Rejoinder: Better Training, Deeper Thinking, and More Policing

Xiao-Li MENG

REPRESENTING AN ENGAGED POPULATION?

Perhaps due to the somewhat unusual nature of my piece—a discussion of Brown and Kass (2009) that ended up longer than the article itself—the Associate Editor (AE) who handled it had an unusual idea: inviting the general public to react to it. The AE's motivation is clear from his/her editorial comments:

> "My thought here is that we too often turn to the "usual suspects" to get reaction to such manuscripts, yet this is an issue that touches all of statistics and all statisticians. It would be interesting to get the perspective of the broader readership on the issues raised by Meng (and potentially Brown and Kass as well, if we wanted to open up that for general discussion). If no comment is elicited by the article, that perhaps says something as well (I'm not sure what)! I think that proceeding in this fashion would potentially open up a forum for a more wide-ranging discussion."

I was intrigued, and particularly liked the idea of testing what reactions (if any) would be generated without any targeted invitation, from a truly self-selected sample! As statisticians we worry deeply—and rightly—about biases in any self-selected samples, but here one could argue that the seven sets of discussants are a reasonable sample of the population of the "engaged participants," as Kotz characterized them. [If this characterization offends you ("I didn't have time to write because I was busy teaching!"), then you are in this population by definition!]

The AE's prediction of "a more wide-ranging discussion" is also accurate. Government, business, industry, and academia are represented by the discussants; so are North America, Europe, and Australia. The representation also contains deeper stratifications: two-year colleges and universities, nonprofit and for profit, on duty and retired, West Coast and East Coast, etc. Even the writing styles cover a whole spectrum, from humorous storytelling to almost a DoW (Declaration of War)! It is indeed quite remarkable that merely seven discussions can have such a broad and deep representation (but of course no claim on *proportional representation*)!

Any author should be grateful for such wide-ranging reactions, even if most of them are disagreements or criticisms (not the case here!). My heartfelt thanks also go to the AE and the Editor, John Stufken, for providing the forum. In addition, John needs to be thanked for gently reminding me not to repeat history by making my rejoinder longer than the discussions. This freed me from trying to have 95% coverage, but instead to focus on 10% tails in either direction. This rejoinder therefore contains mainly stories inspired by discussants' excellent questions, points that received too little attention in my original piece, and responses to discussions that I need to pour myself a glass over because I have been given too much or too little credit. The responses are organized along the three main themes of my piece as highlighted by the discussants.

BETTER TRAINING

Kotz's and Soler's discussions should make us appreciate more the AE's creative idea. We all have seen many discussions over the years about statistical education, from K-12 to Ph.D. programs. Whereas technically two-year colleges have been included in these discussions, some issues Kotz and Soler raised are completely new to me, and I suspect also to many of AE's "usual suspects." For example, the "bizarre scheduling" situation Soler reported is not something most (any?) of us in statistics departments have thought about, yet now I can see how frustrating, complicated, and serious the matter is. Kotz is right that we all should care more about what goes on in two-year colleges, because such issues directly affect our entire profession; I was quite taken aback by the sheer number of students taking statistics merely in Kotz's and Soler's colleges, a combined annual total over 7000! (Speaking of numbers, this year alone we have 15 undergraduate students declare statistics as their concentration (major), to answer a question of Kotz.)

I also cannot have said better than Kotz's two "blunt" statements about the responsibilities on our shoulders, which remind me of a story from a statistician who joined a large pharmaceutical company after years of being a professor. His first task was to analyze a set of pre-clinical data. He told me that the night before his presentation, which he was told would *determine* whether the company should launch an estimated 30-million-dollar clinical trial, he literally felt sick to his stomach: "I was really scared; I had never felt this much responsibility!" I echoed that I could easily imagine how I'd have felt if I had been in his shoes. Retrospectively, however, I have been asking myself: have any of us ever felt sick to our stomachs the night before teaching because of the thoughts about the responsibility of training future generations, which surely should be heavier than any 30-million-dollar study? Of course most of us have not (at least I have not) for a very simple reason: the impact/outcome of our teaching is not immediately tangible or even measurable—I am sure many of us would have if we were told that tomorrow's lecture would *determine* 30 students' career choices. But this very fact should also remind us of the immensity and longevity of our impact through teaching and hence increase our sense of great responsibility. Perhaps it is not inappropriate to further intensify our "pedagogical sensation" by borrowing a phrase about passionate love—especially given that effective teaching also requires passion and love: it can be felt years after the sunrise....

