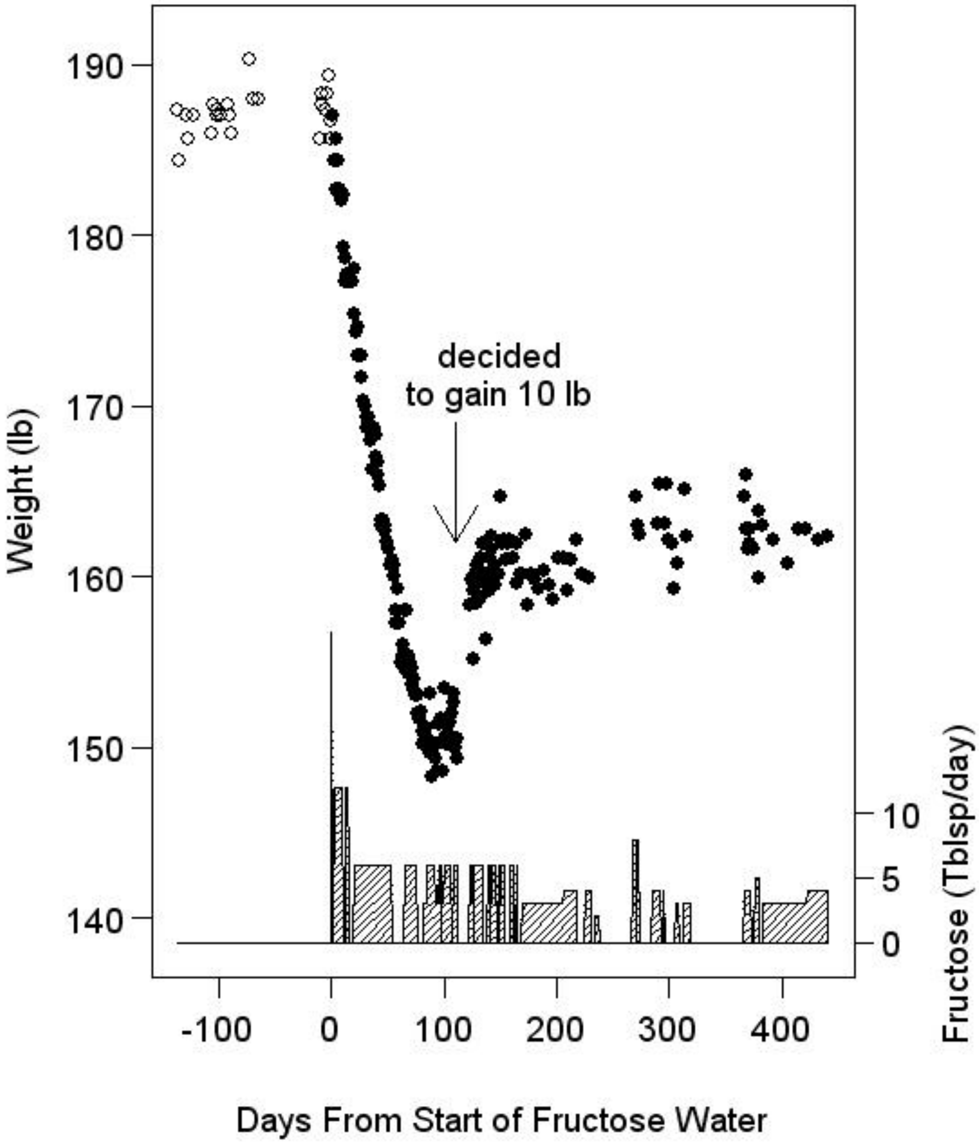

Figure 12. Effect of sugar water on my weight. Each point is the average of three scales. The bars at the bottom show how much fructose I consumed each day (dissolved in water).

Later I used the theory to discover other ways to lose weight. (The best is to eat food while holding your nose shut.) I wrote a book (Roberts, 2006) based on my and other people's experiences.

Drinking unflavored sugar water may not be a solution to the obesity epidemic, but the weight loss shown in Figure 12 suggests there is something seriously wrong with mainstream ideas about weight control, which say sugar is fattening.

Example 10: Weight Loss Methods Compared

In 2001, Alex Chernavsky, a 35-year-old IT manager, wanted to lose weight. He weighed 265 pounds. His Body Mass Index was 38 (40 = morbidly obese). He started to weigh himself daily and record the result.

Over the next 11 years, he tried several ways of losing weight. Figure 13 shows what happened. Here is what he tried:

1. Recording weight. At first he changed nothing besides starting to record his weight. He was eating a lot of sweets and other junk food and not exercising at all. He hoped that paying more attention to his weight would help. He slowly gained weight during this period.

2. Long walks (started 2001). They lasted 1.5-2 hours. Because he lives in upstate New York, he cut down on the walks during winter.

3. Low-carb diet (2002). He lost 50 pounds but then started to regain the lost weight.

4. Vegetarian/vegan (2003). For moral reasons, he became vegetarian and later vegan. This seemed to slow down his weight gain but did not stop it.

5. Long walks resumed (2005). They stopped in the winter.

6. Long walks resumed (2006). They stopped in the winter and did not resume.

7. Shangri-La Diet (2009). He started with extra-light olive oil and sugar water but switched to drinking 4 tablespoons of flaxseed oil per day.

8. Shangri-La Diet modified (2011) To the original Shangri-La Diet regimen he added a heaping tablespoon of coconut oil.

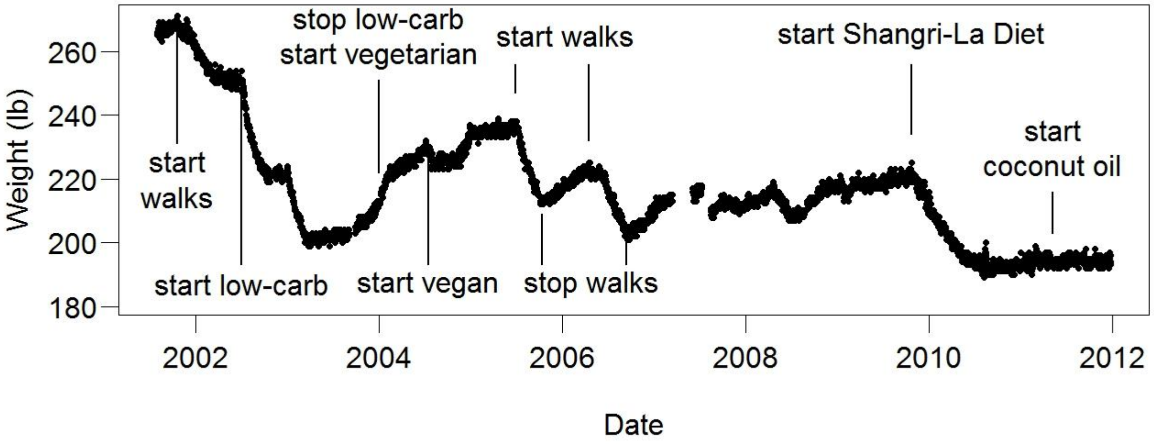

Figure 13. Alex Chernavsky's weight over ten years.

This is personal science as evaluation/customization. Chernavsky did not invent a new way of losing weight. He compared known methods. He wanted to know which worked best. Even this modest goal led to ideas new to me. First, I have seen many times that low-carb diets cause weight loss. I have never before seen if the weight loss is sustained. In this case, it wasn't. Second, weight regain after walking stopped was remarkably slow. Third, his data suggest it takes at least two years to evaluate a weight-loss method. It took that long to show that the weight loss caused by a low-carb diet was not sustained. Almost all published weight-loss studies last less than two years.

Example 11: Brain Function and Flaxseed Oil

The Shangri-La Diet (Roberts, 2006) advocates drinking smell-less oil to lose weight. After it appeared, a few readers wondered if they could do the diet using flaxseed oil capsules. Flaxseed is high in omega-3, which a variety of evidence suggests is good for the brain.

I wondered about the effect of flaxseed oil. One evening in 2006 I swallowed six 1000 mg flaxseed oil capsules. Every morning for two years I had put on my shoes standing up. I stood on one foot while tying the laces of the shoe on the other foot. Even after two years, it was hard. On the morning after I ate the flaxseed oil capsules, it was much easier than usual. Did flaxseed oil improve my balance?

I devised a homespun measure of balance: how long I could stand on one foot on a cutting board balanced on a pipe cap. I chose a diameter of pipe cap that made the task neither too easy nor too hard. I used a stopwatch to measure how long I managed to balance on one foot. A test session was 30 trials and took about 15 minutes. I averaged over the 30 durations to get a single number for each session.

I measured my balance at increasing doses of flaxseed oil. A shift from 2 tablespoons/day to 3 tablespoons/day was followed by clear improvement (Figure 14). The difference is so clear and the experiment so fast that I can easily find the minimum dose/day of flaxseed oil that maximizes my balance. To get sufficient omega-3, the *2010 Dietary Guidelines* from the United States Department of Agriculture recommends eating 8 ounces or more of seafood per week. The results of Figure 14 are especially interesting because I was already eating at least that much seafood per week.

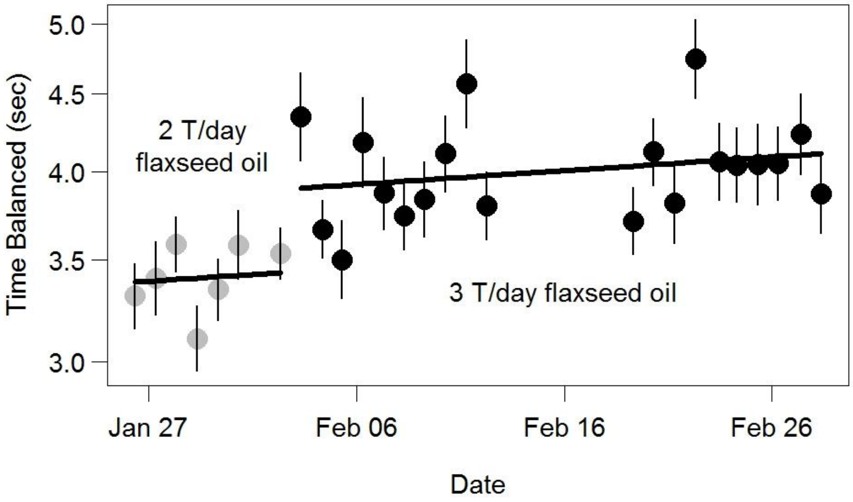

Figure 14. Effect of flaxseed oil dosage on balance.

The results of Figure 14 made me wonder: *What else am I missing?* What other nutrients was I deficient in? I had had no reason to think I was low in omega-3. I was measuring balance only because of my shoelace-tying experience. Maybe I could do better than the balance test, which required special equipment, did not travel well (in a new place scores were much lower), and took 15 minutes.

To find a better test of brain function, I measured the effect of flaxseed oil on four tests:

1. *Balance.*

2. *Digit span* (computer) A common measure of short-term memory. How many random digits can I remember? [how long?]

