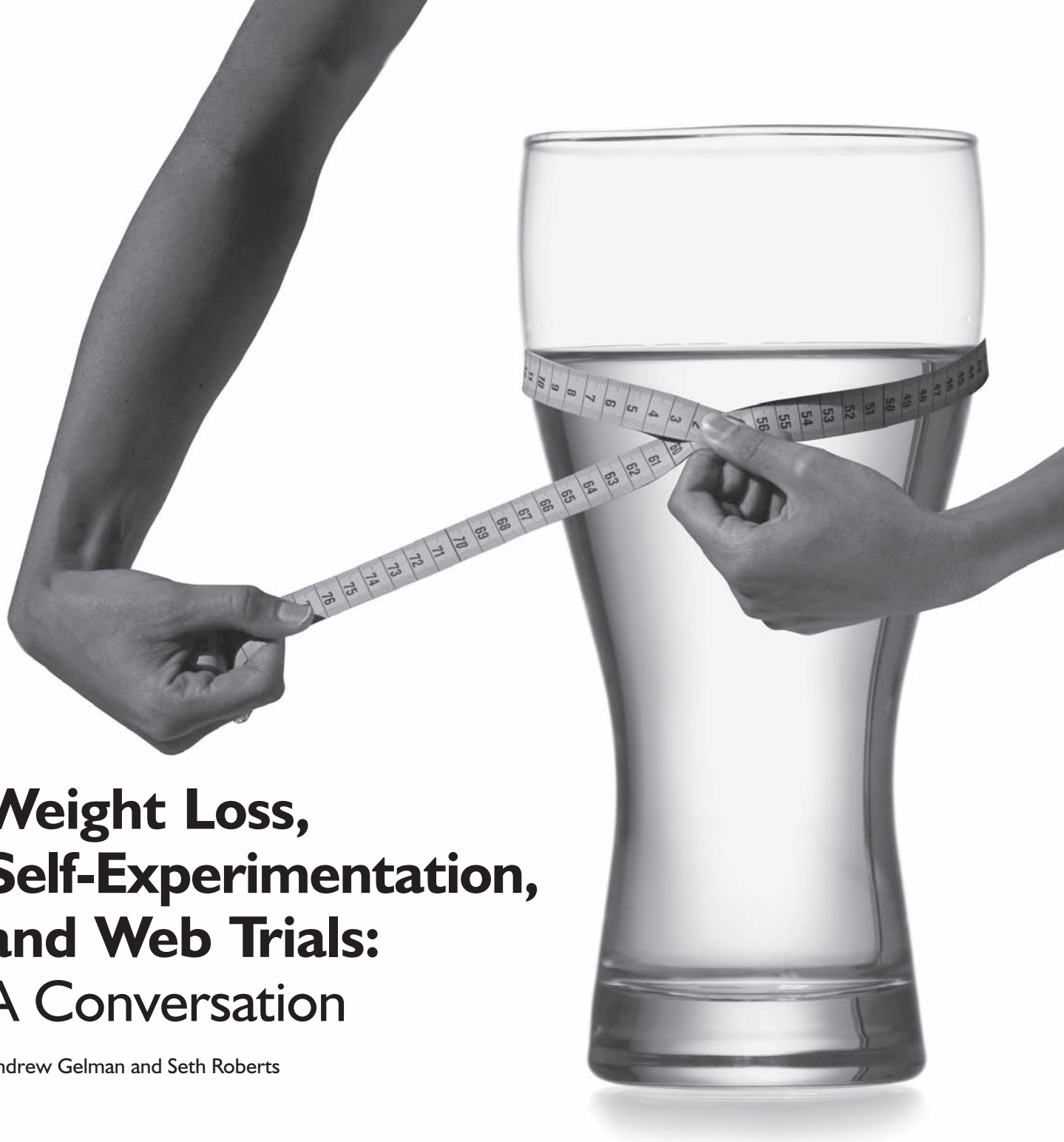


Weight Loss, Self-Experimentation, and Web Trials: A Conversation

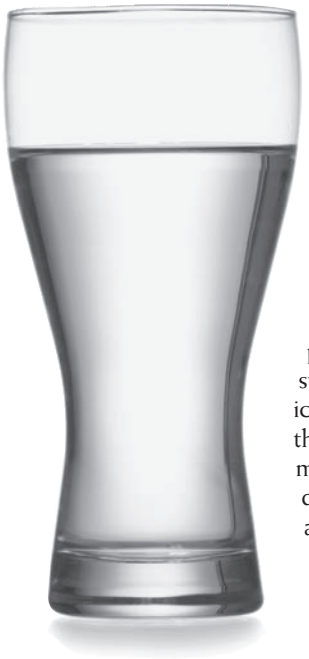
Andrew Gelman and Seth Roberts

The authors of this article have been friends since co-teaching a course on left-handedness more than 10 years ago. More recently, each has started a blog, which has increased their awareness of the benefits of informal exploration of scientific ideas. (Andrew Gelman's blog covers various topics in social science and statistics; Seth Roberts' is mainly about self-experimentation and scientific method.) What follows is an online conversation (slightly edited) from a few days in 2007, in which Gelman and Roberts used instant messaging.

SR: I want to ask your opinion about web trials. People go to a web site, where they choose or are randomly assigned a treatment. Then, they come back and report the results.

AG: Then, the records of their choices and outcomes are made publicly available.

SR: Yes. And there would probably be some summary of the results prepared by experts. It wouldn't be just raw data.



Comparing to Current State of the Art in Medical Research

AG: We could compare to the current state of the art in medical research, which I think is to have some moderately large randomized clinical trials, each of which is published in a journal, followed by a meta-analysis of these trials.

A difficulty with the current state of the art is that sample sizes in clinical trials seem to be simultaneously too small and too large. Too small in that results tend to be just barely statistically significant (and often not significant for subgroups), so that you can't really put your faith in one study, hence the need for meta-analysis. Too large in that each study is unwieldy, takes a huge amount of effort, and doesn't allow for much learning and experimentation during the study.

SR: Richard Doll, a famous epidemiologist, once said that if the effect is strong, you don't need a big study.

