The *Sociological Methodologist*

Newsletter of the Methodology Section, American Sociological Association
Winter 2008

Chair: Ross M. ("Rafe") Stolzenberg *Chair-Elect:* Tim Futing Liao *Secretary/Treasurer:* Guang Guo
Council: Kenneth C. Land, Pamela Paxton, Aimee Dechter, Trond Petersen, Lingxin Hao, and David B. Grusky
Newsletter Editor: Lawrence E. Raffalovich *Assistant Editor:* David P. Armstrong

Contents

Methodological Shamanism and the Review Process[1]

by Ross M. Stolzenberg
r-stolzenberg@uchicago.edu

While we can never erase all doubt about the accuracy of particular statistical e*stimates*, we can be absolutely certain about key characteristics of proper estimation *methods*. I call methods "proper" if their key characteristics are established by mathematical proof. Even the faithless can have faith in proofs. For example, we can prove that certain estimators are consistent, which means that bigger samples tend to give smaller confidence intervals, other things equal. If someone claims otherwise, then they have made a *mistake*—a human behavior that tends unintentionally to produce avoidable, unwanted outcomes.

There's a big difference between estimation errors and methodological mistakes. A lot of mistakes are caused by inattention and ignorance. Statistical mistakes occur when analysts are careless with data or ignorant of the methods they use. For example, notorious mistakes include the analysis of ordinal data with methods designed for interval measures. Mistakes are often inconsequential, but mistakes are not dependably inconsequential. So it is that regression often leads to the same conclusions as ordered probit analysis, but only the ignorant would depend on the inevitability of that fortunate result.

Ignorance and carelessness cause mistakes but do not excuse them. If discovered, ignorant, careless statistical practice usually is punished by public shaming and ostracism. Shaming is done in the "commentary and debate" section of journals, in the audience question-and-answer period at colloquia, and in the formal discussion of presented papers at professional meetings. Ostracism is done by rejecting papers for publication in journals or presentation at meetings. Journal editors, meeting organizers and conveners of proposal review panels keep the wheels of shaming and ostracism turning by selecting reviewers and discussants.

---

[1] *The author is responsible for all opinions and assertions expressed herein. Thanks for comments and advice to Thomas DiPrete, Robert Hauser, Kenneth Land, Lawrence Raffalovich and Michael Sobel.*

Of course, reviewers and discussants are sometimes ignorant and careless too. Fortunately, in professional meetings and scholarly journals, authors can respond to criticism of their published and presented work, when and where criticism is made. In these settings, ignorant or careless critics risk the same humiliations as ignorant and careless authors. The system is often needlessly harsh, but it does motivate practitioners to seek advice, check their work, and thereby overcome the human tendency to do otherwise.

However, there is no regular opportunity for authors to respond to mistaken criticism of papers and proposals submitted for journal publication and grant funding. Reviewers sometimes agree to evaluate papers and proposals without realizing that they lack the time or expertise to do so properly. Editors and proposal decisors often lack the time and expertise necessary to recognize mistaken reviews. Victims of ignorant or careless evaluations sometimes cannot ask for reconsideration, but sometimes they can; those who can ask sometimes do so; and those who do ask sometimes get what they ask for. But reviewer anonymity and editorial confidentiality keep mistaken review problems in the dark. With this problem out of sight, most decision-makers seem hesitant to believe that a mistaken-review problem exists at all, even if some authors complain of it. Rejected applicants and authors are famous for complaining. But I think that most authors are afraid to complain, do not know that they can, or doubt the value of complaint, even when they have good reason to do so.

But the mistaken review problem does exist, and some of us think it is getting worse. A former chair of the Methodology Section recently stated recently that we are witnessing the resurgence of "methodological shamanism," in which ill-informed, unsupported opinions by self-appointed "methods gurus" are given more editorial weight than the mathematically proven properties of well-established statistical methods. Methodological shamanism is the goofy "proof by authority" that was debunked in the age of enlightenment. The former chair may be right: Some time ago, I was shown a decision letter by the editor of a flagship journal that concurred with a wooden-headed reviewer who suggested that statistically significant results would not be significant if they had been obtained from a *larger* sample. In another case, an editor agreed with a Luddite reviewer who objected to the use of ordered probit analysis of an ordinal dependent variable—reviewer and editor agreed that ordinary regression would be a better choice because it was "easier." And, in another case, reviewer and editor opined for unstated reasons that it would be fine to accommodate independent variable nonlinearities with dummy variables, but that substantive theory would be required to justify the same accommodation with fractional polynomials. When I edited *Sociological Methodology*, I heard regularly from flabbergasted authors seeking advice after receiving mistaken reviews at substantive journals.