3. *Memory scanning* (paper and pen). A test invented in the 1960s by Saul Sternberg, a professor of psychology at the University of Pennsylvania. On each trial, I memorized three digits and went through a block of 100 digits, marking each one to indicate whether or not it was in the memorized list. Each test had four trials and took about 4 minutes.

4. *Arithmetic* (computer). I saw a series of simple arithmetic problems, such as 3 + 4, on my laptop screen. I typed the answer and hit Enter as fast as possible. Each test had 100 problems and took a few minutes.

The tests were roughly equal in discomfort (= unpleasantness, ease of testing). I chose the number of trials so that there were as many as possible yet the test still seemed easy. I wanted to find a sustainable test – a test I could do daily for years.

I used these four tests in two experiments. Both had an ABA design. In both, I drank 4 tablespoons/day of flaxseed oil (Treatment A) most of the time. For several days, I switched to something else (Treatment B) – in the first experiment, 4 tablespoons/day of olive oil, in the second experiment, nothing (= I stopped drinking the flaxseed oil) – then switched back to 4 tablespoons/day of flaxseed oil (Treatment A). In the first experiment, I stopped the olive oil after 9 days because the decrement in balance was so large and clear.

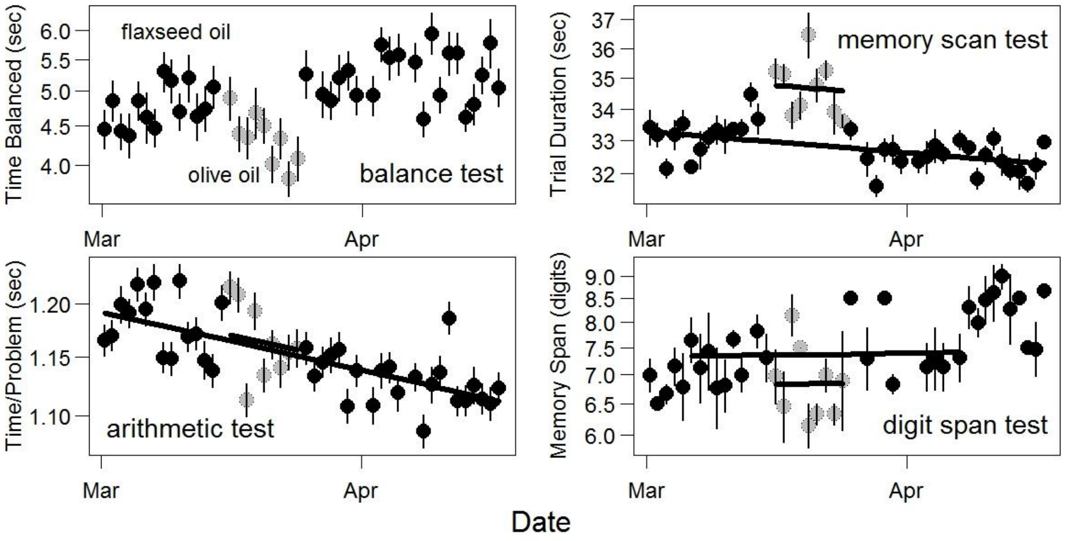

Figure 15. Comparison of flaxseed oil (4 T/day) and olive oil (4 T/day). Each panel shows results from a different test. Each point is a mean. Error bars show standard errors. The lines were fit to the points constraining the two lines to be parallel.

Figure 15 shows the comparison with olive oil, Figure 16 the comparison with nothing. Table 1 shows the *t* values for the comparisons. (Except for the olive-oil balance results, I fit parallel lines to the two sets of data. The *t* value is from a test that the two lines have the same intercept.) The results, especially the memory scanning results, support the idea that flaxseed oil improved my brain function more broadly than just balance.

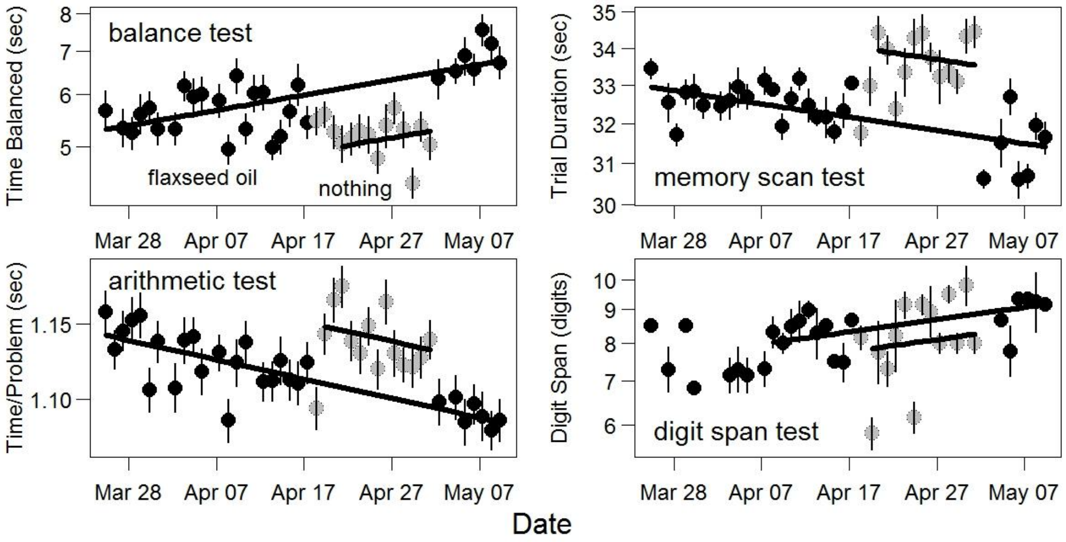

Figure 16. Comparison of flaxseed oil (4 T/day) and nothing. Each panel shows results from a different test. Each point is a mean. Error bars show standard errors. The lines were fit to the points constraining the two lines to be parallel.

	Experiment	
Test	**vs. Olive Oil**	vs. Nothing
Balance	large	7
Memory Scan	8	8
Arithmetic	1	7
Digit Span	2	2

Table 1. Summary of *t* values.

Table 1 suggests that tests equated for discomfort can vary greatly in sensitivity. Based on Table 1, I decided to use the arithmetic test for daily testing. It was not as sensitive as the balance or memory scanning tests but it was far more convenient.

A few years later I returned to the question of how much flaxseed oil to take. At the time I was taking 3 tablespoons/day. I reduced the dose to 2 tablespoons/day. In contrast to Figure 13, I found no change in an arithmetic test – maybe because I had started taking butter in the meantime (Example 12). Then I reduced the dose to 1 tablespoon/day. This *did* make a difference: I became slower (Figure 17). When I went back to 2 tablespoons/day, I returned to my original speed. This is more evidence that flaxseed oil makes a difference and more evidence of the sensitivity of short easy tests.

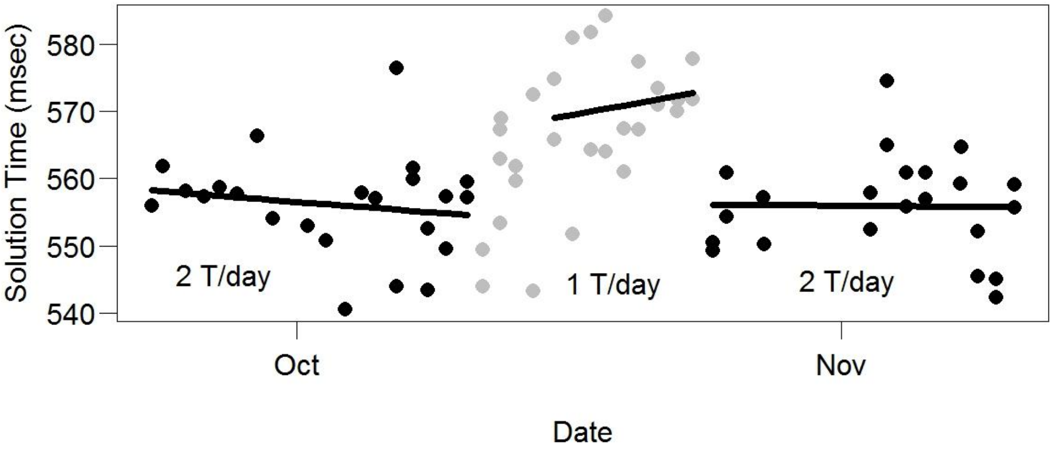

Figure 17. Effect of flaxseed oil dosage on arithmetic speed. 1 T = 1 tablespoon = 15 ml.

Flaxseed oil is high in what is considered an inferior form of omega-3 – short-chain omega-3 (alpha-linolenic acid, ALA). The brain uses long-chain omega-3 fatty acids (eicosapentaenoic acid, EPA, and docosahexaenoic acid, DHA), and almost all experiments about the effects of omega-3, especially those involving the brain, are done with fish oil rather than flaxseed oil. The body can convert ALA to EPA and DHA, but the conversion rate is believed to be low.

Example 12: Brain Function and Butter

After learning that pork fat improved my sleep (Example 4), I tried to eat pork fat every day. One day in 2010 I couldn't get any. As a substitute, I ate a lot of butter (maybe 40 g) at lunch. An hour or so later, I felt a pleasant warmth in my head (the opposite of a headache). I had never before noticed such a feeling and I had never eaten so much butter at once. The large amount of butter had been the only unusual feature of lunch. I concluded that butter influenced brain function.

Because of this, I started to eat large amounts of butter (about 60 g, or 4 tablespoons, half of an American stick of butter) daily. Soon after that, I noticed a sudden improvement in my arithmetic scores (top panel of Figure 17). The next day (Tuesday) I repeated the unusual features of the previous day (Monday). My score on Tuesday was close to my score on Monday (bottom panel of Figure 17). Monday had had about four unusual features. I tested each of them separately. The crucial feature seemed to be the butter.

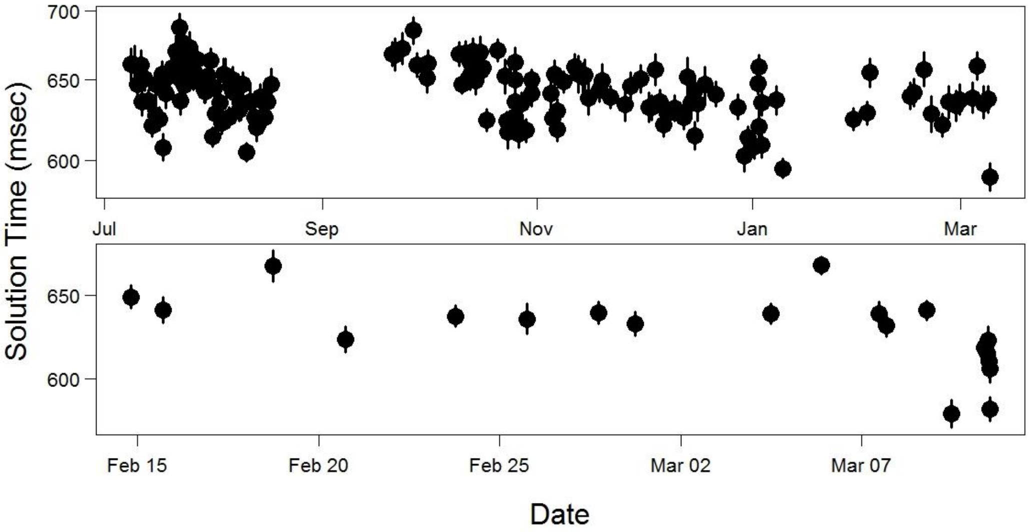

Figure 18. Arithmetic speed over eight months. Top panel: The initial outlier that suggested butter might have an effect. Bottom panel: Repetition of the outlier.

