Rigor Without Rigor Mortis: The APS Board Discusses Research Integrity

February 01, 2012

 Please excuse this further sidetrack from the road we were on in my previous columns. Two months ago, the column I had planned was displaced by a response to the considerable attention that various media paid to a social psychologist's faking of data and the attendant questions about whether psychology was especially susceptible to cheating. The implication seemed to be that many, if not most, of the most striking results in psychology might be bogus.[i] I argued in my previous column that there is nothing special about psychology when it comes to fraud, that meta-analyses suggest that fraud is rare (about 2% of researchers admit to it), and that tools intrinsic to the practices of science, such as replication, help root out "false positives," or Type I errors (concluding that some effect is present when in fact it is not), and produce a science we can believe in.

But can we do better, at least in the sense of encouraging practices that allow science to function more efficiently and effectively? The APS Board took up this and related questions at our retreat in early December. The discussion was animated and (in my opinion) very productive, so we decided that this "Boardologue" should be shared with the broader APS community. Here's the plan we came up with: I would write a column outlining a few of the issues and possible recommendations and then we would begin spilling ideas over to the APS website by having board members share their perspectives. The third step intended is an open forum, refereed for relevance, redundancy and respect for our community. So here goes step one.[ii]

More than two decades ago, I was one of the people invited to help celebrate the 25th anniversary of the University of Minnesota Center for Research in Learning, Perception and Cognition. The speakers were invited to speculate on how the field of learning might or might not change 25 years in the future. The only thing I remember about my own talk was the tongue-in-cheek prediction that in 25 years counter-

balancing would still be a good idea. The audience laughed (probably politely), but later on, a graduate student from my lab, David Thau, told me that after the laughter died down, the graduate student next to him turned and asked, “What’s counter-balancing?”

Well, I still think counter-balancing to control for order effects is a good idea and should be used when the study design permits it. Furthermore, the fact that it may be inconvenient to do so doesn’t strike me as a good excuse for not counter-balancing. Yes, you may have to cut and paste parts of your questionnaire six times when it seems like one order would do, but I think it’s worth it. First, if you find no order effects, you’re on your way to a more robust pattern of results. Second, if you do find order effects, you may open a new line of inquiry, perhaps regarding some sort of priming effect. If you don’t counter-balance but obtain statistically significant results anyway, you won’t know whether you have lucked into the one question order that can produce the result of interest. So the issue is less about “false positives” than it is about a false sense of security surrounding the generality of the results and your interpretation of them.

Let me now turn to other suggestions from my wish list.

1) Counter-balancing (see above).

2) More on methods and procedures. At a time when journals seem to be pushing for streamlined everything, including methods sections, there is a danger that potentially relevant procedural details will be missing. If we know (or think we know) that a messy versus neat experiment room or the presence of an American flag can affect participant’s performance, it seems odd to skimp on details just because the factors are not of current interest. There are tons of studies on priming effects, but we seem to be unperturbed about writing that experimental probes were part of a larger set of tasks that (we assume) are not relevant to present concerns. Given that supplemental materials can be placed online, why not insist on providing the details and letting the entire scientific community judge their relevance?

3) In an earlier column, I suggested that attention to experimenter expectancy effects seems to have fallen out of fashion. Why not require that authors report whether or not the experimenter was blind to the hypotheses? [iii]

4) As<\span> was noted last month, Barbara Spellman, the editor of *Perspectives on Psychological Science*, and others are working to develop an archive of attempts to replicate experimental phenomena.[iv] Why not require authors, again in supplementary materials, to describe any related studies they have conducted for the same hypothesis but have chosen not to publish? [v] (I would make an exception for studies that have blatantly flawed designs.)

5) Another rule with lots of exceptions [vi] might be to include the actual data in supplementary materials. Some journals, such as *Judgment and Decision Making*, already have this rule.

Well, I’m going to stop here because I don’t want to consciously or unconsciously plagiarize other board members. My tentative bottom line is that we could add a touch more rigor to our empirical efforts and that it may be feasible to do so by some slight shifts in publication policies.

But we don’t want rigor mortis.

Some well-established areas of research may be like Phase III clinical trials, in which the methods and measures are settled issues and the only concern is with assessing effect size. Other areas, however, may rely on open-ended tasks in which the dependent variable cannot and typically should not be specified in advance. For example, to analyze people's sortings of (pictures of) different species only in terms of taxonomic relationships would leave researchers blind to alternative organizational schemes (such as sorting according to the habitats where species are found). In her dissertation studies, my former student Sara Unsworth [vii] got a great deal of mileage out of asking rural Wisconsin Native American and European American adults to tell her about "their last encounter with deer."