I think that the problem of careless and ignorant reviews is an infection that grows only in the dark. Sunlight could do a lot to reduce the problem. In brief, editors and grant decisors need to know that mistaken reviews are a problem, and they need to know that decisions based on gross mistakes will be known, even if reviewers and authors remain anonymous. So examples of reviews with gross mistakes need to be published or at least publicized and discussed, rather than stuffed in wastebaskets, shredders and filing cabinets. Perhaps a web site or a blog for concise, egregious examples would be useful. A blog or web site could not grant justice to every wronged party, but it could serve as a reminder of the imperfections of the system, and it could help to establish responsibility for the quality of reviews, and reduce the cover of darkness placed over bad decisions that are based on mistaken reviews. Blind review is not a shield for editors.

I hope that publicizing the problem with real examples will cause decisors to consider reviews with the same healthy skepticism that they give to submitted papers and proposals. I hope that publicity causes authors to be less hesitant to ask editors to reconsider the advice of mistaken reviewers. And I hope that publicity will make editors and decision-makers more receptive to authors who provide reasoned evidence that their reviewers are mistaken. When those things happen, then the review process will be less like a beauty contest with an occasional quirky, myopic judge, and more like the universally skeptical, reasoned evaluation procedure that is the hallmark of scientific inquiry and scholarship.

Contemporary Issues in Causal Inference

by Ken Frank
kenfrank@msu.edu

This article discusses contemporary issues in causal inference and how they are being addressed, especially by social scientists. While it is intended as a general introduction, many of the issues are controversial. The presentation below reflects my own take and interpretations rather than a broad review of multiple viewpoints.

The fundamental problem of causal inference is often defined by the counterfactual. As an example: I have a headache; I take an aspirin; my headache goes away. Is it because I took the aspirin? It is impossible to know for sure. We could be certain only if we could have also observed what happened to me if I had not taken the aspirin. But this control condition is impossible to observe for a single individual. It is counterfactual.

Recognizing our inability to observe the counterfactual as a fundamental problem of causal inference at the individual level, scientists typically infer causality by comparing sets of units that received different treatments. A classic example would be a chemist who divides a single solution (composed of molecules) into two parts, one exposed to a treatment and one to serve as a control. Resulting differences are then attributed to the treatment. This example is often considered to be a gold standard for inferring cause.

But even the gold standard requires certain assumptions for making inference. Namely, that the units receiving the treatment and control are homogeneous. But units could be heterogeneous if the original solution is not properly stirred, if the two solutions are exposed to even small differences in conditions not associated with the treatment (such as temperature), or if two solutions have different levels of purity.

Like the chemist, social scientists often analyze sets of units to make causal inferences. But it is precisely when analyzing an aggregate of individuals that causality is uncertain because there may be baseline differences between those who received the treatment and those who received the control, or the treatment may have a different effect for those who received the treatment than for those who received the control.

In order to reduce the baseline differences, social scientists often randomly assign subjects to treatment and control conditions. As sample sizes increase, randomization reduces differences between treatment and control groups, thus making causal inferences based on differences between the treatment and control group more robust. As a result, randomized control trials (RCTs) are currently considered the gold standard for inference in the social sciences.

But RCTs have three important limitations. First, the experiments on which they are based are often different in important ways from treatment conditions as they naturally occur. For example, in attempting to insure uniformity of treatment, experimenters may educate or engage those who implement the treatment in ways that are unlikely to occur outside the treatment. In education, some evaluators require that the treatment be implemented by 70% or 80% of the teachers in a school, a level of implementation that does not occur for most reforms or innovations.

Second, randomization still requires that people are treated and respond independently of one another. This may not be the case if doctors treat multiple patients. In education, many reforms and innovations are implemented through the coordinated activity of teachers. Therefore teachers within a school are not independent of one another. Typically, many sources of dependencies are accounted for by carefully defining the units that can be considered independent (e.g., schools) but this can dramatically increase the cost of RCTs.

Given the above limitations of RCTs, social scientists often make causal inferences from observational data or quasi-experimental designs. Examples include analyses of large scale data bases that have shown a relationship between smoking and lung cancer, or between job training and employment outcomes, or between socioeconomic status and achievement.