Since then I have eaten butter regularly, about 60 g/day. Figure 19 shows my arithmetic scores over most of this period. The data support the idea that butter produced an improvement.

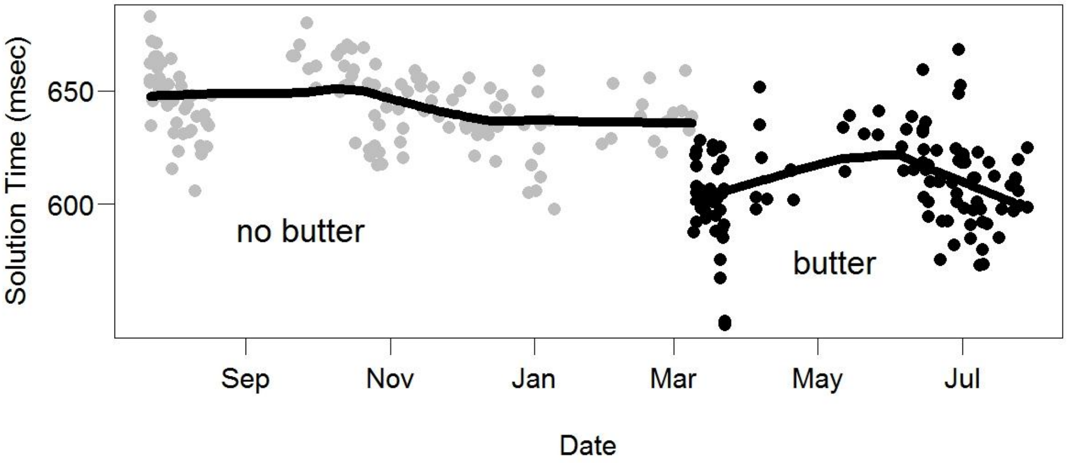

Figure 19. Long-term effect of butter on arithmetic speed. During the no-butter phase, I ate almost no butter (< 1 g/day). During the butter phase, I ate about 60 g/day of butter.

I described these results at a Quantified Self Meetup (meeting). One audience member was Greg Biggers, the founder of Genomera, a Mountain View, California company devoted to "crowd-sourcing health discovery by helping anyone create group health studies" (quotation from their website). He decided to run a test of my results. A Genomera employee named Eri Gentry recruited 45 participants, who were randomly assigned to three groups: butter, coconut oil, and no change. The study lasted three weeks. On each day of the study the subjects did an online arithmetic test that resembled mine. The three weeks were divided into three one-week phases. During Week 1 (first baseline), the participants ate normally. During Week 2 (treatment), the Butter participants added 4 tablespoons of butter (56 g) each day to their usual diet, the Coconut-Oil participants added 4 tablespoons of coconut oil each day to their usual diet, and the No Change participants continued to eat normally. During Week 3 (second baseline), all participants ate normally.

After the experiment ended, Ms. Gentry reduced the data set to participants who had done at least 10 days of testing (her decision). Then she made the data available. At this point I got involved. I wanted to compute difference scores (Week 2 minus the average of Weeks 1 and 3) so I eliminated anyone who had no Week 3 data. I also eliminated four days where the treatment was wrong (e.g., in the sequence N N N N N B B N N B, where N = No Change and B = Butter, I eliminated the final Butter day). That left 27 participants and 443 days of data. Because the scores on individual problems were close to symmetric on a log scale, I worked with log solution times. I computed a mean for each day for each participant and then a mean for each phase for each participant.

Figure 20 shows difference scores (Week 2 minus the average of Weeks 1 and 3). There are clear differences by group. Persons in the Butter group did arithmetic faster when eating butter; persons in the other two groups did not clearly speed up. A Wilcoxon test comparing the Butter and No Change groups gives one-tailed p = 0.006. The results support the idea that butter improves brain function and suggest that coconut oil does not, contrary to what some people think.

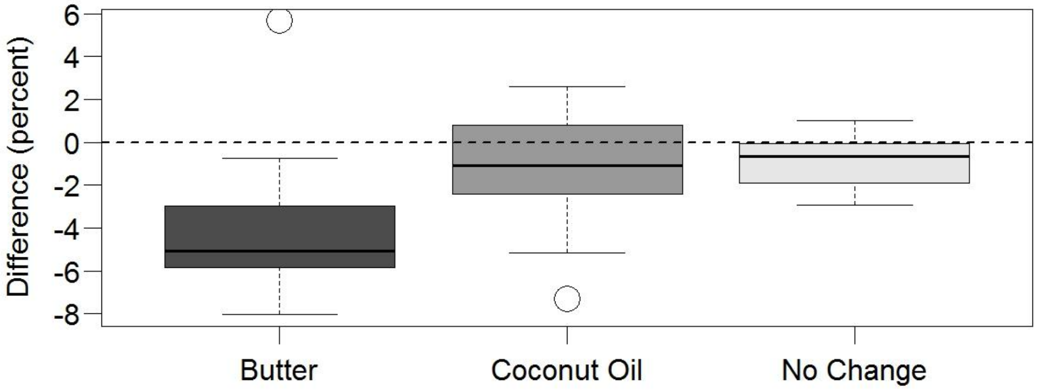

Figure 20. Boxplots of difference scores for three groups: butter, coconut oil, and no change. Difference score = solution time on the treatment week minus the mean solution time on the preceding and following weeks. Negative difference score = faster.

In 2011, I did an experiment to learn about the best dose. Was 60 g/day – the amount I had been taking for a long time -- too much? For 8 days I reduced my dose to 30 g/day. Then I returned to 60 g/day. I measured my arithmetic speed once/day. Figure 21 shows the results. At 30 g/day, I was clearly slower than at 60 g/day (*t*[29.4] = 6.1). The two 60 g/day phases were not reliably different (*t*[19.2]=1.7). The results of Figure 21 suggest that the amount of butter has its entire effect within 24 hours. This is consistent with Figure 18 (outliers), which suggests that butter has a large effect quickly.

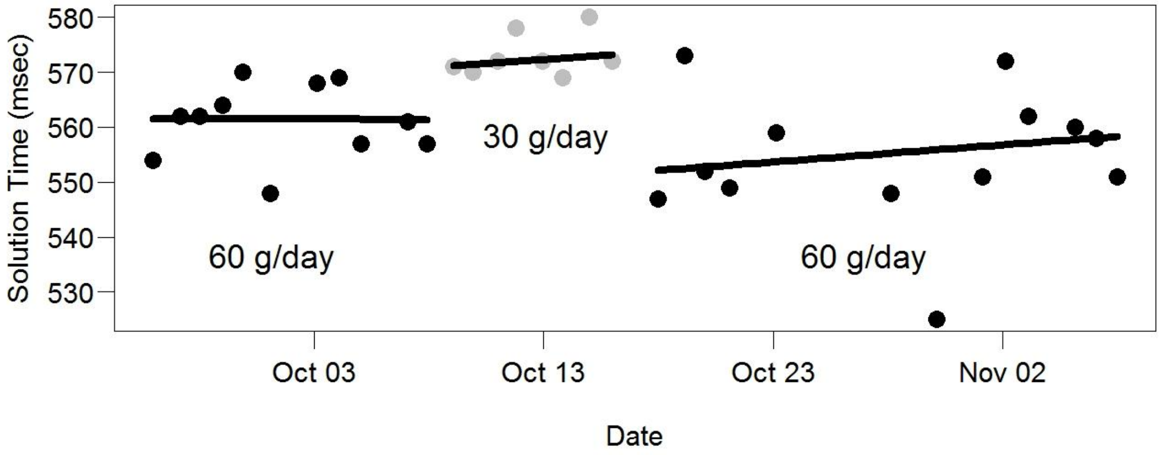

Figure 21. Effect of butter dosage on arithmetic speed.

A 36-year-old New York lawyer who wishes to be called GP was impressed by my butter results (Figure 19) and wondered if butter would improve his brain function. In 2012 he started eating the same amount of butter as me – about 60 g/day (4 tablespoons). To make sure this was not having a bad effect on his cholesterol, he started to measure his cholesterol more often.

He had been measuring his cholesterol with lab tests. To make more frequent measurements, he bought a refurbished CardioChek PA meter. (New ones cost about $700.) The machine is meant for doctors' offices, but it is easy to use. He started measuring his cholesterol about once/week.

He started with Kerrygold butter. After he switched to another butter (from a farmers' market) the results got worse. Then he switched to an Icelandic butter then back to Kerrygold. Then he stopped the butter for three weeks; then resumed the Kerrygold butter for five weeks at double the original dose (120 g/day instead of 60 g/day).

The results (Figure 22) showed that his high-density-lipoprotein (HDL, "good") cholesterol concentration did not decrease; apparently it increased. The CardioChek HDL measurements have a ceiling of 100 so the latest measurements (which equal 100) are probably underestimates. His non-high-density lipoprotein (non-HDL) cholesterol did not clearly change.

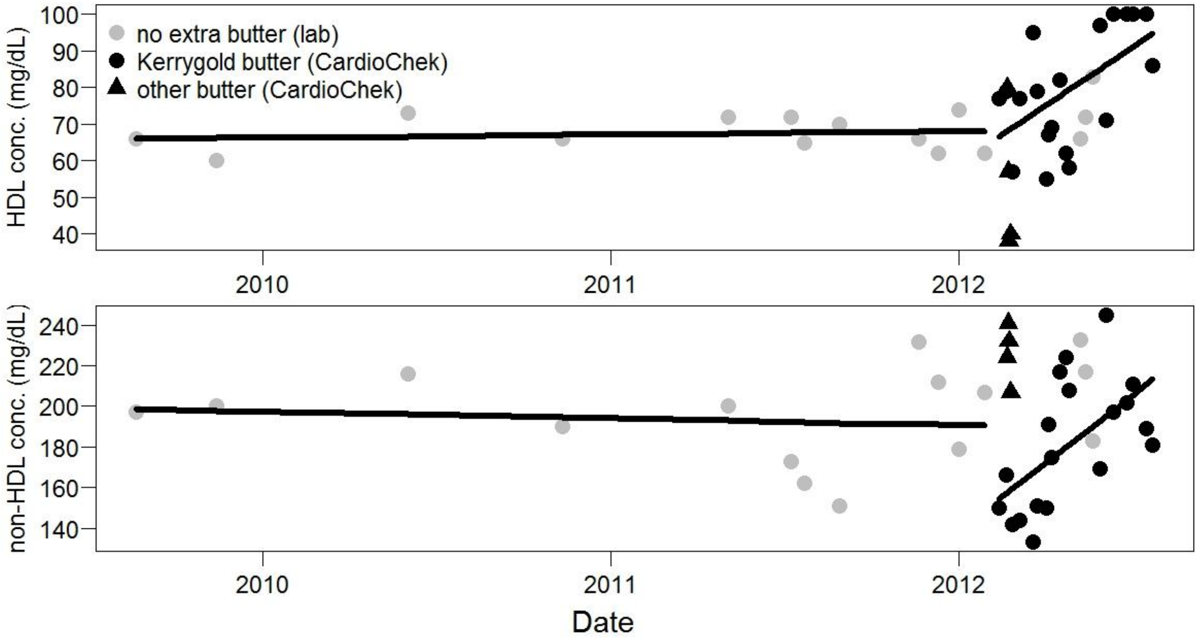

Figure 22. Effect of butter on cholesterol. Upper panel: high-density lipid concentration. Lower panel: non-high-density lipid concentration. The straight lines were fit only to the no extra butter and Kerrygold butter results.

