Discussion of "Statistical Research: Some Advice for Beginners" by Hamada and Sitter

Editor's Note: In the May 2004 *TAS*, Hamada and Sitter provided advice for the statistics graduate student on doing research. We asked several prominent statistics researchers to discuss this article, and *TAS* readers were also invited to contribute discussion. In the following articles, we get some personal insight on how statisticians do research, and we also see that research is an activity for all statisticians, not just the ones in academia.

—James Albert, Editor, *The American Statistician*

On Advice for Beginners in Statistical Research

Donald B. Rubin

I enjoyed reading Hamada and Sitter (HS), and I thank the editorial board of *The American Statistician* for inviting me to contribute a discussion. HS offer many very helpful gems of advice, and I would encourage graduate students contemplating research projects to read it carefully. Nevertheless, my own feeling is that actually conducting research is an extremely personal enterprise and styles for doing it, as well as types of contributions, can vary dramatically across successful researchers. Consequently, I would not necessarily suggest that all budding researchers actually try to follow all of the suggestions presented by HS, but rather only those suggestions that seem to resonate with them.

For example, I essentially always start with a real problem in "science" (broadly defined to include—in addition to the traditional sciences like astrophysics and chemistry—social sciences such as economics and education). Fortunately, I've always been surrounded by such work, partially through collaborations with wonderful colleagues, and partially through consulting projects. Working on a real problem, not a theoretical one, is important to me for at least two reasons. First, even if I fail to make some theoretical contribution along the way, at least I may have contributed to the solution of the real issue, which is satisfying. Second, the pressure to do something really clever technically is not there—some real data problems do have simple and straightforward solutions, and then these solutions should never be eschewed. However, when attacking a real problem without an immediate solution, I always try to devise an approach that addresses the actual problem but does so in broad generality, so that my approach addresses not just the problem at hand but a large class of problems of which the current one is just an example. Work on a specific problem, but think abstractly. (Thanks to George Cobb for this very succinct statement made at the 2004 New England Statistics Symposium.)

In the same vein, I look for outrageously simple but general solutions based on some "grand" conceptualization of the problem. Having physics as my first intellectual love, I am devoted to solutions based on insights arising from symmetry arguments and grand conservation principles (e.g., energy, momentum, angular momentum) from which the solution is "immediate" without tedious technical calculations. Consequently, I always hope to find answers to statistical problems that require hard thought to conceptualize correctly, but once that is accomplished, no real technical argument is needed—perhaps something as simple as using new concepts and notation that will lead to a two-line proof rather than a two-page one. For me, simplicity is never to be avoided in favor of complexity. Specific examples from my own work may be my contributions to missing data (e.g., Rubin 1976, 1987; Dempster, Laird, and Rubin 1977) and to causal inference (e.g., Rubin 1974; Rosenbaum and Rubin 1983; Angrist, Imbens, and Rubin 1996).

Also, I personally avoid reading what anyone else has written or done on the problem at hand until after I think I have arrived at a general approach on my own. I feel that it is far easier to be creative in my attack when my head is not filled with others' approaches to the same problem or others' classifications of the problem. This attitude may be viewed by some as reflecting "lazy scholarship," and so PhD students may want to use this tactic cautiously, but it is the one I have always used, although it is somewhat idiosyncratic.

I fully agree with the HS view that the act of writing is, in itself, part of the creative research process. For me, ideas come to life only after they are written down. Moreover, once written, they can be critiqued, and the writer should be the most fierce critic. When reading something you have just written, try to flush your mind of anything you learned in the process of doing the research for the paper, and see if you, as a critical reader, find the written article interesting and persuasive, and the approach no more complex than needed to address the topic of the paper (e.g., extraneous complex technical mathematics, although possibly appealing to some more mathematical readers, is entirely off-putting to the statistician (and latent physicist) in me, and I think to many readers). This "mind-flushing" is difficult to do, especially when reading the same paper repeatedly in a short period of time, but I see it as essential. If necessary, at some point put the paper aside for a while, maybe even for many months. Of course, this entire process becomes much easier with a good co-author, and I have benefitted tremendously from having had many great ones, especially helpful with books [e.g., Rod Little

Donald B. Rubin is John L. Loeb Professor of Statistics and Chairman, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: rubin@stat.harvard.edu).