Causal inferences from observational studies are tenuous relative to those from RCTs because there could be baseline differences between those who received the treatment and control that could cause differences in outcomes. To account for these, social scientists employ a range of statistical tools. First, social scientists can control for a covariate using the general linear model as in ANCOVA. Second, more complex controls might employ an instrument as an alternative measure of assignment to treatment condition. This may reduce bias in estimation, but requires extra assumptions and has decreased power. Interestingly, a recent meta-analysis [Glazerman, Stephen, Levy, Dan and Myers, David (2003) Nonexperimental versus Experimental Estimates of Earnings Impacts." *Annals, AAPS* (589): 63-85] found that statistical control for a prior measure better estimated effects later obtained from RCTs than estimates using instrumental variables. Third, social scientists have recently started to approximate the counterfactual by matching those who received the treatment with those who received the control based on propensity to receive the treatment.

Even after employing statistical controls for covariates, there may still be concerns that treatment effects could be attributed to uncontrolled baseline differences between treatment and control groups. Defined in terms of the general linear model, my work quantifies how large the impact of an uncontrolled confounding variable would have to be to invalidate a statistical inference.

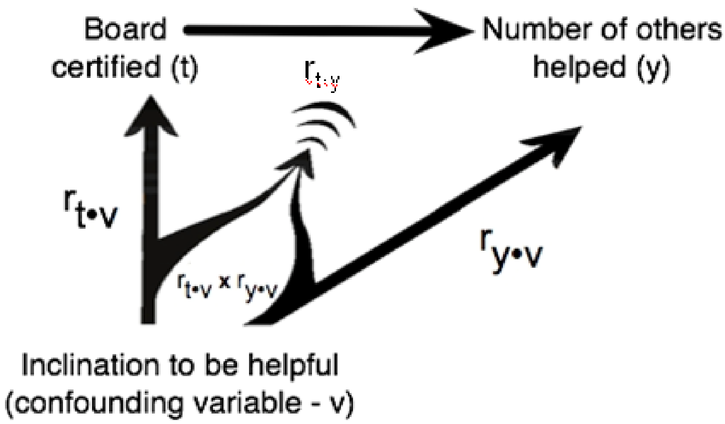
The scenario begins when $r_{t \bullet y}$, the observed sample correlation between t (reflecting assignment to treatment condition) and some outcome y, is statistically significant and is used as a basis for causal inference. Now define the impact of a confounding variable on $r_{t \bullet y}$ in terms of $r_{v \bullet y} \times r_{v \bullet t}$, where $r_{v \bullet y}$ is the correlation between an unmeasured covariate, v, and y; and $r_{v \bullet t}$ is the correlation between v and t. Next define $r^{\#}$ as a quantitative threshold for making inferences from a correlation representing the relationship between a predictor of interest and an outcome. For example, $r^{\#}$ can be defined by a correlation that is just statistically significant or by an effect size. Then, (maximizing under the constraint: impact = $r_{v \bullet y} \times r_{v \bullet t}$):

if the impact of an unmeasured confound
$> (r- r^{\#})/(1- r^{\#}) \rightarrow$ original inference is invalid

if the impact of an unmeasured confound
$\leq (r- r^{\#})/(1- r^{\#}) \rightarrow$ original inference is valid.

The example of the impact of a confounding variable in figure 1 applies to an analysis I recently conducted regarding the inference that attaining National Board certification affects the amount of help a teacher provides to others in her school. Because teachers who are more inclined to be helpful may attain National Board certification, the inference may be invalid. But calculations show that the impact of "inclination to be helpful" would have to be greater than or equal to 0.081 (with each component correlation equal to about 0.28) to invalidate the inference. As a basis of comparison, of the measured covariates, the extent to which a teacher believed leadership would enhance teaching had the strongest impact (.017) on the estimated effect of National Board Certification. Thus the threshold value of .081 shows that the impact of an unmeasured confound would have to be four times greater than the impact of the strongest covariate in our model to invalidate the inference. This suggests the inference that National Board certification affects the amount of help a teacher provides to others in her schools is at least moderately robust with respect to concerns about unmeasured confounding variables. Importantly, robustness indices, as a form of sensitivity analysis, do not alter the initial inference. What they do is to quantify the robustness of the inference to inform scientific debate (for a spreadsheet and sas software for calculating my indices of robustness, and power point and related papers, see http://www.msu.edu/~kenfrank/research.htm#causal).