AG: In some way, the high cost is a good barrier in that people have to think seriously and justify what they want to do. On the other hand, within any particular research plan, it would seem to limit the possibility for innovation.

Speaking generally, a challenge is to integrate clinical judgment (including ideas of experimentation and trying different things with different patients) with scientific goals such as replicability.

Also, there are well-known cognitive illusions in clinical judgment, which is what motivates the evidence-based-medicine movement (for randomized trials, public records of data, etc.) in the first place.

SR: How do web trials fit into the picture you have drawn?

AG: Ideally, web trials are intermediate between controlled randomized trials on one hand, and full recording of observational data on the other. If people are really volunteering to be randomized and they follow the protocol, then this is a clean randomized experiment (albeit not blinded, an issue I'd like to raise with you). In practice, there will be lots of selection, dropout, measurement error, etc., which moves it toward an observational study. The dispersed nature of the data collection is similar to (in fact, more dispersed than) the idea of individual clinicians recording their experiences and outcomes into a centralized database. That is, the data collection is dispersed, the database is centralized.

SR: A web trial would have more regularity—less variation—across subjects than observations collected from individual doctors, because everyone would get the same instructions; whereas, in a usual experiment, different doctors are obviously going to give different instructions (for the same nominal treatment).

AG: Yes. That's why I said the web trial is in between.

Difficulties with Blinding

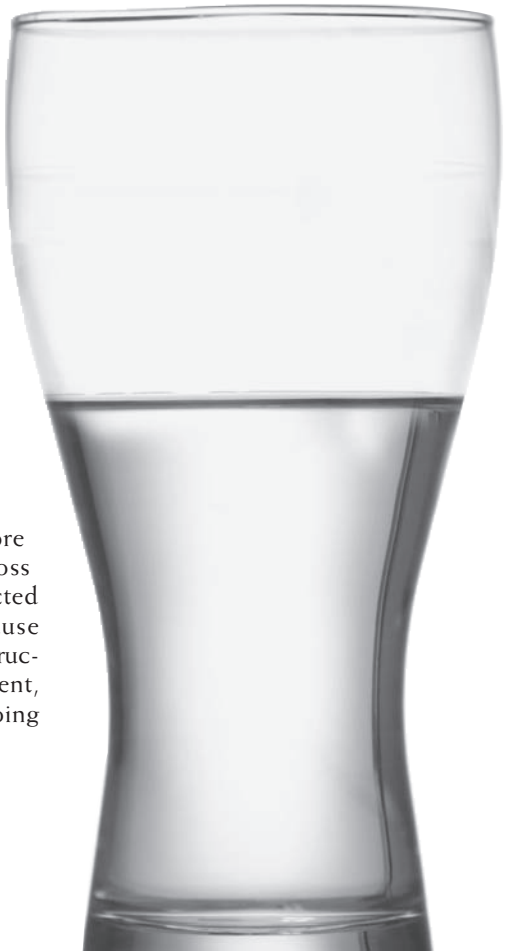
SR: In the area of blinding, I think a web trial would be better than the conventional double-blind clinical trial, if the goal is to guide practice. In practice—in real life—patients are not blinded. Blinding is a tool to equate expectations. Better to equate expectations by comparing different treatments both believed to be effective.