Passion-driven statistics is indeed the central theme of Easterling's humorous piece—I almost wanted to negotiate for a

Xiao-Li Meng is the Whipple V. N. Jones Professor and Chair of Statistics, Department of Statistics, Harvard University 1 Oxford Street, Cambridge, MA 02138 (E-mail: *meng@stat.harvard.edu*). I thank my colleague Joe Blitzstein and my happy team members (Yves Chretien, Paul Edlefsen, Kari Lock, Cassandra Wolos Pattanayak) for many insightful comments and suggestions.

© 2010 American Statistical Association DOI: 10.1198/tast.2010.10003

new car so I could show my dealer what I am made of (but I won't tell him my shoe size)! Easterling is entirely correct that the current generations have a chance (and responsibility) to bring the passion to a new high, and our collective effort can start as simply as better utilizing existing textbooks; as Fox puts it, let's "get it going" rather than "get it perfect." Cleary and Woolford brought in another starting point that was only alluded to in my piece: better training should start with better admissions/recruitment—training with passion from the outset, whenever possible, is obviously more effective than passion injected afterward. Their excellent point on not repeating what we hope others won't do (i.e., equating one to two courses with competence) suggested that the verb in "supplementing graduate curricula with Professional Development Curriculum" should eventually be replaced by "integrating." This will take time, but its ultimate reality is an important assurance for Fox's prediction: "our profession has an incredibly bright future."

Fox also asked an excellent question: what are the minimum standards and competencies for persons deemed suitable to teach statistics? In my original piece, I answered a much easier question: what are the ideal qualifications? Fox's question currently has no enforceable answer. And *that's the problem* (*illa difficultas est*?), to answer Fox's "*Quo Vadis or Quid Agis*?" Surely any minimum standards should include having taken X courses in statistics with $X > 0$, right? As Mark Twain (or Will Rogers) was alleged to have said: "You can't no more teach what you ain't learned than you can come from where you ain't been." Well, even $X = 1$ would disqualify a good number of Soler's colleagues. And that is only for a single two-year college. Thinking about all the two-year colleges, four-year colleges, and AP Statistics, my stomach is now indeed turning.... (I know I am generalizing from $n = 1$, but I have a strong prior!)

Easterling worried about students getting turned away by bad teaching, especially the sharp students. I share that concern, as I detailed in my response (Meng 2009, Part I; 2010, Part II) to Rossman et al. (2009) regarding my observations that some Harvard undergraduates had been turned off by poorly taught AP Statistics, a point on which Kotz also commented. I sincerely hope that Kotz is right that the unintended perception my "Harvard observations" might generate is indeed a misrepresentation, but as I argued in my two-part response, any scientific assessment of the real impact of an educational program needs to study both the "turned-on" population, as Rossman et al. (2009) and Kotz reported, and the "turned-off" population, as I encountered. My two-part responses therefore include a suggestion to ASA to conduct assessment studies, which can help to assess whether the "Harvard observations" are merely local anecdotes or an indication of something far more worrisome.

DEEPER THINKING

This is another point that generates no disagreement, although von Collani asked that "statistical thinking" be replaced by "stochastic thinking," a concept whose meaning I yet need to find out. All discussions below, therefore, still center on statistical thinking as I understand it.