Figure 1: The potential impact of a confounding variable on a regression coefficient



Causal inference is one of the most rapidly growing areas for social science methodologists. As a result, it is impossible to comprehensively discuss contemporary issues in causal inference in the limited space provided. For more comprehensive or alternative discussions, see work by James Heckman, Charles Manski, Donald Rubin, Paul Holland, Thomas Cook, Steve Raudenbush, Guanglei Hong, Michael Sobel, Philip Dawid, and Paul Rosenbaum. At Michigan State, Jeff Woolridge (Econ) has an excellent discussion of causal inference in his textbook *Econometric Analysis of Cross Section and Panel Data*, Nigel Paneth and Jim Anthony (both in Epidemiology) have written on causal inference especially with respect to experiments, Claudia Holtzman (Epidemiology), teaches a course in causal inference, Daniel Patrick Steel discusses the philosophy of causal inference, and my colleague within education, Barbara Schneider, has written on issues of causal inference and scale-up. Barbara and I are teaching a seminar (CEP991B: section 3) in Spring of 07 on causal inference).

For an excellent introductory and intermediate text, see:

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, NY: Houghton Mifflin.

For a starting point on the web, try http://www.wjh.harvard.edu/~winship/cfa.html.

A Practical Appendix on Curvature in Regressions

by Arthur L. Stinchcomb
a-stinch@northwestern.edu

1. The basic idea of such tests is that, since the multiple regression equation with (say)

Z= b-sub 0 + b-sub 1x + b-sub 2 y + ... + error

is still controlled for x and y when adding a squared term, x-square or y-square or both, or adding the product term xy, to the independent variables, so the added terms only tell whether part of the error term in the linear-only equation above is explained by curvilinearity. Hence the test on the quadratic coefficient will be the same if that term is "residualized" with respect to x, y, or both, as appropriate; I advocate residualizing, then standardizing, all the product or squared terms. That is, regress the products or squared terms on the linear ones, and subtract off the predicted value from the curvilinearity variables. then enter these residualized curvilinearity terms. This leaves the constants and coefficients of the linear equation unchanged and retains the meaningfulness of the statistical tests on the linear effects. Since the controlled linearity variables have no substantively meaningful metric, one might as well have the beta coefficient, which has the usual meaning.

2. If x and y are correlated, then curvilinearity in the relation of x, y, or xy to z all can show up in any of the others. One can show this easily algebraically, or visualize it geometrically. Geometrically, one can almost always find (if the important term is a product of xy to detect interaction effects) a hyperboloid which, in the region of the data, has about the same curvature as a paraboloid, and vice versa. It is, in short, fairly likely that the residualized variables for x-square, y-square, and xy, controlling for the relevant linear ones, will be correlated. So controlling for one kind of curvilinearity (e.g. for a parabolic curve) changes

the meaning of estimates of another (e.g. an interaction effect).

3. None of the curvature metrics, if a product or squared term is added to the regressions is immediately comprehensible. If the squared terms are divided by the standard deviation of the linear variable, then the squared term will have a metric similar to its linear counterpart, so the coefficient might be comparable. One can get a sort of mixed metric by dividing the xy product term by the product of the square roots of the standard deviations of the linear variables. But as mentioned above, I think that one does not usually want a substantive interpretation of the degree of curvature, and it is hard to standardize a curvature metric. I guess Einstein did it.

4. Since one is fitting a surface, the best-fitting curvature in the area of the data depends on the coefficients of all the variables in its formulation (and of course on where one has placed the zero in metricizing the variables). Thus adding any quadratic or product terms tends to create wild variations in the linear coefficients and in the constant term, while the regression tries to find a surface that fits wherever the data happen to fall. This is much reduced by residualizing the quadratic and product terms. Social scientists recognize this in their intestines, so they hardly ever even mention the wild coefficients or wild constants, and never interpret them. Since unless you have residualized, your interaction terms and curvilinearity terms are highly collinear. Only if the measures of x and y have theoretically meaningful zeroes, can we interpret the linear coefficients once we have controlled for the highly collinear quadratic terms, except for their signs. The test of the added variance is OK for curvature, but the tests on the linear coefficients and the constant are now meaningless. I guess the failure of social scientists to use the coefficients confesses that they do not know what a real zero is, in the world they are investigating. They fairly often try to interpret the now meaningless statistical tests of the constants and linear coefficients.