The idea that butter can improve HDL levels is new, at least to me. It contradicts the usual claim that butter is bad for heart disease.

Example 13: Brain Function and Dental Amalgam Fillings

Since the 1800s, dentists have filled cavities with an amalgam that is half mercury. My childhood cavities were filled this way. Small amounts of mercury can damage health and this practice has been criticized. In 1990, *60 Minutes* questioned it. In 2004, a brochure from the Dental Board of California ("The Facts about Fillings") said "scientific evidence and research literature in peer-reviewed scientific journals suggest that otherwise healthy women, children, and diabetics are not at an increased risk [of bad health] from [mercury] amalgams in their mouths." The most recent review I can find (Roberts and Charlton, 2009) says "conclusive evidence is lacking that directly correlates [mercury] amalgam with adverse health effects." However, alternatives are available. By 2009, Norway, Denmark and Sweden had banned mercury-amalgam fillings (perhaps because of disposal problems).

In 2004, my dentist at the time gave me a gold crown next to a tooth with a mercury amalgam filling. This was a mistake. When different metals touch, they generate current (galvanism) that may cause mercury to be released. Hair tests from 2008 to 2009, done by a company named Doctor's Data, suggested I had more-than-usual amounts of mercury in my hair. The percentiles were 73, 75, 70, and 70, with larger values meaning more mercury. (The reference population was not described.) I decided to have my amalgam fillings replaced with non-metallic fillings. On July 28, 2010 I had two amalgam fillings replaced.

A month later, I moved from Berkeley to Beijing. About two months after the move, I noticed that my arithmetic scores had improved and I wondered what caused the improvement. My first guesses involved differences between Berkeley and Beijing: 1. *Walnuts*. Maybe I ate more walnuts in Beijing. Walnuts are supposed to improve brain function. 2. *Heat*. It was much hotter in Beijing than Berkeley. Maybe warmth improves brain function. 3. *Vitamins*. I took fewer vitamin supplements in Beijing. Maybe they harmed brain function. 4. *Other differences between Berkeley and Beijing*. I tested each of these ideas: 1. I stopped eating walnuts. My arithmetic score did not clearly change. 2. Winter came, it got much colder. The improvement persisted. 3. I took the same vitamins I'd taken in Berkeley. My arithmetic score didn't change. 4. I returned to Berkeley. The improvement persisted. So none of these explanations made correct predictions.

After I had gathered evidence against these explanations, I realized there was another one: removal of my amalgam fillings. I checked my records to find out when they were removed. I saw that the arithmetic improvement had started close to the date of the removal. Figure 23 shows the data.

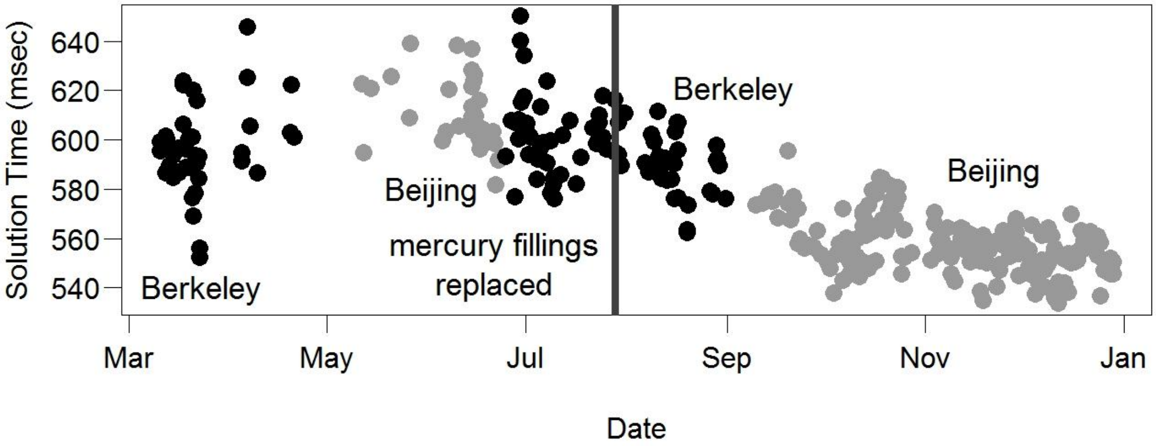

Figure 23. Arithmetic speed over ten months. The vertical bar shows the date that mercury-amalgam fillings were replaced with fillings that did not contain mercury.

The evidence that removal of mercury amalgam fillings improved my arithmetic score includes three things: 1. Four other explanations of the improvement made incorrect predictions. 2. The improvement, which took months to reach maximum, started within a few days of the removal. Improvements that take months to reach maximum (not due to practice) are rare. Over five years of these tests, this is the only one I've noticed. 3. Mercury is known to harm neural function ("mad as a hatter").

Eventually I had the rest of my mercury amalgam fillings removed. My arithmetic scores did not noticeably change.

My mercury amalgam fillings were from childhood. The gold crown, however, was recent. Had I been tracking my brain function before it was put in, I might have been able to detect that it had a bad effect.

Discussion

The examples suggest that personal science can improve on expert advice. "Science is the belief in the ignorance of experts," said Richard Feynman in a talk to high-school science teachers (Feynman, 1969). The examples support this in the sense that they show that a type of science (personal science) found important things that experts didn't know. Maybe Feynman meant that science has been motivated by discoveries that revealed expert ignorance. That was true here. I began personal science because it reduced my acne more than my dermatologist's advice alone. I continued personal science because, with reasonable amounts of time and effort, it continued to improve on expert advice (it found new ways to sleep better, be in a better mood, and so on). I have included many (13) examples because a few examples might be luck.

The examples show the usual benefits of DIYization: 1. *Spread*. The benefits of science spread. Science includes subject-matter knowledge, knowledge about experimental design, data analysis, and so on. I put this knowledge to new use. 2. *Innovation*. All of the examples involved innovation in the sense of non-obvious practical benefit. In my acne experience (described in the Introduction) and Example 10 (weight-loss methods compared, the innovation was choosing among existing treatments (customization). In Examples 1 (blood sugar and walking), 6 (resistance to infection) and 13 (brain function and amalgam fillings), the innovations were close to existing ideas. In the rest of the examples, the innovations were unlikely new methods of improvement (e.g., a new way to sleep better).

The examples suggest that personal science can help improve a wide range of health problems (blood sugar, sleep, weight, etc.). Because most of them are my work, they raise the question of whether others can do this sort of thing. No doubt my job (psychology professor) made personal science much easier for me than most people. On the other hand, what I did can be divided into four parts: measurement, choice of treatment, experimental design, and data analysis. Three of them (measurement, experimental design, and data analysis) can surely be done by many people. With measurement, you use a scale or tool that someone else has made. This is how Jon

Cousins measured mood (Example 7), for instance. My experimental designs were very simple. Data analysis varied from simple (plotting measurements versus day) to slightly complex (the standard errors in Figure 18, butter outliers, were calculated adjusting for difficulty) but in most of the examples plotting simple averages of the data was sufficient to learn something useful. When I began personal science, I had no doubt that I could do these three elements (measurement, experimental design, and data analysis) sufficiently well. I also believe these three elements can be supplied or taught to many people.

The hard part, I believed and still believe, is choice of treatment. The treatments you decide to test need to have a reasonable chance of having an effect. When I studied my sleep, for example, it was easy to measure my sleep, do experiments (test ways of improving my sleep), and analyze the data. *Much* harder was choosing treatments to test that had a reasonable chance of helping. Nothing I had read about scientific method said how to do this. If I did 100 experiments and none showed an effect, I might have stopped. I could not copy what has worked for others because I did not know *any* treatments that reduced early awakening. I was pushed into the unknown – I needed to find new effective treatments – in a way that professional scientists never are. (*Requiring* a professional scientist to find a new treatment for this or that would be seen as asking too much. Such discoveries are too unpredictable.) When I began I had no confidence that I would succeed sufficiently often. I did succeed often enough, so I continued.

All of the examples involved some sort of beneficial change. They can be divided into two groups: 1. *Copy* (Examples 5 and 10). The beneficial treatment was copied. I copied use of Vitamin D3 in the morning from Tara Grant (Example 5). Alex Chernavsky copied the weight loss methods he tried from other people (Example 10). 2. *Accident* (the other 11 examples). The beneficial treatment was found by accident, usually facilitated by long-term self-tracking (long-term records of blood sugar, sleep, etc.). In Example 1, for instance, I discovered the benefit of walking by accident. These 11 examples can be divided into two groups: (a) *Theory involved* (Examples 8 and 9). In Examples 8 and 9, what I did that led to the accidental discovery was heavily based on theory. (b) *No theory involved* (the other 9 examples). The accidental discovery of Example 1 (walking lowers blood sugar), for instance, did not involve any theory. I doubt many people can use theory to make accidental discoveries (Category 2a). Even psychology professors rarely do this. But I believe many people can find useful treatments by copying (Category 1) or theory-free accident (Category 2b).

The growing power of personal science, illustrated by the examples, derives from more than new technology (new hardware, software, and instrumentation). Three other changes also helped: 1. *Treatment*. More treatments are available. In Example 1 (blood sugar and walking), it was easy for me to walk one hour per day partly because I had a home treadmill (cheaper now than in the past). Example 8 (mood and morning faces) required a video recorder. 2. *Science*. Science has improved. More is known. Tools are better (e.g., loess). I couldn't have discovered the Shangri-La Diet (Example 9) without the work of other scientists (especially Michel Cabanac, Israel Ramirez, and Anthony Sclafani) done in the 1970s and 1980s. 3. *Spread of knowledge*. It is easier to get access to what's known. Before the Internet were cheap photocopies. Example 5 (sleep and Vitamin D3) was inspired by one person's unusual experience, which I learned about from her blog. 4. *Audience*. It is easier to do something if you can tell others about it. Professional scientists can give talks about, teach and publish their work. Until recently, there was no audience for personal science. (Richard Bernstein had great difficulty publicizing his work.) Now personal science can reach an audience via blogs and Quantified Self events.