This sort of work raises different challenges with respect to rigor, as typically it just isn't feasible to specify a coding scheme in advance. I'm not sure what we know about the science of developing coding schemes, and our standards for establishing inter-rater reliability, in my opinion, remain underdeveloped.[viii]

I guess this is all part of what makes our field so exciting. We have a large advantage over other sciences in that our focus on human cognition and behavior naturally includes researchers and the psychology of their practices. We are intrinsically part of that which we study, and that is why rigor without rigor mortis not only advances our science but is part of it as well.

All of the Board members participated in the December discussion. Here are representative comments from a few:

Popularity Shouldn't Define Scientific Significance

 1) Recently, there has been a premium on “innovation,” “transformation,” and “paradigm-changing” research. This is important, of course, but it overlooks the importance of “normal” science, in the Kuhnian sense. Grant applications are now not being funded, merely because they are incremental. Not everything has to be paradigm shifting to be valuable.

2) There seems to be a blurring of boundaries between popular and scientific impact. Until recently, most scientists did not care whether or not their work was communicated to the public. This was a problem of course, but now the pendulum seems to have swung in the opposite direction: Sometimes it appears as if we care too much, and the science suffers for it.

Scientists now have competing goals. One is to publish work that is newsworthy (e.g., to be mentioned in the *New York Times* science section). A second is to publish work that is theoretically important and makes a significant contribution to the scientific question at hand. These are not necessarily the same, and so should not be confused. But they often are. Findings in papers are often hyped in a way that is more appropriate in a press release than in a scientific paper. Students now cite popular books (which are, at best, a secondary source) as evidence of some finding or effect , instead of citing the scientific papers. Often papers are triaged (in *Science* for sure, and some even claim this is happening in *Psychological Science*) because they are not newsworthy or splashy even though they are quite scientifically important.

Often, when we try to communicate things to the public (e.g., calling freezing behavior “fear” and calling the acquisition of freezing to a tone via classical conditioning “fear learning”), this filters back into the science itself in a way that is not helpful (e.g., the belief that “fear” has a unified biological cause).

3) The public still does not have a good grounding in the value of science and science education. Hence, they believe that there should be applied value in research that delivers right away. They often don’t understand that a *theory* is not a speculation or a hypothesis — it is a scientific explanation that is well established with data — or they confuse an *effect* with a *theory*.

4) Many psychology students no longer receive education in philosophy of science, and this limits the scope and validity of their theory-building attempts.

– Lisa Feldman Barrett

Continue the Conversation–Click Here to Make A Comment

Technology Could Help

 In the interest of encouraging replication and promoting transparency in evaluating methods, I suggest that each published paper include a video of the experimental protocol (faithfully reproducing the context, stimuli, spatial layout, experimenter intonation, gaze, pacing, feedback, etc.). This would essentially serve the purpose of what current methods sections are intended to do (permit others to replicate one's research) but would use current technology to capture much more detail and nuance than is possible with a brief verbal description. This small step would potentially have several benefits: (a) replication attempts would be more uniform, and the effects of slight procedural variations would be easier to measure; (b) methodological flaws in items or procedure would be more apparent; (c) unconscious cuing of participants may be detectable; and (d) researchers may be encouraged to be more accountable in ensuring that procedural details are thoughtfully considered in the design phase of the research and uniformly followed during data collection. There are serious issues to be addressed regarding how to maintain realistic fidelity without introducing IRB concerns re confidentiality, but I think these issues are solvable.

– Susan A. Gelman

Continue the Conversation–Click Here to Make A Comment

Universal Rules Could Be Problematic

 I’m all in favor of rigor and view my own work as high on the appropriate scales, whatever they may be. That said, I think that attempts to capture best practices by a set of rules are almost certainly doomed to fail, given the diverse nature of psychological science. Psychophysical experiments, for example, have been published with an *N* on the order of 2, possibly with only the authors (who obviously know the hypotheses) being willing to undertake the tedious hours of data collection with a repetitive task. That may not be the norm, but it illustrates why restrictions shouldn’t be expected to apply universally. My own work often uses instruments that can measure the positions and forces people exert over time, with the possibility of dependent variables exploding accordingly. If I discover that a variable affects jerk (the third derivative of position) rather than acceleration (the second derivative), am I prohibited from publishing?