 DOI: 10.1198/000313004X1576

(Little and Rubin 1987, 2002); Andy Gelman, John Carlin, and Hal Stern (Gelman et al. 1995, 2003)].

Finally, this brief discussion on conducting research would not be complete without stating the huge role that "wiser and older" heads can have. My own views and work have been heavily influenced, primarily by my extremely sagacious yet practical PhD advisor, Bill Cochran, and secondarily by a variety of memorable conversations during my formative years with George Box, David Cox, and Art Dempster. Thanks to them and many others who have contributed so substantially to current and future generations of statistical researchers.

REFERENCES

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion) *Journal of the American Statistical Association*, 91, 444–472.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data Via the EM Algorithm" (with discussion) *Journal of the Royal Statistical Society*, Series B, 39, 1–38.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995), *Bayesian Data Analysis*, New York: Chapman and Hall.

——— (2003), *Bayesian Data Analysis* (2nd ed.), New York: CRC Press.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: Wiley; translated into Russian in 1991: *Finansy and Statistika* Moscow, Publishers, Andrei Nikiforev, translator.

——— (2002), *Statistical Analysis with Missing Data* (2nd ed.), New York: Wiley; translated into Chinese in 2003: China Statistics Press, People's Republic of China.

Rosenbaum, P., and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

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

——— (1976), "Inference and Missing Data" (with discussion) *Biometrika*, 63, 581–592.

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Discussion of "Statistical Research: Some Advice for Beginners" by M. Hamada and R. Sitter

Alan E. GELFAND

Hamada and Sitter, in the tradition of recent "advice" articles in *The American Statistician* (e.g., Murphy 1997; Stasny 2001; Vardeman and Morris 2003) have provided a well-written, well-organized, well-researched, valuable overview of the process of doing statistical research. They have covered much ground in an engaging manner. My comments here are intended as a supplement, elaborating a few themes that, in my experience, help to illuminate the *game* of research. The use of the word "game" is not intended to be cynical, pejorative, or demeaning with regard to scientific endeavor. Building a successful research career requires learning the "rules" and developing an "effective strategy" and, in that sense, becoming skilled at the game.

Currently, statistics is a rapidly evolving field. Perhaps more than most fields, it has been a huge beneficiary of the wide availability of inexpensive, high-speed computational power. As a result, for many statisticians, the nature of research has changed. A modern perspective recognizes that contributions can be made in many, not necessarily disjoint, ways—methodology, mathematical rigor, modeling, serious application, computation. There is no hierarchy or pecking order in these contributions and it may prove frustrating and unsatisfying to attempt to be a master of all of them. In this regard, it can be useful to identify your strengths and pursue a research agenda that is consistent with these strengths.

Furthermore, statisticians increasingly find themselves involved in interdisciplinary work. Collaborating with scientists in other fields often provides wonderful opportunities for novel statistical work and, at the very least, will make you a richer scientist. A frequent by-product of interdisciplinary work is the development of research teams, drawing scientists from several fields, bringing in post-docs and graduate students. Team meetings typically result in exciting and exhilarating brain-storming experiences and provide a wonderful mechanism for revealing to junior researchers what is arguably the most critical component of the research enterprise. In my view, aspiring researchers should actively seek out such opportunities.

A caution may be appropriate here. Research teams produce multi-authored papers that will often find homes in subject matter journals. How useful will such papers be to you, as a beginning researcher? This will likely depend upon the environment you are working in. Some groups reward such activity more than others. An obvious strategy is to strike a balance. Devote a portion of your research effort to interdisciplinary projects and, also, a portion to problems where progress will be reportable in statistics journals, where you work by yourself or with a statistics colleague. Still, the value of the interdisciplinary experience should not be underestimated. Often, the novel statistical work will find its way to a strong statistical journal (with a reduced set of co-authors). Also, such experience facilitates opportunities for pursuing and securing external funding. In many environments, such funding is an important yardstick for research performance.