5. So practical advice:

5a. Measuring x and y as deviations from their means makes the squares and product approximately orthogonal to the linear variables. Sometimes this is enough to make the coefficients and constants behave.

5b. Since one is not going to interpret the new coefficients anyway, on can just not report them and report the test and the signs of those coefficients that are needed to interpret the kind of curvature.

5c. Since testing one kind of curvilinearity against another is not going to occur to your reader, who can only think of at most one curve at a time, one can ignore the alternative of a curvilinear effect of one or both correlated variables, and conversely for the curvilinearity implied by the product term if one is interpreting the squared variables. Just test for the interaction or curvilinearity, and no one will notice the alternatives. I predict that in 25 years this will no longer be OK, maybe because solid analytic geometry will come back into the curriculum. I remember going from two to three axes was really hard, especially when I had to draw a graph.

Arthur L. Stinchcombe
Fellow in Sociology at Princeton University and Emeritus Professor of Sociology at Northwestern University.

DNA Testing in a Traditional Social Survey: Report for the Panel Study "Peer Impacts on Attitudes and Drinking Behavior"

from Guang Guo and Greg Duncan
gguo@email.unc.edu

Background

A key component of our proposal was a genetic study in which we would take DNA samples from interviewed students and test for interactions between candidate genes for alcoholism and drinking vs. non-drinking roommates. An important issue is the feasibility of the DNA data collection: will college students consent to providing saliva DNA samples that enable the research team to conduct the genotyping?

To investigate this issue, we conducted a pilot study at the beginning of the 2006-7 school year. Here we report on compliance rates from the pilot study.

The Pilot Study

A seed grant of $25,000 from a large southern public university enabled us to conduct the pilot test. We have now successfully completed the pilot project. Of the 200 randomly-selected first- and second-year students at the university whose names and contact information were provided to us by the Student Housing Department, 191 are deemed eligible participants. Excluded from the study were three individuals under age 18, four individuals who did not live on campus, and two individuals who did not list their email addresses in the student information file provided to us by the Student Housing Department.

Of the 191 eligible participants, 77.0% (147) completed a web-based survey that was similar in length and sensitivity of subject matter to the one we envisioned in our proposal. They were paid $5 for their efforts. Of the 147 who completed the web survey, 84.3% (124) visited our office, provided a saliva DNA sample, and received a $15 incentive payment. In terms of the 191 eligible students, 65.0% provided saliva samples. Outright refusals appeared rare in both cases; we are not aware of any individual who refused to complete the web survey, while one individual who had completed the survey refused to provide the saliva sample.

The BioSpecimen Processing Facilitating Center at the University has processed the DNA samples. The saliva DNA collection from all 124 participants was successful. The average yield of DNA is 144.5 micrograms/participant with a standard deviation of 125.4. The minimum yield is 15.6 and the maximum 684.8. Even the minimum yield is more than sufficient for all genotyping proposed in the main study.

The Genotyping Core at the University carried out the test genotyping in December 2006. We genotyped one SNP in the dopamine D2 receptor gene (*DRD2* TaqIA, dbSNP reference: rs1125394, LocusID: 1813) and a second SNP in the Catechol O-methyltransferase (COMT) val met SNP (dbSNP reference: rs4680; LocusID: 1312). The genotyping was carried out using the Applied Biosystems TaqMan® genotyping technology. The two SNPs were typed for 128 DNA samples including 4 blind duplicated controls. Five ng DNA was used for each SNP. Both SNPs were successfully detected and the call rate was 99.2% and 98.4% for rs1125394 and rs4680 respectively. Negative controls were tested as well and no signal was detected. In conclusion, the SNP genotyping test was successful and DNA quality was very good.

There are good reasons to believe that both the web and DNA response rates can be improved in our main study. The data collection phase of the pilot study was launched on October 4, 2006 and ended on November 21, 2006. The responses to our web survey request were generated by an initial email and five email reminders, the last of which was sent on November 3, 2006. The DNA collection visits to our on-campus office were also generated by email reminders alone, the last of which was sent on Sunday evening, November 19, 2006. Our DNA collection response rate was also negatively affected by the limited hours during which our office was open. For the large majority of the time, our office was only open 10:30-2:00 Monday through Thursday because of the budget constraint. The hours may not fit some of the students' schedules.