When I say personal science is becoming more powerful, I mean more powerful relative to professional science. The first three changes (treatments, science, spread of knowledge) have also made professional science – and expert advice based on professional science – more powerful. Why would they disproportionally benefit personal science?

To answer that, I propose that the rate of scientific progress (progress per person per unit time) depends on the product of five factors:

progress = knowledge/skill * resources * 1/cost * freedom * motivation.

Looking at the product of these factors is plausible because you need all of them. If any are zero, progress is zero. As I said in the Introduction, professional scientists have the advantage in the first two factors (skill/knowledge and resources), whereas personal scientists have the advantage in the last three (1/cost, freedom and motivation). But that is just the base state of affairs – true for a long time.

Examples 1-13 suggest that personal science is overtaking professional science in certain ways in that (a) personal science found important effects that professional science had missed and (b) most of the examples would have been much more difficult or impossible 30 years ago. To explain the overtaking, consider how the five factors (knowledge/skill, etc.) have changed in the last thirty years. In the two factors where personal science has been behind, it is catching up.

1. *Knowledge/skill*. The professional/personal difference has gotten smaller because personal scientists have much more access to accumulated scientific knowledge.

2. *Resources*. Personal scientists have much more access to high-quality measurement devices, a wide range of treatments, statistical software, and so on. All of the examples show that useful research can be done very cheaply and the necessary materials are

widely available. Because personal scientists can test many treatments that professional scientists cannot, more availability of treatments may have increased the advantage of personal scientists in this area.

In contrast, I see little change in the last three factors for professional science or personal science. In all three, personal science retains its immense advantage

3. *Cost*. Personal science remains much cheaper per experiment than professional science.

4. *Freedom*. Personal scientists continue to have much more freedom than professional scientists.

5. *Motivation*. Personal scientists remain much more motivated to produce actual benefit than professional scientists.

That is my attempt to explain the most important feature of the examples -- that personal science improved on expert advice. In this sense, personal science "worked" – was a good investment of time and effort. This suggests that personal science will grow in popularity, increasing demand for statistical expertise. Most of the rest of this discussion is about the other way personal science may influence statistics: Statisticians can learn from it.

The Underselling of Statistics

My personal science persuaded me that statisticians, in my contact with them (books, classes, articles), were radically undervaluing their subject. This belief began with my acne research, which I described in the introduction. In a few months it reduced my acne much more than my dermatologist's advice. The statistics involved was means. Afterwards I thought of my statistics teachers (including Tukey, 1977): *Why hadn't they said this was possible?* It wasn't *exactly* statistics that had reduced my acne – I had measured my acne, done an experiment, and so on – but statistics had played a big part.

Each of the examples raises the same question: *Why hadn't they said this was possible?* The examples are about subjects of great interest (e.g., diabetes, sleep, mood, weight loss). Placed against professional research on these topics, the research described here, in terms of cost, effort and time, is a drop in the bucket. Yet it substantially helped. Apparently a little data – with the help of data analysis -- can go a long way. I have never heard a statistician make this point.

In 2012 I asked William Cleveland about this. My point was: your methods (e.g., Cleveland, 1993) are capable of producing great drama, yet your books do not show this. He replied:

> Local high drama. If I come in and show a previous result was wrong or missed information in the data, which is in both of my books, I can assure you there is a lot of drama and consternation on the receiver's side if they know about it. Go back and have a look, and imagine you are the author of a paper I address. For example, you are the author of the paper where I show the behavior is manifestly nonlinear, and that your fit of the linear is just completely wrong. I guarantee you'd feel drama. One hopes the reader feels "There but for the grace of ... ".
>
> Global high drama. There is NASA missing the ozone hole for a decade because an automated data-cleaning program kicked out the "bad" observations. Had they made some simple plots they would have seen the yearly cycle in the hole and they would have known instantly it was real. The British eventually discovered it. This is stuff of the newspapers. I suppose I had this in mind earlier. [This example is not in Cleveland (1993) but he includes it in talks.]

The NASA example makes the point I am saying is not made. In his talks, Cleveland is speaking about cutting-edge analysis and graphics. My acne research taught me that *means* can be a powerful and radical tool.

How Statistics Helped

How can statisticians help personal scientists? You might think all I did was plot data. Here's what else helped:

1. *R*. About seven years ago, I started using R, which turned out to be much better than earlier statistical software I'd used (Statgraphics, Splus). Compared to them and other alternatives, its virtues for my work are better graphics, command-line interface, few bugs, zero cost (sharing is much easier), loess implementation, and ability to create the reaction time tests of Examples 11-13 (currently restricted to Windows – the program that measures response latency is only available for Windows).

2. *Flexibility of approach*. I learned there is no one "right" way to analyze a dataset.

3. *Loess*. Flexible summary of time series was a constant need, which loess fulfilled well.

4. *Adjustment*. In the arithmetic tests (Examples 12 and 13), a large fraction of the variation between tests was eliminated by adjusting for problem difficulty.

5. *Transformation*. Transformation of the data had the usual virtues. For example, it reduced the influence of extreme values and made graphs more informative.

6. *Statistical tests*. In my professional science, *p* values and statistical tests matter greatly, as they do for most scientists. In my personal science, explicitly

computed *p* values and statistical tests were almost never important. (Table 1 is an exception – *t* values helped compare the sensitivity of different tests.) Perhaps this is because (a) the effects I studied were stronger than those I study in my professional science and/or (b) data is much cheaper, so that I can simply collect enough data until the answer is clear. However, statistical tests and explicit *p* values improve communication. I can show data where it is apparent to me that *p* is very small (e.g., 0.0001) so I don't bother to compute it. Yet someone may ask "is that effect reliable?" Then it helps to compute *p*.

These are just examples. They come, of course, from a much larger culture/methodology. Statisticians have created a language, methods, and sets of simplifications (called assumptions) that help transform data into conclusions. The methods can be as simple as computing means. In my first personal science, about acne, computing means made it much clearer that the antibiotic my doctor had prescribed did not work. I believe that professional scientists use the language, methods and simplifications because they work, where *work* means allow their careers to prosper. For example, (a) it is not too easy nor too hard to meet the criteria for publication, (b) published results can be repeated by other scientists often enough and (c) are useful to non-scientists often enough that scientific research continues to be funded. Examples 1-13 show that the methods also work for personal science, where *work* means improve my health (most of the examples) or other people's health (Examples 7, 10, and 12).

How to Find Ideas Worth Testing

A professional scientist might say that idea testing is the essence of science ("it is the very essence of science to propose hypotheses and to devise tests to see if they can be falsified," wrote Heaney, 2008, p. 1592) but that view overlooks the necessity of idea *finding* – the need to find ideas plausible enough to be worth testing. Without a source of new ideas worth testing, scientists end up testing existing ideas over and over, with diminishing returns. That the term *hypothesis testing* is common and you have never seen the term *hypothesis finding* (I invented it) may reflect the influence of professional science (much more interested in the former than the latter) on statistics. Examples 1-13 contain many instances of idea finding. Maybe they can teach us about it.

Statisticians know little about idea finding, as far as I can tell. I've seen a few dozen introductory statistics textbooks. All of them spend many pages on how to test ideas and say little or nothing – usually nothing – about how to find ideas plausible enough to test. An exception is Tukey (1977), which advocated graphing your data in various ways. The graphs might reveal unexpected structure, which could generate a new idea. The new idea would be plausible because it was suggested by data. Scientists also know almost nothing about idea finding. The state of the art is illustrated by Root-Bernstein (1989), which drew a wide range of lessons about idea finding from instances of it. The various lessons seemed unrelated and I found the evidence for them unconvincing. If a scientist understood how to find ideas worth testing – knew something new and useful -- presumably he or she would use that to find many ideas worth testing. None of Root-Bernstein's examples included this result.

For most of my scientific career, I was equally ignorant. My professional and personal science have run in parallel. Both started in graduate school. In both, I took Tukey's (1977) advice and plotted my data as much as possible. In my professional science, this seemed to work. The new ideas worth testing I managed to find did come from graphing data. But there were only a few of them. In my personal science, in contrast, graphing data made little difference. It helped detect outliers (e.g., Example 12), but most of the important outliers were detected without it (e.g., Examples 1, 2, 4, and 8). Although my professional science generated far more data (tens of millions of numbers) than my personal science (thousands of numbers), it found far fewer ideas worth testing. Graphing data is a good way to find ideas, yes, but why did my personal science find many more ideas worth testing than my professional science? Not because of graphing data.

For a long time I had no idea why my personal science found many more ideas worth testing than my professional science. Now I think I can explain it. The essence of the explanation is: *To find ideas worth testing, lower the cost per test.* (This is half of the explanation. For the other half, see the next section, Power-Law Distribution of Progress.) My personal science provided very cheap tests. My professional science did not.

What makes a new idea (X causes Y) worth testing? Sufficient plausibility. Tests are costly. Suppose X costs $10 and Y is worth $1000. Something like this (we have X, we want Y) is the usual state of affairs. Learning that X causes Y (input $10, output $1000) is much more valuable than learning that X does not cause Y (input $10, output $0). If you think it very unlikely that X causes Y, a test of whether X causes Y may be a poor investment. For example, suppose you believe the chance that X causes Y is 0.0001 (very implausible). To learn if X causes Y costs $100,000. At that price, you are unlikely to find out.

For professional scientists, tests cost more than money. You can only do a few per year. If too many fail, your career is in danger. The effect of these costs is that ideas must be quite plausible to be worth testing. If an idea is *too* plausible, it cannot be published ("we already knew that"). The usual solution, to reach high but not too high plausibility, is to take an idea with very high plausibility (already confirmed by you or someone else) and alter one element (X or Y), thus lowering its plausibility.

The effect of such cost-benefit analyses ("is this worth testing?") by professional scientists is that many ideas with low but positive plausibility have not been tested. Some of these ideas may be true. My personal science managed to search among the ideas that professional scientists cannot afford to test. That's how I found a relatively high number of ideas worth testing (with expensive tests).

My personal science found several new ideas worth testing because *it provided much cheaper tests*. It did so in two ways:

1. *It provided essentially free tests*. The free tests came from two features of the situation. (a) *Long baseline measurements*. I measured myself regularly, even when not doing an experiment. For example, I recorded my sleep every day. This was so easy it was essentially free. It provided free tests by taking advantage of variation in my life. Suppose I started eating a new food. My sleep records tested the idea that the new food affected my sleep. In this way, I tested many low-plausibility ideas. A few gained enough plausibility from the baseline test to be worth a visible test (e.g., Example 2, sleep and breakfast, and Example 4, sleep and pork fat). (b) *Many dimensions effortlessly tracked*. We constantly track our well-being in many ways without actual records. We have a rough memory of how well we usually sleep, what our skin usually looks like, how we usually feel after a meal, and so on – dozens of dimensions. Making a deliberate change in one's life (eating a new food, going to a new place, and so on) provides a rough test of the idea that what has been changed affects each of these dimensions. When I did an experiment to learn if *X* causes *Y,* and explicitly measured *Y,* I also learned to some extent whether *X* affected dozens of other variables that I was not explicitly measuring. This is how I got the first clue that morning faces affect mood the next day (Example 8) and that flaxseed oil improves brain function (Example 11).