Returning to my earlier remark regarding the need for an effective research strategy, it is vital to be a *finisher*. No negative connotation is attached to this term. Work must be written up, submitted, revised, and resubmitted or perhaps sent to a different journal, and so on, all in a timely fashion, in order to make its way through the pipeline to publication. This is not to say that research should be rushed, but rather to state the obvious—other scientists will not likely be aware of what you are doing, invite you to present what you are doing, or cite what you are

Alan E. Gelfand is J. B. Duke Professor of Statistics and Decision Sciences, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251 (E-mail: alan@stat.duke.edu).

doing unless it is available. A corollary to this is that, if a problem is substantial, it need not be solved in its entirety before it is ready to be written up and submitted. It is often useful to divide the work on a problem into portions and to target different portions to suitable journals. Within the portions, detailed outlines and multiple drafts are invaluable (Hamada and Sitter noted this as well) both to clarify what needs to be done to complete a manuscript as well as to help articulate the partitioning into portions.

A related notion is that a successful research program requires "multi-tasking." You should plan to be involved in more than one project, to work on more than one problem. The variety will make the research component of your life more interesting, will help you through lulls in any particular project, and will ensure that you have several manuscripts proceeding through the publication pipeline at any given time.

Though at first thought it may seem peripheral, it is very important to be visible in the community or at least in the subcommunities that are working in the areas you are interested in. Of course, this means readily accepting seminar invitations and regularly traveling to conferences. Smaller, more focused gatherings are particularly worthwhile. Participation in such meetings through poster sessions enables you to show your work to others (and, of course, enables you to see what others are doing). Stimulating conversations ensue, research relationships are facilitated, friendships are made.

I conclude with a comforting thought. We are fortunate that statistics is a very supportive discipline. We do not tend to be protective, competitive, or ruthless with regard to our work. This may be due to the fact that we are not a "glamor" field, we are not a field where the prize is to the first one there. But we tend to be a nurturing field. Established researchers eagerly draw junior people into projects and happily attempt to provide mentoring. You certainly need not travel your research path alone.

REFERENCES

Murphy, J. R. (1997), "How to Read the Statistical Methods Literature: A Guide for Students," *The American Statistician*, 51, 155–157.

Stasny, E. A. (2001), "How to Get a Job in Academics," *The American Statistician*, 55, 35–40.

Vardeman, S. B., and Morris, M. D. (2003), "Statistics and Ethics: Some Advice for Young Statisticians," 57, 21–26.

The Importance of "Why?"

Elizabeth Thompson

Hamada and Sitter are to be congratulated on an article that should be required reading for every beginning research student in the statistical sciences. Although no two researchers will agree on every point, every point made is worth discussion among students and with their mentors. Above all, the emphasis is that the beginning, the progress, the finishing, and the defending are responsibilities of the student not of the advisor. This is something that too many students discover too late.

The authors do an excellent job of addressing many aspects of "How?", in finding problems, reading literature, doing the research, and presenting it both orally and in writing. However, I believe there is a question far more fundamental than "How?" that researchers must constantly ask themselves. This is "Why?"

"Why?" can be asked at many levels, and in all aspects of research. In reading the literature, it stretches from "Why does this line follow from the previous one?" through "Why are these assumptions sufficient for this result?", "Why are these assumptions necessary for this result?", "Why does the writer want to show this result?", and "Why has the writer chosen this model?", to "Why has the writer chosen this general approach?" and "Why is this paper interesting or important?" A student who asks "Why?", and who attempts to discover a satisfactory answer to this question, is doing research. Someone who does not ask "Why?" will never become a good researcher.

Having learned to apply "Why?" in reading, it is equally necessary to apply "Why?" to every aspect of one's own research, again from the details of results, to the collection of results, to the overall approach, and to the thesis as a whole. If you can explain why your results follow from your assumptions, why your assumptions are necessary as well as sufficient, why your approach and model are the appropriate ones to address your problem, and why your problem is an important one to answer, then you can defend your thesis.

Once one has learned to ask "Why?", the "When to start?" is already accomplished, and the "How?" and the "What?" will follow.

Elizabeth Thompson is Professor, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195 (E-mail: thompson@stat.washington.edu).