AG: One of the difficulties with your self-experimentation is that there's no blinding at all, which is similar to these trials. Some of it is the nature of your treatments, but perhaps with some effort, you could come up with blinded versions.

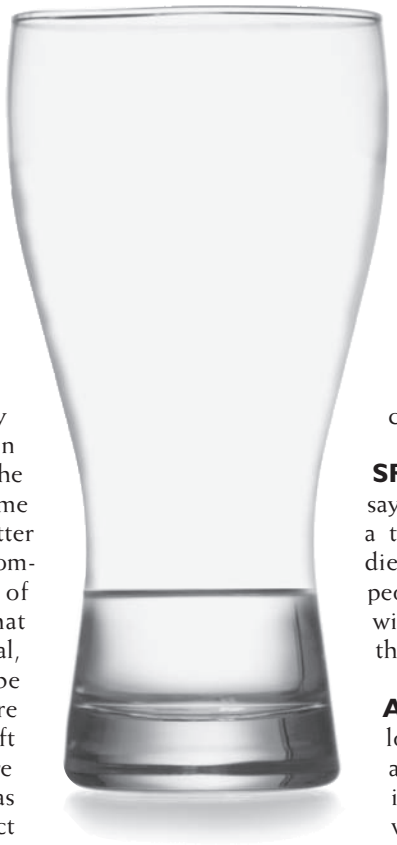
SR: In my self-experimentation, the expectations are equal in the different conditions, in many cases.

AG: For example, consider the recent self-experiment you describe on your blog, where you try different oils and measure your balance. I'd believe these

"Richard Doll, a famous epidemiologist, once said that if the effect is strong, you don't need a big study."



results a lot more if you blinded the treatments.



SR: Sure, blinding would help in that case. I agree. I plan to do something like that. But blinding is not necessary to equate expectations. For example, I tried many ways of losing weight. In every case, I expected the treatment to work. Some ways worked much better than others. It is this comparison of the effects of different treatments that is interesting. In general, expectations cannot be very powerful or there would be no problems left to solve. Expectations are powerful in a few areas and seem to have no effect in many areas. I don't mean we should ignore them, but to emphasize them as a big deal is not what the evidence suggests. In any case, in web trials, the participants would only be randomized (or choose) treatments they thought might work.

AG: There's some work by statisticians and economists on "broken randomized trials," which can more generally be thought of as experiments that have partial randomization.

SR: I think of web trials as giving "entrants" (or subjects) a choice: To be or not to be randomized. Then, when it's all over, you compare the two groups.

AG: That makes sense. You'll still have some problems, such as subjects failing to follow the protocol, bias resulting from the failure to blind subjects to treatments, and possibly other problems.

SR: Well, these are equal for all conditions, so they shouldn't distort anything.

AG: In a controlled trial, you can deal with some of these things. You have more opportunity for interaction with the experimental subjects, which may improve compliance with the protocol. In a controlled trial, you can (sometimes) ensure blindness. In general, I don't think you can get away with assuming that biases cancel out.

SR: I think you are saying there could be a treatment-by-obedience interaction—people more obedient with some treatments than others.

AG: Failure to follow protocol can be a serious problem in clinical trials, as well. For example, if one treatment has an unpleasant taste or side effects and the other doesn't, then compliance could easily depend on the treatment.

Analyzing Data from Web Trials

AG: Your web trials should give us a big, juicy source of data that can be thrown at a statistics PhD student as a thesis project! My intuition as an amateur sociologist of applied statistics is that an exemplary applied analysis is a good way to kick-start the study of a statistical problem.