2. *It provided very cheap tests.* I can test any idea amenable to self-experimentation much more cheaply than professional scientists can test the same idea.

Plotting your data as much as possible fits this framework because plots provide very cheap tests. Making a plot costs almost nothing and any plot tests some idea. For example, a scatterplot of *X* and *Y* tests the idea that they are connected.

To summarize: *How to find ideas worth (expensive) testing? By cheap testing*. Some fraction of cheap tests will have positive results, making those ideas plausible enough to be worth expensive tests. This way of thinking implies that that every area of science needs a set of tests ranging widely in cost. The cheapest tests can search most widely. The ideas they make more plausible can be assessed by the second cheapest test. Ideas that pass the second cheapest test can be assessed by the third cheapest test, and so on. People who classify science as "good" and "bad", in my experience, fail to understand this. Their "good" science (e.g., randomized clinical trials) is more expensive than their "bad" science (e.g., epidemiology). My personal science did a good job of finding ideas worth testing partly because it provided much cheaper tests than professional science and partly because it provided two layers of testing: essentially free and very cheap.

Power-Law Distribution of Progress/Surprise

The history of science is full of single observations with great consequences, such as Galvani's discovery of galvanism and Fleming's discovery of penicillin. Scientific progress is the opposite of smooth, in the sense that a tiny fraction of all observations produce a large fraction of the progress. My area of expertise, an area of experimental psychology called animal learning, has shown the same pattern on a smaller scale. Progress has depended almost entirely on the discovery of new cause-effect relationships. These discoveries are a very small fraction of all research. On a still smaller scale, my own research, both my professional and personal science, has also shown this pattern. Examples 1 (blood sugar and walking), 4 (sleep and pork fat), 8 (mood and morning faces) and 12 (brain function and butter) derived from single observations, for instance.

What is the distribution of progress per observation? If I looked only at my professional science, I would have a hard time answering the question because the average progress per observation is very close to zero. It is hard to see how such tiny quantities are distributed; it is like asking about the distribution of thicknesses of the hairs on one's head. I might guess the distribution is bimodal: there are discoveries (tiny fraction), which produce a detectable amount of progress, and non-discoveries (everything else), which produce much less, nearly zero progress. *The Structure of Scientific Revolutions* (Kuhn, 1962) argued for a bimodal distribution (paradigm-breaking science versus normal science).

My personal science, however, suggests a power-law (Pareto) distribution of progress. If the term *progress* is vague, think of it as surprise (= amount of change of beliefs). The more surprise, the more change in beliefs, the more progress. In my professional science, even the largest amounts of progress/surprise are not very large (alas), making it hard to see what the whole distribution looks like. In my personal science, however, the largest amounts of progress/surprise have been quite large. The key observations of Example 8 (mood and morning faces) were the most surprising, followed by the key observations of Example 9 (Shangri-La Diet). The key observations of Examples 4 (sleep and pork fat), 12 (brain function and butter), 3 (sleep and standing), and 11 (brain function and flaxseed oil) were also quite surprising, but less so. And so on. There is no clear separation between discoveries and non-discoveries. Consider Example 5 (sleep and Vitamin D3). Because of Tara Grant's earlier observations, I wouldn't call the most surprising observations of that example a discovery. However, they still produced considerable progress/surprise because of the novelty of the idea they support (the time of day you take Vitamin D makes a big difference).

Because the largest amounts of progress/surprise generated by my personal science were relatively large, it became possible to make an informed guess about the shape of the whole distribution. This paper, of course, describes only a tiny fraction of my personal science data, omitting an enormous number of less helpful observations. Looking at all of them, I could see that, as one might expect, there was a continuous gradient from nearly zero surprise to the large amount of surprise in the key observation of Example 8 (mood and morning faces). Conceptually, the scale divided into three regions: (a) results that agreed with what I expected (low surprise), (b) results that contradicted expectations but weren't shocking (medium surprise), and (c) shocking results (high surprise). But there was no trimodality. Everywhere along the scale, more surprising observations were less likely. A very small fraction of observations produced a large fraction of the progress. For these reasons it looked like a power-law distribution.

Once I saw that my personal science data had a power-law-like distribution of progress, I could see something else: The slope could change. Two comparisons suggested this:

1. *Personal science versus professional science*. After I noticed the power-law-like distribution of my personal science, I looked back at my professional science data. Progress per observation was so small that I could only make out that the distribution was consistent with a power-law distribution. Assuming that was the actual distribution, comparison with my personal science data showed that it had a much steeper slope on log-log coordinates. My professional science collected far more data than my personal science but made much less progress overall.

2. *Personal science after 1990 versus personal science before 1990*. Before 1990, my personal science made relatively little progress. Example 2, determining that breakfast made my sleep worse, was the turning point. That was the first set of observations that could be called a discovery. After that, discoveries were more common. Because Example 2 was "accidental", it is plausible that the distribution of progress that generated Example 2 had a power-law distribution. The increase in rate of progress (before versus after Example 2) can be explained by assuming that Example 2 improved the slope of the distribution. This is plausible. Example 2 suggested to me that many health problems may be due to non-Stone-Age aspects of our diet. This idea considerably narrowed my later choices of what treatments to test (Roberts, 2004).

These two comparisons – professional versus personal science, and before and after Example 2 – both suggest that the slope of the power-law function can vary greatly from one situation to another, even when the two situations involve the same scientist.

Here we have six pieces of information. 1. The distribution of progress/surprise produced by my personal science resembles a power-law distribution. 2. The distribution of progress/surprise produced by my professional science is consistent with a power-law distribution. 3. The history of science is consistent with a power-law distribution. 4. The history of the study of animal learning is consistent with a power-law distribution. 5. My personal science has been far more productive (of progress/surprise) than my professional science. 6. After Example 2, my personal science became more productive.

They can be explained with two broad assumptions:

First, *the distribution of progress/surprise (per observation) has a power-law distribution* (Figure 24). Supporting this view of scientific progress, the distribution of citations to a scientific paper resembles a power-law distribution (Gupta et al., 2005; Redner, 2005). Meyer (2007) tells many stories of medical progress via rare "accidental" discoveries, each of which supports something like a power-law distribution.

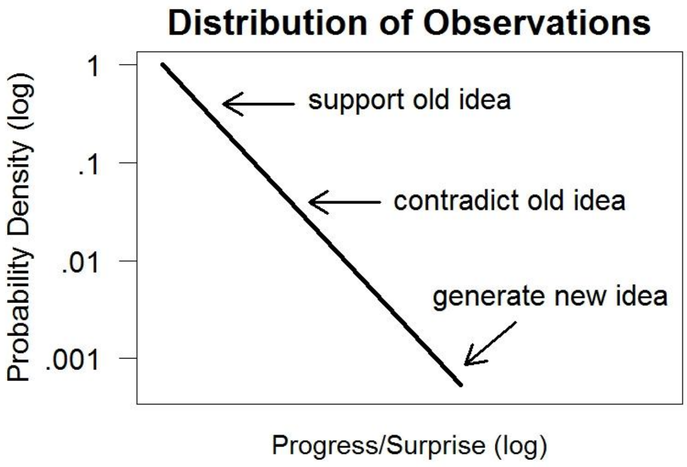

Figure 24. Power-law distribution of surprise/progress per observation.

Second, *the slope (parameter) of the distribution depends on novelty, subject-matter knowledge, and scientific skill* (Figure 25). Let me explain each of these.

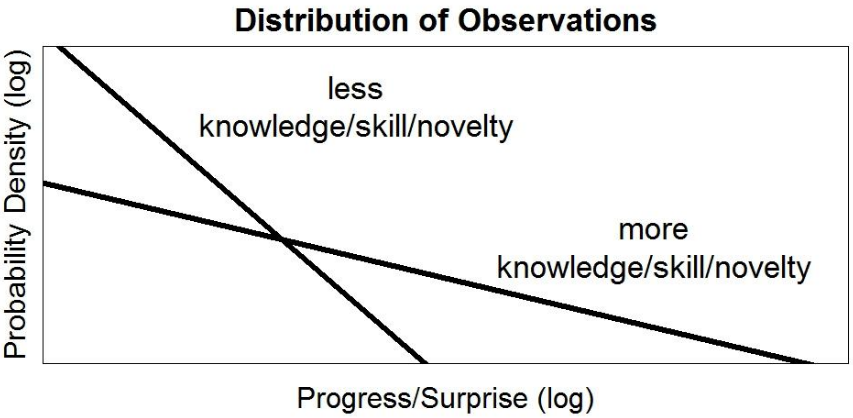

Figure 25. Effects of knowledge, skill and novelty on the slope of the distribution of surprise/progress.

(a) *Novelty* means the novelty of the observations, which depends on the novelty of what's measured, what's varied, and the context. The more novel an observation, the flatter the slope of the surprise/progress distribution. My professional science collected far more data than my professional science but was far less novel. I measured lab rats pressing a lever for food and tested a relatively narrow range of treatments. In contrast, my personal science studied less-studied treatments (e.g., no breakfast, standing, morning faces) and less-studied dimensions (e.g., human sleep) and did so in the context of my ordinary life, a less constrained situation than laboratory life. My life varies more day to day than a lab rat's life.

(b) *Subject-matter knowledge*. The more I knew about what I was studying (e.g., sleep), the flatter the slope. The more I knew, the better I could choose what to vary (the better I could guess what treatments would have an effect) and the better I could understand strange results. Example 8 (mood and morning faces) illustrates both advantages. Because of subject-matter knowledge, I tested an unusual treatment (morning faces on TV). Because I knew about the connection between depression and circadian rhythms, the strange result of looking at morning faces (the next morning I felt happy, eager, and serene) made sense to me. I explain the pre-1990/post-1990 difference by an increase in subject-matter knowledge. In 1990, my surprising breakfast results (Example 2) suggested to me that many common health problems are due to differences between modern life and Stone-Age life (Roberts, 2004). Plenty of data support this idea. Pasteur's dictum "chance favors the prepared mind" is close to what I am saying here.

(b) *Scientific skill*, meaning skill in measurement, experimental design, and data analysis. Better measures will be less noisy and less likely to suffer from floor or ceiling effects. A better experimental design will provide more interpretable data. Better data analysis will find learn more from the data.

This theory of science predicts a bright future for personal science. Personal scientists, who can study anything, have a great novelty advantage over professional scientists, who must do experiments not far from experiments already done because they need a steady stream of publications. Comparison of my personal science to my professional science – a comparison in which subject-matter knowledge and scientific skill are roughly equal in the two things being compared – suggests what a big difference novelty can make. Personal scientists are almost surely inferior to professional scientists in terms of subject-matter knowledge and scientific skill (e.g., a person studying his own sleep knows less about the sleep than a professional sleep researcher), but they can catch up in these areas. If they are studying sleep, for example, they can read countless articles about it. Their scientific skills can improve by studying available examples of personal science, which should become more common.