Hoerl and Snee correctly emphasized that statistical thinking should be coupled tightly with statistical engineering, a notion that was not discussed in my article but was advocated by John Tukey (if anyone can locate a specific quote, please let me know). A key component of this coupling, as I see it, is *efficiency*, a critical element that I wish the ASQ's definition of *statistical thinking*, as Hoerl and Snee quoted, had recognized, in addition to *process*, *variation*, and *data*. The maturity of a scientific discipline is measured not only by its accumulated coverage but also—and arguably more critically—by its demonstrated ability to establish *limits*, that is, the optimality and impossibility given constraints. I therefore like Hoerl and Snee's repeated emphases on how statistical engineering—like any other engineering—is about how to "best utilize" concepts, principles, theory, methods, etc. It is this adverb *best* that separates professionals from amateurs, and it is the quest for doing the best given the practical constraints that requires deeper thinking. Most people do not need to take a course in experimental design in order to try out "one-factor-at-a-time" (unless, of course, you are Easterling's unfortunate Sandia colleague!). But to be able to design optimal or even just cost-saving experimental designs given a variety of real-life constraints requires far deeper understanding of the principles of statistical experiments and modeling than most people are naturally equipped with; this kind of ability can have high societal impact, but it can be acquired only via a good dosage of interweaving statistical thinking and statistical engineering, to echo Hoerl and Snee's key point.

A local example illustrates well the importance of understanding optimality/impossibility in defining one's professional identity and hence being desired. My CS (computer science) colleagues here have been teaching all sorts of wonderful algorithms and programming for computing least-squares solutions and alike. However, they found themselves unable to explain satisfactorily the statistical models and principles underlying these solutions, nor could they answer seemingly simple questions such as "Why take squares?" I was thus invited last year to provide a guest lecture to one of their introductory courses. The 90-minute lecture was fully packed, proceeding from Gauss and Galton to the meaning of statistical models to the concept and wonder of MLE. The punch-line that the least-squares estimator is the MLE under the normal model, something we statisticians all take for granted, was an eye opener to both the students and my CS colleagues. It is particularly intriguing to them that once the normal assumption is made (an assumption few of them ever questioned), "taking squares" is the best one can do—as one of them told me: "This is really cool—I've got to look into this MLE thing!" Perhaps the best indication that the lecture got CS students' attention was the course evaluation comment, "you guys teach CS really well, but you should really leave statistics to statisticians," as one of the course instructors relayed to me in the following semester.

If you are thinking that I am using this example to show off how statisticians think more deeply than computer scientists, then bear with me for the other half of the story. Because of this guest lecture, I sat through the one immediately preceding it. It was equally an eye opener to me! Just as statisticians are well-versed in the limits of inference and the like, it is my CS colleagues' cup of tea to tell what is possible and impossible with algorithms and programming, among others. The lecture taught me that it is impossible to have an algorithm/program that can debug every other program correctly. Whereas logically it might not be hard to suspect such an "almighty" algorithm cannot exist, what demonstrates well the deep thinking by computer scientists is their ability to identify problems that seemingly have no connection whatsoever but in fact are equivalent to the impossible debugging problem. And hence they can immediately tell any amateur, "don't even try!" just as we statisticians can tell CS students not to waste their time trying to beat MLE asymptotically.

Hoerl and Snee also asked about what approaches are taught in Harvard's Stat 399 about attacking deep, broad problems that require more than one technique to solve. As I mentioned in my piece, the course was a result of responding to students' request that we help them to better prepare for Ph.D. qualifying examinations. Over the years our qualifying examination format has changed considerably, but one theme has remained—the problems are *not* designed around a set of textbooks or courses; rather, they come out of faculty members' research project problems or problems that teach deep thinking in statistics, such as applying the principle of bias-variance trade-off to investigate what is possible and what is not possible. That is, the problems are often multipart "nano research projects," mimicking their real-life counterparts yet doable in an examination setting. Such examination formats provide a forum for an intensified dialog between students and faculty, before, during, and after the examination. See the report by Blitzstein and Meng (2010) for detailed examples and discussions of the usefulness of "nano research projects." It is also worth emphasizing that the ultimate goal of repeatedly using real-life problems, as in Stat 399 and Stat 105, is not just to showcase the ubiquity of statistics, but more importantly—as Hoerl and Snee also emphasized via the cited Bryce's course—to demonstrate how statistics operates as a scientific discipline with a set of core principles, theories, and methods that can be applied to address an exceedingly wide range of problems.

MORE POLICING

This point is more debated, as several discussants expressed concerns about whether the label "police" would carry a passive image that we only react when someone does something wrong. Retrospectively I wish I had chosen a term that would not conjure such an image, because the whole message of my piece is how we can be more active than reactive, moving from everyone's back yard to the front yard and even living room. Perhaps it is my Chinglish, not understanding well all the connotations of the term "police." When I wrote that I am proud to be labeled as a "statistical police," what I had in mind was "We serve and protect"—a slogan seen on every police car in Chicago (where I spent 10 years)—we provide service to others and we protect them from mistakes.