SR: What's an example of such a kick-start? That's an interesting point.

AG: I'm thinking of the hierarchical models that were fit by Novick et al. (1972), Lindley and Smith (1972), and others in the late 1960s through early 1980s to educational data. These provided examples for people to follow—templates—as well as demonstrations that these methods really worked. There were various interesting discussions about these models in the stat literature, in particular I'm thinking of a paper by Rubin (1980) on law school validity studies that had several discussants.

SR: Yes, it is true the data from web trials would be complex and interesting in new ways and accessible to everyone.

AG: Yes, having available data is another plus—that's really a new feature that should help. Now back to the warnings. A very well-known example is the Nurses Health Study, an observational study that found that taking post-menopausal drugs was associated with lower heart-attack risks (and lower death rates). But when a big randomized experiment was done, they found that taking the drugs slightly increased risks of cancer, heart attacks, and strokes.

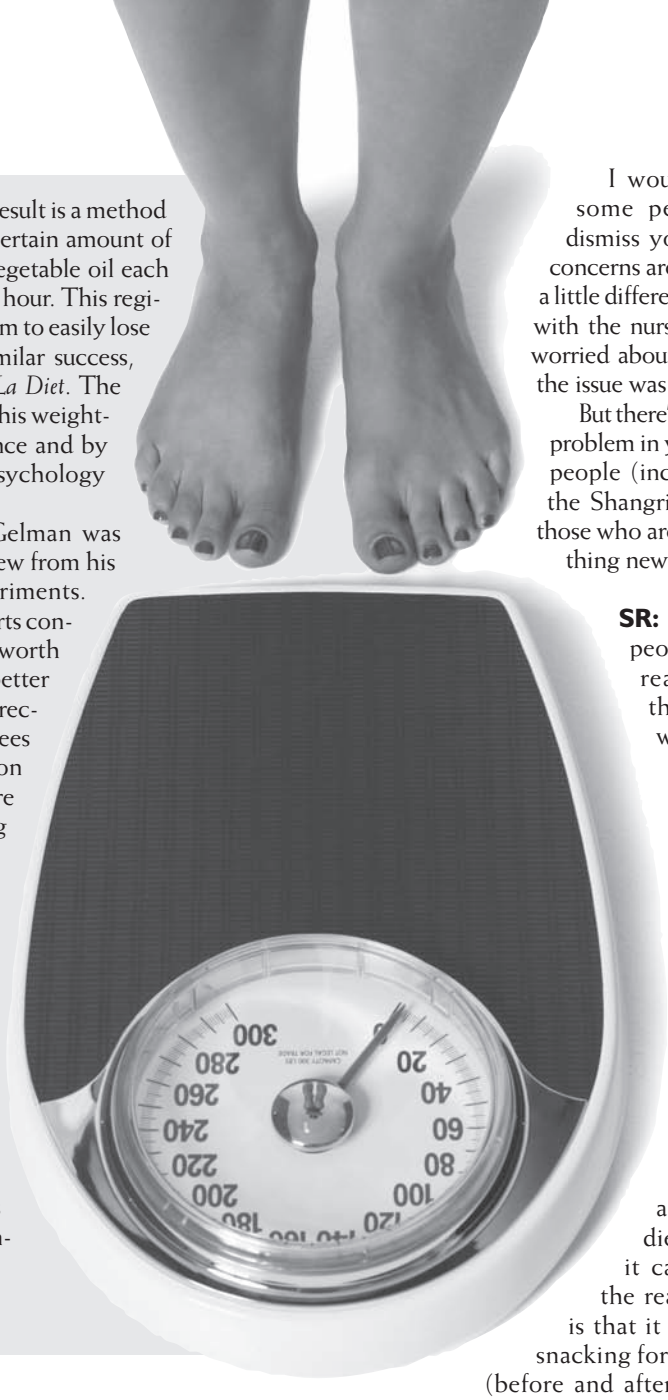
I talked with various people about this, and there are different potential explanations for the discrepancies. One story is that the women who took the drugs were otherwise healthier, more health conscious, etc., even after controlling for whatever pre-treatment variables they controlled for. Another story is that the populations of the two studies were different (in particular, in their average ages), and perhaps the drugs are beneficial for some ages but not others. (Incidentally, the drugs were not originally intended to reduce heart-attack risk. This was an unexpected effect [or noneffect], I believe.)