This view of science may interest statisticians because it is closely connected to Bayesian ideas. By adding constraints (plausible assumptions) , it makes Bayesian ideas more powerful. In addition, it offers a new explanation (or description) of the value of statistics: it flattens the distribution of progress/surprise. I wonder why statistics has this effect. Do some aspects of statistics do so more than others? Can the flattening produced by specific statistical tools be measured?

"Correlation Does Not Equal Causation" is Misleading

Many statisticians have said something like "correlation does not equal causation." (By *correlation* they mean non-experimental associations. By *equal* they mean always imply.) For example, Barnard (1982, p. 387), writing about causation, said "that correlation is not causation is perhaps the first thing that must be said." Holland (1986, p. 945) said "correlation does not imply causation." Utts (2003, p. 74) said that "statistical literacy" included understanding "when it can be concluded that a relationship is one of cause and effect, and when it cannot, including the difference between randomized experiments and observational studies."

This has a grain of truth. Newspaper articles sometimes make a *correlation equals causation* mistake. They may say that a study that found X associated with Y found that "X causes Y" when other interpretations are plausible. Utts (2003) gives an example. However, Barnard (1982), Holland (1986), and Utts (2003) were addressing scientists, not journalists. I have never seen a scientific article confuse empirical result (correlation) with explanation (X causes Y).

How much association of X and Y (correlation) increases the plausibility of *X causes Y* (causation) depends on only one thing: the plausibility of other explanations. If other explanations are absolutely certain, then association of X and Y doesn't increase the plausibility of *X causes Y* at all. In practice, it is rare that you can be sure that an alternative explanation is correct. In practice, then, observation of a correlation between X and Y usually *does* make *X causes Y* more plausible. The statement *correlation does not equal causation* ignores all this. It ignores the crucial thing: the plausibility of alternative explanations.

In my experience, the plausibility of alternative explanations varies enormously. If someone says "This correlation makes *X causes Y* more plausible" it is helpful to respond with information about the plausibility of alternative explanations. To reply "correlation does not equal causation" is like saying “shut up” or (more politely) "I don't want to think about it."

Does this widespread misunderstanding (taking seriously *correlation does not equal causation*) illustrate a larger problem? I suspect it reflects two larger problems. One is human nature. Some people enjoy criticizing others. For them, statements that facilitate criticism (such as *correlation does not equal causation*) are less carefully scrutinized than other statements. The larger problem this creates is unbalanced evaluation: too much *you cannot conclude X from this data*, too little *what can we learn from this data*. The lack of balance – not restricted to statistics – is reflect in the fact that the term *critical thinking* is common but you’ve never heard the term *appreciative thinking*. The other larger problem may be distance from the practice of science, especially science about low-plausibility ideas. In my personal science, correlations – what a statistician might call "weak evidence" -- were precious. They were hard to find. (It took me ten years to find something that correlated with early awakening. Example 2 describes what I finally found.) They led to tests of possible explanations. Almost all the examples (e.g., Examples 1 and 2) illustrate the enormous value of correlations in the search for causation and how unhelpful it would have been to have greeted the correlations I managed to find with *correlation does not equal causation*. When you are searching among a thousand ideas, all with very low plausibility, to find weak evidence for one of them is great progress.

Three Types of Scientific Method: Experiment, Survey, and Wait and See

Many of the examples – in particular Examples 1 (blood sugar and walking), 3 (sleep and standing), 4 (sleep and pork fat), 6 (resistance to infection), 7 (mood and mood sharing), 12 (brain function and butter), and 13 (brain function and dental amalgam) – illustrate a scientific method that I have never seen mentioned in the statistical literature or discussions of scientific method. I call it *wait and see*. You repeatedly measure something and hope that there will eventually be a puzzling outlier. If you observe an outlier, you try to learn what caused it. In Example 1, for instance, the outlier was an unusually low blood sugar reading. I guessed that it was caused by walking more than usual. Subsequent events supported this.

Successful examples of the wait and see method illustrate the broad point that different methods are good at different things and we need all of them. In contrast, evidence-based medicine advocates that say some types of evidence are “better” than other types — implying one-dimensional evaluation. (For example, they say that randomized experiments are “better” than surveys.) In assessing the value of this or that method, at least three dimensions matter. One thing scientists want to do is *test cause-effect ideas* (does X cause Y?). The wait and see method doesn’t do that at all. Experiments do that well, surveys are better than nothing. Another thing scientists want to do *is assess the generality of cause-effect ideas*. The wait and see method doesn’t do that at all. Surveys do that well (it is much easier to survey a wide range of people than do an experiment with a wide range of people), multi-person experiments are better than nothing. A third thing scientists want to do is *come up with cause-effect ideas worth testing*. Most experiments are a poor way to do this, surveys are better than nothing. The wait and see method is especially good for that.

Department of Scientific Method

Statistics blends mathematics and psychology. Statisticians understand how mathematics helps. They have been slower to understand how psychology helps. Growth of interest in graphics is an example. Maybe interest in graphics derived from experience (graphs often helped) but behind the efficacy of graphics was a psychological fact: Humans are visual.

Behind the efficacy of personal science, illustrated by the examples, lies another psychological fact: Professional scientists have much less freedom than personal scientists. Personal scientists have just one goal: better health. In contrast, professional scientists are not single-mindedly trying to maximize (other people’s) health. Health improvement is one goal among several: status, job security, the need to publish enough to get tenure, and so on. The other goals push professional scientists to be cautious, which reduces how much progress they make. The more statisticians understand the psychology of scientists, the more they will understand how science works and how the psychology of scientists has shaped statistics. (The caution of professional scientists has led to neglect of idea finding, a central part of science.)

The more statisticians understand how science works, the more easily departments of statistics can become departments of scientific method. If statisticians decide to help personal scientists analyze their data, they can also help them design measurements and experiments – they need help with those, too.

Statisticians may see personal scientists as low-status customers – at least, lower status than professional scientists. If so, they may want to reflect on the lessons of Christensen (1997), which describes how, in one industry after another, failure to make tools for low-status customers doomed industry-leading companies. The disk-drive industry is one of Christensen's examples. When all disk drives were large, the main customers were universities and large companies – high-status customers. Then personal computers arrived. They needed much smaller disk drives. The industry-leading companies – the ones that sold the most large disk drives -- did not enter this market at least partly because they didn't want to deal with low-status hobbyists. It was a fatal mistake. The future belonged to small disk drives.

A Healthy Scientific Ecosystem

An area of science is an ecosystem in the sense that research builds on other research. In an ordinary ecosystem the various organisms help each other. Animals help plants spread their seeds; plants provide animals food and shelter, and so on.

Scientific ecosystems benefit from diversity just as ordinary ecosystems do. The examples of this paper do not show that personal science is *better* than professional science. That would be like saying one part of an ecosystem is better than another or that the steering wheel of your car is better than the tires. The examples of this paper are the joint product of personal and professional science. For example, Jon Cousins measured his mood (Example 7) using a tool developed by professional scientists. I discovered the mood-raising effect of morning faces (Example 8) because of clues provided by professional science. The examples show that the *combination* of personal and professional science does better than professional science alone.

Personal and professional science differ greatly, just as animals and plants differ greatly, which is why they can be so helpful to each other. Just as animals are much more mobile than plants, personal scientists are much more flexible in what they study than professional scientists. Just as a plant spends its whole life in one place, a typical professional scientist spends his whole career studying one thing. The mobility of animals allows them to spread seeds much more widely than plants alone can. Likewise, personal scientists can take ideas of professional scientists and apply them much more widely than professional scientists can by themselves. Animals and plants get energy from entirely different sources. Plants get energy from light, animals get energy from plants and animals. Likewise, personal and professional scientists have different goals. Personal scientists want to improve their own health. Professional scientists want job security, status, and so on. Improving their own health is not a goal. Animals react much faster to environmental change than plants. Likewise, personal scientists can react much faster to new findings than professional scientists.

Personal science varies along a continuum of innovation, from barely innovative to highly innovative. At the barely-innovative end, personal science tests known cause-effect relationships and customizes existing treatments. Testing the two drugs my dermatologist prescribed for my acne is an example; so is comparison of different ways of losing weight (Example 10). At the highly-innovative end, personal science finds new cause-effect relationships. Most of the examples illustrate this. These two ends vary in their relationship to professional science. To barely-innovative personal science, professional science provides ideas to test. In return, barely-innovative personal science provides better testing of these ideas (because personal scientists are biased neither for nor against them and test them under a much wider range of conditions). To highly-innovative personal science, professional science provides tools, knowledge and skills. In return, highly-innovative personal science provides new ideas for professional science to test. Figure 26 illustrates this.

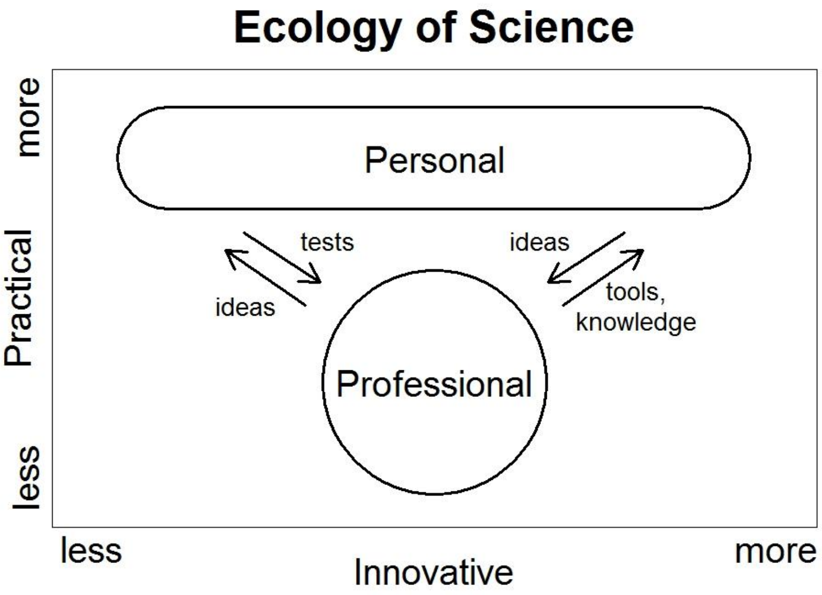

Figure 26. Relationships between personal science and professional science.

Science and Literacy

Spread of ability to do science may resemble the spread of literacy (ability to read). Reading is a kind of primitive science in the sense that it is information gathering. Like a scientific instrument (e.g., microscope), literacy increases how much you can observe. It allows you to learn more, with fewer intermediaries, about almost everything. Ability to do science – so new and rare there is no word for it – allows you to learn how things work, including how your body works, with fewer intermediaries. We now take the value of literacy for granted. There may come a day when it is widely accepted that everyone should be able to do science.

