Special Issue:
Bayesian Probability and Statistics in Management Research

Journal of Management
Vol. 41 No. 2, February 2015 632–643
DOI: 10.1177/0149206314525208

Editorial Commentary

The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective

Andrew Gelman
Columbia University

We connect the replication crisis in social science to the default model of constant effects coupled with the flawed statistical approach of null hypothesis significance testing and the related problems arising from the default model of constant treatment effects. We argue that Bayesian modeling of interactions could lead to a general improvement in the communication and understanding of research results. Moving to Bayesian methods (or, more generally, multilevel approaches that incorporate external information) offers an opportunity for introspection into how social scientists interact with social phenomena to produce knowledge.

Keywords: *hierarchical modeling; multilevel modeling; null hypothesis significance testing;* p *values; regression analysis; variation*

Living With Variation and Uncertainty

For the past few years, there has been a growing crisis in social science, and in biology, medicine, and other statistics-dependent fields, that many claimed research findings are fragile, are unreliable, cannot be replicated, and do not generalize outside of the lab to real-world settings (Ioannidis, 2005). Arguably the crisis is most pronounced within psychology (see Pashler & Wagenmakers, 2012). While some view this crisis as being due to factors such as

Acknowledgments: Editors Michael Zyphur and Fred Oswald provided so many suggestions that they should be considered as coauthors of this paper. We also thank three anonymous reviewers for helpful comments and the National Science Foundation for partial support of this work.

Corresponding author: Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University, New York, USA.

E-mail: gelman@stat.columbia.edu

publication bias (see Francis, 2012, 2013). We see many of these problems as the result of applying 20th-century statistical models in a 21st-century research environment. As discussed below, we recommend moving beyond the worldview in which effects are constant and unvarying in their essentials.

An essential part of science is that people can come to agreement about new knowledge. But this is threatened by a torrent of published studies of varying quality. To simplify somewhat, it is natural for people (including scientific researchers, including ourselves) to feel that the world should be real with clear objective properties (indeed, even children have been found to have an "essentialist" view of the world as having deep structure despite surface appearances; see S. Gelman, 2005), and it is also intuitively appealing for us to think these properties can be studied in a deterministic fashion. When previously trusted studies do not replicate, this can be upsetting: We feel that science should give us facts because that is how we want it to work, and we are unsatisfied when the world does not conform to our obsession with using modern tools for dicing up the world into individualized knowledge-objects (or, as we say, "stylized facts").

Beyond all this, we as social scientists have difficulty following scientific practice as generalized from the world of classical physics, for (at least) two reasons:

- Psychological and social processes show much more variability than the usual phenomena in the physical sciences. The strength of an iron bar might be calculable to a high degree of accuracy from a formula derived from first principles, but there will not be anything like the same precision when predicting prices, political opinions, the contact patterns of friends, and so on.
- Even to the extent that patterns can be discovered from social data (for example, the Yerkes-Dodson curve in psychology, the Phillips curve in macroeconomics, the seats-votes curve in politics, and various predictive models of consumer behavior), these patterns can and do change over time, and they can look different in different countries and for different groups of people. Social science even at its best is contingent in the sense that physical models are not.

These concerns relate to the problem of nonreplication of social science research. Instead of focusing on ways to procure the certainty often attributed to the hard sciences, perhaps we as social scientists can address the nonreplication problem by changing what we want to get out of our research, accepting that we can gain knowledge without the certainty we might like, either for individual predictions or for the larger patterns we seek to discover. If effects are different in different places and at different times, then episodes of nonreplication are inevitable, even for very well-founded results.

Once we realize that effects are contextually bound, a next step is to study how they vary. In policy analysis, this implies different effects for different people under different scenarios (Dehejia, 2005), and it connects to goals such as personalized medicine (or, less pleasantly, individually targeted marketing pitches). From the perspective of statistical inference, however, fitting models of varying treatment effects presents technical challenges, as interactions are typically estimated with less accuracy than main effects.