Anyway, the people I trust on these matters believe the difference is because of "selection" (i.e., the drugs don't really reduce heart-attack risk). But the observational study led people to recommend the drugs. So, this is a big example where the observational study was misleading.

Meanwhile, the nurses study continues to operate and make headlines such as "Obesity Protects Against Breast Cancer" and "Grandkids Can Make You Sick," so this is a live issue with this study and observational studies in general.

SR: Did the randomized study conclusively rule out the effect size seen in the correlational study, or did it simply find no effect? J. Ioannidis, A. Haidich, and J. Lau compared 24 observational studies of various treatments with 24 experimental studies of the same treatments and found the effects were roughly the same size.

AG: In this case, the experiment actually contradicted the observational study—a statistically significant negative effect



Roberts' best-known self-experimental result is a method for losing weight based on drinking a certain amount of unflavored sugar water or unflavored vegetable oil each day, separated from meals by at least an hour. This regimen reduced his appetite and allowed him to easily lose 30 pounds. After other people had similar success, he wrote a book about it, *The Shangri-La Diet*. The book describes how Roberts developed his weight-loss method through personal experience and by reading nutrition and experimental psychology research.

As many statisticians would be, Gelman was skeptical of the conclusions Roberts drew from his nonrandomized, unblinded self-experiments. Coming from the other direction, Roberts concluded that, for generating new ideas worth testing, small-scale experimentation is better than the large clinical trials statisticians recommend as the gold standard. Roberts sees a similarity between self-experimentation and exploratory data analysis: Both are nimble methods for learning something new in contrast to large, formal, randomized experiments and traditional hypothesis testing as are often done in medical research. These large, preplanned studies can be clunky, inflexible, and discouraging of innovation. Roberts is coming from psychology, where—as in industry—researchers usually make progress by doing many small experiments, rather than a single huge clinical trial that is intended to be definitive. Web trials—treatment tests done via the web—have elements of both self-experimentation and conventional clinical trials.

for one and a statistically significant positive effect for the other. It wasn't just that there was significance for the experiment and no significance for the observational study.

SR: I'd like to return to the issue of blind versus don't blind. You believe any experiment where subjects are not blind to the treatment has a problem?

AG: Yes, if knowledge of the treatment could affect the outcome (for example, through motivation). I worry about it for your diet and depression studies.

SR: Well, in much research, the first question is whether there is a useful effect. Later experiments deal with mechanism. I was under the impression that what matters is to equate expectations across conditions and that blinding is just one way to do this.

AG: Maybe you're right. I'm not actually up on this literature. I think P. Rosenbaum discusses these issues in *Observational Studies*.

More on Blindness: Considering the Shangri-La Diet

AG: My knowledge of blindness is not particularly sophisticated. For your diet and depression studies, there are obvious stories based on motivation. I wouldn't go so far as some people and simply dismiss your results. But the concerns are natural, I think. It's a little different than the problem with the nurses study. Here, I'm worried about motivation; there, the issue was selection.

But there's a possible selection problem in your study, too. The people (including you) doing the Shangri-La Diet might be those who are ready to try something new and lose weight.

SR: There are a lot of people who are always ready to try something new and lose weight.

AG: Again, this could be tested with a blinded study. For example, half the people get the oil apart from a meal and half get the oil with the meal. Not that this would solve all problems of interpretation … For example, I told a friend about the diet and she believes it can work, but that the reason why it works is that it stops people from snacking for a two-hour period (before and after the oil) and also focuses people on their snacking.

SR: If anyone thinks that—and it is a perfectly reasonable thing to think if you are just starting to learn about it—then they can replace the oil with water and see if they continue to lose.

AG: In response to your comment, "There are a lot of people who are always ready to try something new and lose weight," yes, I remember you saying this before, and this is a big reason I wouldn't dismiss your results immediately. But, still, people willing to try this wacky new thing might be special (on average). To put it another way, I expect there were similar successes with people trying Scarsdale, Atkins, etc.

SR: I'm sure people who try my diet are unusual early adopter types. I think Atkins has some truth to it, some reasons it would actually work. I don't know enough about Scarsdale to comment. My theory says that merely changing what you eat (to foods with unfamiliar or at least less familiar flavors) should lower your set point.