I of course agree with Hoerl and Snee's "good cop" role, which is similar to Fox's "embedding" approach to work from within. Again much of my piece is about how to provide better quantitative training for future generations for other disciplines, which aims at helping others to move faster on their endeavors in the first place. However, there is either an apparent contradiction or a troublesome implication in the following statement of Hoerl and Snee's: "Meng rightly points out that statisticians can play a useful role in society by limiting the claims made by other scientists based on faulty statistical studies. He refers to this function as playing the "statistical policeman" role. We call this playing the "bad cop" role, in that bad cops fundamentally slow down the research of other disciplines." I do not see how slowing down the research of other disciplines is "a useful role in society," nor how avoiding bad/faulty statistics would slow down research. Isn't the whole purpose of avoiding or stopping mistakes, statistical or otherwise, to speed up real research progress? If someone can move his/her research faster by using *bad/faulty* statistical studies, would that logically imply that the *good/sound* statistical studies are actually antiscientific?

Hoerl and Snee could not possibly have meant what their sentences appear to imply, just as my being proud as a "statistical police" could not possibly mean that I am proud of being a "bad cop," as Hoerl and Snee characterized it. I surmise that what Hoerl and Snee really had in mind was that we should avoid making others feel that we are only interested in criticizing them, not helping them. This message I certainly agree with. We should be strategic in delivering the "bad news" so that we *consult*, not *insult*; and this is where effective communication skill plays a critical role. But this does not mean that we should avoid our "policing role" (though I certainly want to avoid all the negative connotations of "policing" if a better term can be found!). As some readers may have noticed, I tend to put significantly more emphasis on things I believe to require encouragement than on things that already come with good incentives. As I discussed in my original piece, "policing" is typically a thankless and creditless job at the individual level. But at our professional level, I believe it is a part of our identity that will remain unique to us even if "other disciplines have been seizing opportunities" away from us, precisely because we carry out the role for our discipline's integrity, as a critical part of general scientific integrity. Putting it differently, if someone is able and willing to carry out this role on a routine basis, I will have no trouble in considering him/her my fellow statistician. And in that role we sometimes do need to stop someone, not in his/her research, but in the potential harm s/he can do to others and indeed to an entire field. (Incidentally, the article "Fatal Flaws in Cancer Research" in the most recent issue of *IMS Bulletin* (2010, January, page 5) demonstrates vividly how faulty statistics can do harm to our society and how good "policing/forensic" work can stop it.)

A good example is in the literature of climate change, where decades of efforts have been made to understand and interpret apparent oscillations in running correlations among different climate time series/indices, often with conclusions that they represent some fundamental underlying climate dynamics in Mother Nature. However, Gershunov, Schneider, and Barnett (2001) demonstrated via simple simulations that such oscillation phenomena exist even if the two time series are completely independent white noises! As a part of our statistical thinking (and here it is not even a very deep one), any reasonably trained statistician would be concerned with the potential artifacts introduced by the *overlapping moving windows* used for computing the running correlations in the first place. Decades of efforts have literally been misled, but to make the matter worse, when an entire field is on a wrong track for a long time, it often would take an even longer period to put it back on the right track. No one likes to be told that he/she has wasted his/her (professional) life, so the force to defend the established answer or to at least find ways to "save some face" is often very strong (see Robinson, de la Pena, and Kushnir 2008 for a brief summary of the history of this debate). Of course the truth always prevails (let's hope!) and strong arguments, even or especially the wrong ones, could help us to think more deeply. Nevertheless, the real scientific progress in such cases is clearly delayed, not because policing or self-policing duties were carried out too soon, but rather because they were carried out too late.