-Roberta L. Klatzky

Continue the Conversation–Click Here to Make A Comment

Impact Factors Have Too Much Influence

 There are three main criteria by which we judge scientific work: rigor, importance in the sense that it makes a significant empirical and theoretical contribution, and general interest. It is right to focus on the first of these criteria because it essentially is the only one to which a set of rules or procedures can be applied — but it is the one that causes the least trouble. Fraud or failures to replicate do not arise because the studies were lacking in rigor, at least not insofar as a panel of experts could judge. Many of the suggestions regarding practices that would facilitate judgment of

scientific rigor are good ones, such as publishing raw data (though we already have a system in place that requires us to make raw data available on request). However, allocating journal space or cyberspace to indicate failures to replicate adds noise to a system (how are we to distinguish poorly executed studies from proper ones?), and requiring a statement from authors as to whether the successful study was accompanied by many nonsuccessful ones would seem to invite evasion, if not mendacity.

The more difficult problem concerns the other two criteria, since there is a strong subjective element to both. In order to deal with this subjectivity, the scientific community has tried to introduce a measure of objectivity. Citations and their derivatives, such as the *h*-index and impact factors, have assumed a measure of importance out of all proportion to their usefulness, so that rather than merely taking the pulse of scientific discoveries, they are used to prescribe a scientific regimen.

It is easy to see how we've arrived at this state of affairs. Citations, which are meaningful indices only *after* an article has been published, have been subverted to determine the fate of an article *before* publication. Here's how it works. Journal editors and publishers used citation counts as a way of determining the impact an article published in a given journal has on the field and derived impact factors based on that. Once this was in place, articles were judged not only on their own merit, but on the impact factor of the journal in which the article was published. Because of competition among journals to keep impact factor high, articles came to be judged not only on the basis of the first two criteria — rigor and importance – but also on the basis of the third — general interest, which has little scientific merit aside from drawing public attention to the article. As an analogy, consider a criminal trial in which the jury is instructed to take into account the effect their verdict would have on public opinion before rendering a decision. This mind set is reflected in a journal style in which all or part of the Method section, where rigor is judged, is relegated to the back of the paper and, more recently, to a supplementary section that is available only online and for which a separate search is required. In addition, to entice high-impact scientists to contribute to high-impact publications, reviews had to be rapid and turn around short, both militating against careful scrutiny of the publication. We quickly went from using citations as an imperfect measure of a paper's impact to having them determine ahead of time what kinds of papers will be published.

Most scientists can tell which way the wind blows, and if some are obtuse, tenure committees, granting agencies, and government ministries will make sure their senses are sharpened. Promotion and funding to individuals, departments, and universities (see the example in the UK and France) is based increasingly on these measures. Knowing that they are judged by these "objective" measures, many scientists, myself included, have succumbed to the lure of publishing short, eye-catching papers that will get them into high-impact journals, rather than submitting a paper with an extended series of experiments. We have seen this trend in our own flagship journal, *Psychological Science*. Our boast of having over 2,000 submissions a year reflects not only the quality of the journal, which is high, but also the fact that its impact is high and its articles are short. One or two experiments, rather than a series of them, will get you in.

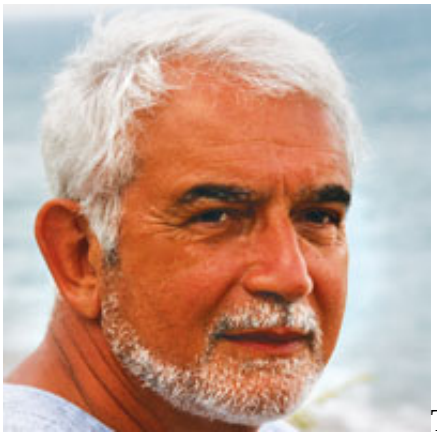
When I was a post-doc, an eminent psychologist who sat on the scientific review panel of the Canadian equivalent of NIH told me that in the 1950s and 1960s, a publication in *Science* or *Nature* was given no more credit than a book chapter and far less than a publication in a specialty, archival journal. The reason was that it was difficult even then to know on what basis the article was accepted for publication in *Science* or *Nature*, and given how short it was, it was difficult to judge the rigor of its methods. I

doubt we can return to that time, but we can downplay the importance we attach, not to citations, because they occur after the fact, but to journal impact factors. To increase rigor, we can return to requiring a series of experiments on a topic before we accept it for publication, even in a journal like *Psychological Science*.