AG: Sure, but you had another point, which was that these were people for whom nothing worked before. I was just using these diets as examples of other things that worked when nothing worked before. It relates to the historical perspective of new diets as things that will work for a few years before burning out—possibly because the new diets can motivate people.

SR: I tend to think they burn out because the new food becomes familiar.

AG: I'm not saying this is necessarily true of your diet—yours might be different—I'm just giving a historical control to give insight as to how there could really be motivational issues.

SR: That's true. Research to distinguish my explanation of the burn-out and a motivational one could be done, but of course, hasn't been.

AG: Your theory, "they burn out because the new food becomes familiar," is plausible. It's also plausible that it's easier to motivate yourself with a plan that's new and different.

SR: I hope there will be studies of whether the theory behind my diet is correct. These would essentially be studies that test the prediction that familiarity matters. This is a prediction that other theories do not make.

AG: Based on reading the appendix to your book, there's still some research synthesis that needs to be done (presumably with the help of animal studies).

SR: I agree.

Back to Web Trials

SR: Web trials are relatively early in the research chain and are relatively practical. In these cases, you don't worry a lot about mechanism. You worry much more about efficacy: Is there an effect?

AG: Regarding the analysis of web trials, it would be interesting to look at other examples of partially randomized experiments. J. Barnard and others worked on a study of school choice where they looked into some of these issues. It was a study that randomized some aspects of which kids went to which schools, but parents had some choices, too.

In medicine and also in economics/public policy, there has been a lot of interest in trying to get inside this sort of study, rather than just relying on the "intent to treat" or explicit randomization.

SR: "Get inside this sort of study." What do you mean?

AG: You should look at the treatments actually chosen by the subjects, not just at the treatments to which they were assigned. This way, you can learn about the process of selection. You can also use methods such as principal stratification to estimate the effects of the treatment among different subpopulations, such as compliers and noncompliers.

SR: Could you sum up why you like the idea of web trials?

AG: Web trials have the potential for gathering lots of data. In addition, getting people's active participation motivates them to randomize, to apply the treatment, and to record results. More generally, web trials have the potential to get people involved in the project as participants, not just 'subjects.'

SR: Those are good points.

AG: I'm still struggling with the question of size. Are medical experiments too small (because they don't have the power to give definitive results) or too big (because they don't allow for innovation during the time of the study)?

SR: Maybe different goals need different tools—and different-sized experiments. One goal is to come up with new ideas worth testing; another is to test those ideas.

Acknowledgments

We thank Albyn Jones for helpful comments and the National Science Foundation, the National Institutes of Health, and the Applied Statistics Center at Columbia University for financial support.

Further Reading

American College of Obstetricians and Gynecologists. (2004). "Frequently Asked Questions About Hormone Therapy. *www.acog.org/from_home/publications/press_releases/nr10-01-04.cfm*.

Barnard, J.; Frangakis, C.E.; Hill, J.L.; and Rubin, D.B. (2003). "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association*, 98:299–324.

Harvard Nurses Health Study. (2006). *http://harvardscience.harvard.edu/node/765*.

Ioannidis, J.P.A.; Haidich, A.B.; and Lau, J. (2001). "Any Casualties in the Clash of Randomized and Observational Evidence?" *British Medical Journal*, 322:879–880.

Lindley, D.V. and Smith, A.F.M. (1972). "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society B*, 34:1–41.

Novick, M.R.; Jackson, P.H.; Thayer, D.T.; and Cole, N.S. (1972). "Estimating Multiple Regressions in M-Groups: A Cross Validation Study." *British Journal of Mathematical and Statistical Psychology*, 25:33–50.

Roberts, S. (2001). "Surprises from Self-Experimentation: Sleep, Mood, and Weight." *CHANCE*, 14(2):7–12.

Roberts, S. (2004). "Self-Experimentation as a Source of New Ideas: Examples About Sleep, Mood, Health, and Weight." *Behavioral and Brain Sciences*, 27:227–262.

Roberts, S. (2006). *The Shangri-La Diet: The No-Hunger Eat Anything Weight-Loss Plan*. New York: Putnam.

Rosenbaum, P. (2002). *Observational Studies (2nd Ed.)*. New York: Springer.

Rubin, D.B. (1980). "Using Empirical Bayes Techniques in the Law School Validity Studies (with discussion)." *Journal of the American Statistical Association*, 75:801–827.

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