For our larger study we propose these plus additional efforts to boost response rates. In particular, we intend to supplement email messages with phone calls for the web survey. For the DNA saliva collection, in addition to increasing manned office hours, we intend to attempt to contact sample members with both phone calls and dorm visits to the participants. It appears that some students were reluctant to take the time to visit the central-campus office. Our dorm visits would be structured to avoid this problem by setting up DNA collection stations in the dorms on a rotating basis and then knocking on dorm doors to inform sample members of the nearby station.

The main study is funded by the William T. Grant Foundation and will go into the field in the spring of 2008. The main study will web-survey approximately 3,000 students and collect saliva DNA from 2,000 of them.

Conference Announcement: Mid-Winter Methodology Section Meeting

James Moody is hosting the mid-winter methodology section meeting this year.

Date: Saturday, Feb. 23rd
Place: Duke University
Sub-Themes: Networks, Methods and Global Health

The meeting will be a "general" methods meeting, but there will be some panels/sessions that focus on specific topics.

For more information please contact:
James Moody
Associate Professor of Sociology
Duke University
Email: jmoody77@soc.duke.edu

Conference Announcement: The Society for Research on Educational Effectiveness

The Society for Research on Educational Effectiveness
Promoting research important for education

First Annual Conference: Research on Educational Effectiveness – Pragmatic Decisions and Critical Outcomes

Sunday, March 2nd through Tuesday, March 4th, 2008
Hyatt Regency, Crystal City, Virginia

The first annual conference for SREE will feature research that examines the effects of educational interventions on the following important educational outcomes:

- ⇒ Reading, writing, and related language skills
- ⇒ Mathematics and science achievement
- ⇒ Social and behavioral competencies
- ⇒ Dropout prevention and school completion

Recognizing the need to further knowledge of research methodology, the conference will also include presentations focused on advances in research design and data analysis.

Keynote speakers include:

Thomas D. Cook
Professor of Sociology, Psychology, Education and Social Policy
Northwestern University

Judith M. Gueron
Independent Scholar in Residence
President Emerita
MDRC

Grover (Russ) Whitehurst
Director, Institute of Education Sciences
United States Department of Education

Conference registration is now available via the conference web site:
www.educationaleffectiveness.org/conferences/2008/

SREE Web Site: www.educationaleffectiveness.org

Member News: Guillermina Jasso

1. The paper that Guillermina Jasso presented at the Winter Meetings of the Methodology Section in 2006 has now been published. Co-authored with Samuel Kotz, it is titled, "A New Continuous Distribution and Two New Families of Distributions Based on the Exponential" and appeared in the August 2007 issue of *Statistica Neerlandica*.

2. Guillermina Jasso Recently worked with a large team of interdisciplinary collaborators—Richard Freeman, Gary Gereffi, Ben Rissing, and Vivek

Wadhwa—to develop a methodology for estimating the number of skilled immigrants who are in line for an employment-based visa, including those already in the United States on temporary visas. An initial report published by the Kauffman Foundation is available at SSRN: http://ssrn.com/abstract=1008366.

3. Congratulations to Guillermina Jasso he was elected a Fellow of the American Association for the Advancement of Science. Notification was in October 2007, and formal induction will be at the annual meetings of the AAAS in February 2008.

Member News: Billie Gastic

Congratulations to Billie Gastic she has been named a 2008 Hispanic Faculty Fellow by the Association of Hispanics in Higher Education.

Notes from the Editor

Early in my career, a colleague asked for help communicating with a journal editor. This editor had just rejected the colleague's paper based on an incompetent review of the statistical methods. I wrote the editor, explaining why the reviewer's criticisms were wrong, with citations to several widely used statistics texts. The editor responded to the effect that the journal received far more submissions than could be published, and understood that the author was disappointed in the outcome of the review process. Lesson learned. Like many of us, I came to accept decisions based on incompetent reviews as an unavoidable injustice. In this issue's lead article, section chair Rafe Stolzenberg suggests that we post or otherwise publicize instances of review incompetence. I hope that Council will consider this at the Section meeting this summer.

I've edited this newsletter for several years, and it's time for me to step down. My last issue will be Summer 2008, following the ASA meetings. Please consider volunteering for this responsibility. I will, of course, do what I can to smooth the transition. You can contact me about this and I'll pass it on to Rafe, or you can contact Rafe directly. Any questions should be directed to me. For the summer issue, please send section news, news about section members, and brief essays to me at L.raffalovich@albany.edu.