It is also worthwhile to emphasize that the impact of such "policing work" can go far beyond what is originally intended. For example, Gershunov, Schneider, and Barnett (2001) work has apparently also influenced researchers in solar physics, where running correlations and alike are used to measure certain solar activities. In particular, Elias and de Artigas (2008) provided a detailed account of how a "spurious quasi-biennial cycle" induced by running correlations may be similar to the reported QBO (quasi-biennial oscillations) of the stratospheric equatorial zonal winds, parallel to Gershunov, Schneider, and Barnett (2001) findings; also see Elias and de Artigas (2006). I am especially delighted to see that Elias and de Artigas (2008) was featured as the leading Expert Commentary in the book on *Solar Physics Research Trends* (ed. Wang, 2008), and its abstract ends with what I consider as a good example of self-policing: "The results shown here do not rule out a physical origin, but point out that a result obtained after a statistical analysis carries, in addition to the physics behind, the spurious byproducts of the method applied."

Compared to Hoerl and Snee, von Collani gave me too much credit. Von Collani labeled my article as a "milestone in the development of science," and credited me as a revolutionary in policing science. While flattered, I must confess that I am at least a mile away in seeing the pictures von Collani is painting, and I am not sure if I would make a half turn or full turn in my grave if someone puts "Chief of Science Police" on my tombstone. Von Collani apparently is questioning the entirety of modern science and statistics and wants to replace everything by "stochastic thinking," a discussion topic that is the furthest from my original piece, certainly beyond my reflection antenna. But my statistical thinking compels me to express skepticism of just about any claim of one size fitting all, especially when it comes to matters as complex and grand as science and statistics.

THE POWER OF COLLECTIVE WISDOM AND ACTION

The experience starting from reading Brown and Kass (2009) to preparing this rejoinder reminded me once more of the power of collective wisdom. No matter how thoughtful, articulate, and well-intentioned each of us is, our individual contributions are inevitably idiosyncratic and can even carry ironies that are obvious to everyone but ourselves. For example, I almost choked on the great wine I was enjoying with the discussions when I read Kotz's question "Why did Meng stop with scientists and policy makers?" Indeed! Although I probably would not go as far as to include Easterling's car dealers and shoe salesmen, how could I forget to include "whole generations of future teachers" in an article that is mostly about teaching future generations?

This brings me to my concluding point, the same as the one in my original piece, and on another historic occasion, the first days of the new decade. I share Fox's and others' enthusiasm and optimism that our future is very bright, but to ensure that such enthusiasm and optimism will be carried over to future generations, we need collective action. So please do anything you can to help build the "statistical leadership" that Hoerl and Snee articulated: one lecture at a time, one speech at a time, one consultation at a time, and one publication at a time.

And a toast to the new decade: may we all have that sick-to-our-stomach feeling at least once before lecturing/speaking/consulting/publishing!

[Received January 2010. Revised January 2010.]

REFERENCES

Blitzstein, J., and Meng, X.-L. (2010), "Nano-Project Qualifying Exams: An Intensified Dialogue Between Students and Faculty," technical report, Harvard Statistics. *The American Statistician*, to appear. [28]

Elias, A. G., and de Artigas, M. Z. (2006), "The Quasi-Decadal Modulation of Running Correlations Involving the QBO," *Journal of Atmospheric and Solar-Terrestrial Physics*, 68, 1980–1986. [29]

——— (2008), "The Quasi-Biennial Oscillation in Time Series of Solar Activity Parameters," in *Solar Physics Research Trends*, ed., P. Wang, Hauppauge, NY: Nova Science Publishers, pp. 3–13. [29]

Gershunov, A., Schneider, N., and Barnett, T. (2001), "Low Frequency Modulation of the ENSO-Indian Monsoon Rainfall Relationship: Signal or Noise," *Journal of Climate*, 14, 2486–2492. [28,29]

Meng, X.-L. (2009), "AP Statistics: Passion, Paradox, and Pressure (Part I)," *Amstat News*, December, 7–10. [27]

——— (2010), "AP Statistics: Passion, Paradox, and Pressure (Part II)," *Amstat News*, January, 5–9. [27]

Robinson, L. F., de la Pena, V. H., and Kushnir, Y. (2008), "Detecting Shifts in Correlation and Variability With Applications to ENSO-Monsoon Rainfall Relationships," *Theoretical and Applied Climatology*, 94, 215–224. [29]

Rossman, A., Peck, R., Franklin, C., Hartlaub, B., and Scheaffer, R. (2009), "Letter to Editor," *Amstat News*, November, 5. [27]