Bayesian methods are relevant in this effort for several reasons. Most directly, Bayesian analysis uses the prior distribution, allowing the direct combination of different sources of information, which is particularly relevant for inferences about parameters that are not well estimated using data from a single study alone. Bayes plays well with uncertainty: The posterior distribution represents uncertainty about any set of unknowns in a model. And, in a

setting where the appropriateness of the model is itself uncertain, it is useful that Bayesian models are "generative"; that is, one can simulate replicated data sets from the model and compare them to the observed data (Gelman et al., 2013). That said, information has to come from somewhere: In a small-sample-size study, a Bayesian analysis can require strong prior information if it is to go beyond a simple recognition that there are high levels of uncertainty. When prior information is included, we think it should be represented as a continuous distribution on the magnitude of effects and interactions, not as the prior probability that an effect is "true" or nonzero or positive. Given the variation inherent in all human actions and measurements, it will not in general make sense to consider effects as exactly zero, and our concern is not the "false discovery" of zero effects. Rather, we are bothered by claims of large effects that do not generalize to other subpopulations of interest.

Before further discussing Bayesian methods, we first consider general concerns about unreplicable research.

Unreliable Findings

We can roughly divide the problems of nonreplicability into three categories:

1. *Fringe research* that follows scientific practices but is not generally accepted outside its subfield, for example, the attempts of Bem (2011) to prove that he found ESP and the claims of Kanazawa (2005, 2007) that nurses are more likely to have daughters, big strong men are more likely to have sons, and so on. Undoubtedly *some* of these sorts of taboo-busting efforts will turn out to be correct, but we can consider them as fringe research because they fall outside the usual bounds of scientific communication. The individual research articles look real enough (indeed, Bem's article was, notoriously, published in a top psychology journal), but the research programs run in parallel with, rather than as part of, mainstream science: As sociologist Jeremy Freese (2007) writes, these hypotheses are "perhaps more 'vampirical' than 'empirical'—unable to be killed by mere evidence" (p. 162). The nature of statistical analysis is that, by looking at data, it is possible to find support for some version of just about any research hypothesis. This does not mean that these hypotheses are false, but it does cast a shadow on published statistical claims. Another way of putting this is that, if the published claims in any of these subfields were to be taken literally, they would imply effects that are highly variable in sign and magnitude and appear in some settings and not in others. That is, the usual scientific paradigm of consistent, repeatable effects does not seem to apply in these settings, which (to take a charitable view) one could explain by a model in which true effects are small and highly variable.
2. *Outright fraud, scientific misconduct, and simple mistakes*, for example, Diederik Stapel's fake data, Mark Hauser's questionable data interpretation, and the data misalignment that Anderson and Ones notoriously refused to retract (see Goldberg, Lee, and Ashton, 2008a, 2008b, for the story and Anderson and Ones, 2008, for the nonretraction). The concern here is that most published work never gets checked. Add to this the pressure to publish in high-prestige journals, and you get a motivation to publish false or sloppy claims. We have published mistakes ourselves (Gelman, 2013a), not from any pressure but because we cannot think to check everything. (We found that particular error only in the process of trying to replicate and extend our own previous study.)
3. *Systematic biases* arising from implicit or explicit statistical reporting rules. Most notoriously, multiple analyses can be applied on the same data set, and successes are more likely than failures to be published. As explained by Ferguson and Heene (2012), selection of statistically

> significant results is more serious a problem than was formerly believed based on earlier, simplified models of the file drawer effect. Meanwhile, the "statistical significance filter" results in the magnitudes of published effects or comparisons being larger, on average, than the magnitudes of true averages or population differences (Gelman & Weakliem, 2009). But, as we have been learning from the research of Simmons, Nelson, and Simonsohn (2011); Francis (2013); and others, the problem is much worse than that. It is not just a matter of researchers reporting their best results; rather, it seems that you can pretty much always get statistical significance if you look hard enough, and indeed this can happen even without researchers realizing this at all (Gelman & Loken, 2013).

One usually talks about only one of the above three issues at a time, but they are all related. There is a growing recognition of a large gray zone in the interpretation of legitimate scientific work, and while this provides an opening for questionable research or outright fakery to slip in, there are more systematic factors that help produce the crisis. In addition, these concerns arise in mainstream research as well, in subfields, such as brain imaging, where the central ideas are generally accepted and misconduct is not generally an issue, but large problems still remain with unreplicated research and specific claims that are implausible (Vul, Harris, Winkielman, & Pashler, 2009).

Moving From Null Hypothesis Significance Testing to Interaction Models

My topic of discussion here is how this crisis in the *practice of science* is connected to *statistical* problems in the modeling of interactions. It goes like this: The two usual foundations of statistical modeling in the presence of uncertainty are (a) the null hypothesis and (b) the model of constant treatment effects. It is relative to the null hypothesis that we get all those statistically significant findings that cause so much trouble, and it is via the model of constant effects that we end up with all those precise (but often wrong) confidence intervals.