Literacy alone is worthless. There has to be (a) something worth reading that (b) you can obtain. Then literacy has value. Likewise, the ability to do science is worthless by itself. There needs to be (a) something worth learning (via science) that (b) you can learn. Sure, personal computers and other technology make science easier. But they do not imply (a) and (b). That is what the examples try to show – that (a) there are useful things about health that the experts don't know and (b) they can be learned with reasonable amounts of knowledge and effort.

References

BARNARD, G. A. (1982). Causation. In S. Kotz, N. Johnson, C. Read (Eds.), Encyclopedia of Statistical Sciences 1 387–389.

BERNSTEIN, R. (2003). *Dr. Bernstein's Diabetes Solution: The Complete Guide to Achieving Normal Blood Sugars Revised & Updated*. Little, Brown, New York.

CHRISTENSEN, C. M. (1997). *The Innovator's Dilemma*. Harvard Business School Press, Cambridge, MA.

CLEVELAND, W. S. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.

EADY, E. A., JONES, C. E., TIPPER, J. L., COVE, J. H., CUNLIFFE, W. J., and LAYTON, A. M. (1993). Antibiotic resistant propionibacteria in acne: need for policies to modify antibiotic usage. *BMJ* 306 555-556.

FEYNMAN, R. (1969). What is science? *Physics Teacher* 7 313-320.

GILL, J. M. and COOPER, A. R. (2008). Physical activity and prevention of Type 2 diabetes mellitus. *Sports Med*. 38 807-824.

GRANT, T. (2011). http://www.primalgirl.com/2011/11/01/nprimalgirl-sleep-issues-vitamin-d/

GUPTA, H. M., CAMPANHA, J. R., PESCE, R. A. G. (2005). Power-law distribution for the citation index of scientific publications and scientists. *Braz. J. Phys*. 35 981–986. Available at http://www.sbfisica.org.br/bjp/files/v35_981.pdf.

HEANEY, R. P. (2008). Nutrients, endpoints, and the problem of proof. *J. Nutr*. 138 1591-1595.

HOLLAND, P. W. (1986). Statistics and causal inference. Journal of the American Statistical Association 81 945-960.

HU, G., LAKKA, T. A., KILPELÄINEN, T. O., TUOMILEHTO, J. (2007). Epidemiological studies of exercise in diabetes prevention. *Appl. Physiol. Nutr. Metab.* 32 583-595.

KEYS, A., BROŽEK, J., HENSCHEL, A., MICKELSEN, O., and TAYLOR, H. L. (1950). *The Biology of Human Starvation* (2 volumes). University of Minnesota Press, Minneapolis, MN.

KUHN, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

MAXMEN, A. (2012). Vitamin D on trial. *The Scientist*. Available at http://the-scientist.com/2012/03/01/vitamin-d-on-trial/.

MORRIS, S. (2011). Bipolar illness: My ever changing moods. *The Independent*. Available at http://www.independent.co.uk/life-style/health-and-families/features/bipolar-illness-my-ever-changing-moods-2248854.html

MEYERS, M. A. (2007). *Happy Accidents: Serendipity in Modern Medical Research*. Arcade Publishing, New York.

MISTLBERGER, R. E. (1994). Circadian food-anticipatory activity: formal models and physiological mechanisms. *Neurosci. Biobehav. Rev.* 18 171-195.

NOLAN, D. and TEMPLE LUCE, D. (2010). Computing in the statistics curricula. *Amer. Statist.* 64 97-107.

RAMIREZ, I. (1990). Stimulation of energy intake and growth by saccharin in rats. *J. Nutr.* 120 123-133.

REDNER, S. (2005). Citation statistics from 110 years of *Physical Review*. *Physics Today* 131 49-54. Available at http://physics.bu.edu/~redner/pubs/pdf/PT.pdf.

REID, K. J., BARON, K. G. , LU, B., NAYLOR, E., WOLFE, L., and ZEE, P. C. (2010). Aerobic exercise improves self-reported sleep and quality of life in older adults with insomnia. *Sleep Med.* 11 934-940.

ROBERTS, H. W., & CHARLTON, D. G. (2009). The release of mercury from amalgam restorations and its health effects: A review. *Oper. Dent.* 34 605-614.

ROBERTS, S. (2001). Surprises from self-experimentation: Sleep, mood, and weight. *Chance* 14 7-14.

ROBERTS, S. (2004). Self-experimentation as a source of new ideas: Sleep, mood, health and weight. *Behav. Brain Sci.* 27 227-262.

ROBERTS, S. (2006). *The Shangri-La Diet*. Putnam, New York.

ROBERTS, S. (2010). The unreasonable effectiveness of my self-experimentation. *Medical Hypotheses* 75 482-489. Available at http://dx.doi.org/10.1016/j.mehy.2010.04.030

ROBERTS, S. (2012). The reception of my self-experimentation. *Journal of Business Research*. Available at http://www.sciencedirect.com/science/article/pii/S0148296311000476

ROOT-BERNSTEIN, R. S. (1989). *Discovering: Finding and Solving Problems at the Frontiers of Science*. Harvard University Press, Cambridge, MA.

SINGH, N. A., CLEMENTS, K. M., and FIATARONE, M. A. (1997). A randomized controlled trial of the effect of exercise on sleep. *Sleep* 20 95-101.

SZALAI, A. (1972). The Use of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries. The Hague: Mouton Publishers.

UTTS, J. (2003). What educated citizens should know about statistics and probability. *Amer. Statist.* 57 74-79.

VON HIPPEL, E. (2005). *Democratizing Innovation*. Oxford University Press, New York.

WAINER, H. and VELLEMAN, P. (2008). Looking at blood sugar. *Chance* 21 56-61.

YOUNGSTEDT, S. D., O'CONNOR, P. J., and DISHMAN, R. K. (1997). The effects of acute exercise on sleep: a quantitative synthesis. *Sleep* 1997; 20:203-14.

Figure Captions

Figure 1. Blood sugar over time. Each point is one measurement (one day). The lines are loess fits.

Figure 2. Early awakening over one year. Each point is one morning. The *y* axis indicates the time from waking up to falling back asleep. If I did not fall back asleep within 6 hours, *y* = 6 hours. The line is a loess fit. The ABA experiment began in 1993. From Roberts (2004).

Figure 3. Effect of standing on early awakening. Early awakening = Fell back asleep within 6 hours after getting up. Vertical segments show standard errors assuming a binomial distribution. Upper panel: Between-phase differences. Lower panel: Within-phase differences. Standing durations were divided into three categories: fewer than 8.0

hr; 8.0-8.8 hr; and more than 8.8 hr. The probability of early awakening for each category is plotted above the median of the durations in that category.

Figure 4. Effect of amount of one-leg standing on rested ratings. When I awoke I rated how rested I felt on a 0-100 percentage scale where 100% = completely rested, not tired at all, 99% = 99% of tiredness gone, and so on. Averages are means. Error bars show standard errors. The conditions baseline(2), baseline(3), and baseline(4) happened before and during the randomized experiment. The conditions random(2), random(3), and random (4) are from the randomized experiment. The condition baseline 4 was after the randomized experiment.

Figure 5. Effect of amount of one-leg standing on sleep duration. Averages are means. Error bars show standard errors. The conditions baseline(2), baseline(3), and baseline(4) happened before and during the randomized experiment. The conditions random(2), random(3), and random (4) are from the randomized experiment. The condition baseline 4 was after the randomized experiment.

Figure 6. Pork belly outlier. Sleep duration as a function of day. The rightmost point is from a night that followed a day on which I ate a lot of pork belly. I had eaten no pork belly (and little animal fat) on all previous days.

Figure 7. Effect of pork belly meal on rested rating. During the first phase of the experiment, I did 2 one-leg standings each day; during the second phase, I did 4 each day.

Figure 8. Effect of Vitamin D3 dosage on rested rating. The right-hand column of integers ("days") are the numbers of days each condition was in effect. Averages are means. Error bars show standard errors.

Figure 9. Health, standing, and morning light. Top panel: Time and duration of colds (upper respiratory tract infections). Middle panel: Duration of standing as a function of day. The lines at the bottom of the panel indicate days away from home. Bottom panel: Exposure to artificial morning light as a function of day. Day 1 = November 27, 1989.

Figure 10. Jon Cousins's mood over five years. He measured his mood daily.

Figure 11. Effect of morning faces on mood. Mood ratings over 17 days in 1999. Upper panel: Mood at 4 pm day after day. Lower panel: Time course of the effect. Mood was rate on three scales: unhappy/happy, irritable/serene, and reluctant/eager. Each scale ranged from 5 = extremely negative (e.g., extremely unhappy) to 50 = neutral to 95 = extremely positive (e.g., extremely positive). In both panels, each point is an average of the three ratings (one per scale). Each line in the lower panel is a separate series of measurements. The data in the lower panel start about 12 hours after the treatment because that is when the treatment began to have an effect.

Figure 12. Effect of sugar water on my weight. Each point is the average of three scales. The bars at the bottom show how much fructose I consumed each day (dissolved in water).

Figure 13. Alex Chernavsky's weight over ten years. The treatments are described in the text. He measured his weight daily.

Figure 14. Effect of flaxseed oil dosage on balance. 2 T = 2 tablespoons = 30 ml.

Figure 15. Comparison of flaxseed oil (4 T/day) and olive oil (4 T/day). Each panel shows results from a different test. Each point is a mean. Error bars show standard errors. The lines were fit to the points constraining the two lines to be parallel. Lines were not fit to the balance results because the olive oil results were not at any time parallel to the flaxseed oil results.

Figure 16. Comparison of flaxseed oil (4 T/day) and nothing. Each panel shows results from a different test. Each point is a mean. Error bars show standard errors. The lines were fit to the points constraining the two lines to be parallel.

Figure 17. Effect of flaxseed oil dosage on arithmetic speed. 1 T = 1 tablespoon = 15 ml.

Figure 18. Arithmetic speed over eight months. Top panel: The initial outlier that suggested butter might have an effect. Bottom panel: Repetition of the outlier. Error bars show standard errors computed from the residuals of a regression that adjusted for several factors, such as size of answer.

Figure 19. Long-term effect of butter on arithmetic speed. During the no-butter phase, I ate almost no butter (< 1 g/day). During the butter phase, I ate about 60 g/day of butter.

Figure 20. Boxplots of difference scores for three groups: butter, coconut oil, and no change. Difference score = solution time on the treatment week minus the mean solution time on the preceding and following weeks. Negative difference score = faster.

Figure 21. Effect of butter dosage on arithmetic speed.

Figure 22. Effect of butter on cholesterol. Upper panel: high-density lipid concentration. Lower panel: non-high-density lipid concentration. The straight lines were fit to the no extra butter and Kerrygold butter results (omitting the other butter results). The grey points come from lab tests, the black points from CardioChek measurements.

Figure 23. Arithmetic speed over ten months.

Figure 24. Power-law distribution of surprise/progress per observation.

Figure 25. Effects of knowledge, skill and novelty on the slope of the distribution of surprise/progress.

Figure 26. Relationships between personal science and professional science.