– Morris Moscovitch

Continue the Conversation–Click Here to Make A Comment

We Need to Work on the Bigger Questions



The majestic production of papers based on fictive data produced by someone who was assumed to be a very respectable member of our community and published in very "respectable" journals has been a major source of reflection.

I shall take this opportunity to draw attention to an issue that provides a possible account for the undetected flourish of the extraordinary event that came to light. It is the theoretical as well as "phenomenal" permissibility that our science and some of our prestigious publication outlets encourage. The absence of a true paradigm in the Kuhnian sense, the absence of truly integrative theory, the absence of a problem that requires collective attention and research is undoubtedly one of the contributory factors allowing this type of misconduct to pass undetected for such a long time. The fractioning of the quest for knowledge to sound bites is becoming the criterion by which quality and significance are being judged, and our graduate programs are becoming increasingly sophisticated in training the next generation with these goals in mind. This means that we have to reflect and work upon the bigger questions that capture the imagination of many competing for the answer for the answer's sake. This means that we have to train the next generation to identify big questions, teach them to separate the big ones from the seduction of sound bites, and to learn to work in teams.

The recent revelation of misconduct, the full magnitude of which we shall only hear closer to spring of this year, is also diagnostic of what we value and why we confer high accolades in our profession, since the culprit in question had accumulated all possible honors in his field of practice and beyond. The shift from the individual to the team, a process that is in the making, will also contribute to a rethinking of the distribution of rewards as well as of the administrative and organizational structures we have to adopt in order to bring about these changes that are essential for our science to progress and reduce the hiccups we occasionally experience.

– Gün R. Semin

Continue the Conversation–Click Here to Make A Comment

Replication Will Expose Cheaters

 I believe three points should be considered in this discussion:

1) Cheating and scientific misconduct sadly happen in all fields of science and take many form,s from the outright forging of data to not reporting all of the data that have been collected. Psychological science is not different in this regard, and we need to come to terms with the fact that there are dishonest people in our field.

2) Replication, a distinguishing feature of science, ultimately ferrets out cheaters — it just takes time. While it is important that we take steps as a field when possible to prevent scientific fraud, it happens, perhaps by the way data are handled and reported. I hope the field does not substitute regulation for replication in its attempt to legislate this bad behavior. Replication remains our chief tool for eventually exposing cheaters.

3) The overwhelming majority of scientists in our field are honest and diligent, and these honest people are our ultimate tool for countering cheating — they sense when something isn't right, and as long as our institutions maintain an open and non-intimidating atmosphere, our honest colleagues will expose the cheaters. This happened in the case that triggered this discussion.

– Joseph E. Steinmetz

Continue the Conversation–Click Here to Make A Comment

Footnotes

[i] Given that our field is an empirical science, I'll just note that this (dubious) claim can be tested.
Return to Text

[ii] The recommendations listed at the end of the recent Simmons, Nelson, and Simonsohn (2011) paper also constitute good material for discussion. For example, they suggest that authors should be required to decide the rule for terminating data collection before data collection begins and report the rule in their article. I can see the value of this principle in certain areas of research, but it may not be so practical in other ones. For example, in the cultural research conducted in my lab, our informal rule is something like "let's run a few pilot participants to see how variable the data are going to be and then interview enough informants so that we can detect fairly large differences." Return to Text

[iii] Of course, there are many situations where blindness or double blindness is not feasible. My aim is just to increase the practice when it can be done. Return to Text

[iv] I should have added that Harold Pashler and Barbara Spellman are collaborating in this effort, coordinating what started out as two independent projects. Return to Text

[v] A postdoctoral fellow in my lab, Sonya Sachdeva, told me about attending a talk where at some point the speaker mentioned that "it took me ten studies to finally produce this effect." Return to Text

[vi] A case in point involves rich data sets (e.g., video observations) that might be analyzed in multiple ways for different purposes or to ask different questions. Here, authors should probably be given some reasonable amount of time to explore their own data before making them publicly available. Return to Text

[vii] Sara is now an Assistant Professor at San Diego State University. Return to Text

[viii] For example, "acceptable reliability" standards strike me as a bit arbitrary. I wonder, for example, if some variation on signal detection theory might be applied to adjust for inter-rater differences in criteria for saying some code is present. Return to Text