What is the problem? We can say it in a number of different ways:

- A *statistical* hypothesis (for example, $\theta = 0$ or $\theta_1 = \theta_2 = \ldots = \theta_k$) is much more specific than a *scientific* hypothesis (for example, that a certain comparison averages to zero in the population, or that any net effects are too small to be detected). A rejection of the former does not necessarily tell you anything useful about the latter, because violation of technical assumptions of the statistical model can lead to high probability of rejection of the null hypothesis even in the absence of any real effect.
- There is always measurement error. Suppose, for example, you find a 53% success rate in an experiment where pure chance would yield 50% with a standard error of 1%. This is three standard deviations away from zero, enough to reject the statistical null hypothesis but, in many social science settings, not enough to reject a scientific null hypothesis that would allow for systematic measurement error. A measurement error of 3% is small enough to come from various uninteresting sources. As Edwards and Berry (2010) note in the context of management research, methodological advances can simply make it easier for researchers to uncover statistically significant but scientifically unimportant effects.
- Treatment effects vary. To put it another way, all models have "error terms." The same measurement taken on two people will give two different results. The same measurement taken on one person is likely to give different results if done at two different times or under two different

conditions. Consider the notorious study that found that people are more likely to "think outside the box" when you put people outside of a physical box (Leung et al., 2012). There is little doubt that where you put someone will have an effect on how he or she thinks, but there is also every reason to suspect that such an effect is situational: It will differ (in both magnitude and sign) under different conditions and for different people.

- The very existence of an error term in models represents a tacit acceptance of varying effects. Yet, when we as researchers test our hypothesis, we compare to the silly comparison point of zero effects. And when we fit our models, we fit a constant effect (that is, a single regression coefficient, or the equivalent as formulated as the inverse of a family of hypothesis tests). This is bad news because the model excludes the possibility of varying effects, and thus any attempt to consider generalization beyond the population at hand must be done outside the model that was used to fit the data.
- To put it another way, once we accept that treatment effects vary, we move away from the goal of establishing a general scientific truth from a small experiment, and we move toward modeling variation (what Rubin, 1989, calls response surfaces in meta-analysis), situation-dependent traits (as discussed in psychology by Mischel, 1968), and dynamic relations (Emirbayer, 1997). We move away from is-it-there-or-is-it-not-there to a more helpful, contextually informed perspective.

The presumption of constant effects corresponds to a simplified view of the world that can impede research discussion. To do helpful science with transportable findings will require grappling with the world's complexity. Unfortunately, a constant-effects worldview makes that hard to obtain. So, the problem is more complex than simply changing methods. It requires changing mindsets.

As an example, consider the continuing controversy regarding the "hot hand" in basketball. Ever since the celebrated study of Gilovich, Vallone, and Tversky (1985) found no evidence of serial correlation in the successive shots of college and professional basketball players, people have been combing sports statistics to discover in what settings, if any, the hot hand might appear. Yaari (2012) points to some studies that have found time dependence in basketball, baseball, volleyball, and bowling, and this is sometimes presented as a debate: Does the hot hand exist or not?

A better framing is to start from the position that the effects are certainly not zero. Athletes are not machines, and anything that can affect their expectations (for example, success in previous tries) should affect their performance—one way or another. To put it another way, there is little debate that a "cold hand" can exist: It is no surprise that a player will be less successful if he or she is sick, or injured, or playing against excellent defense. Occasional periods of poor performance will manifest themselves as a small positive time correlation when data are aggregated.

However, the effects that have been seen are small, on the order of 2 percentage points (for example, the probability of a success in some sports task might be 45% if a player is "hot" and 43% otherwise). These small average differences exist amid a huge amount of variation, not just among players but also across different scenarios for a particular player. Sometimes if you succeed, you will stay relaxed and focused; other times you can succeed and get overconfident.

Whatever the latest results on particular sports, we cannot see anyone overturning the basic finding of Gilovich et al. (1985) that players and spectators alike will *perceive* the hot hand even when it does not exist and dramatically *overestimate* the magnitude and

consistency of any hot-hand phenomenon that does exist. In short, this is yet another problem where much is lost by going down the standard route of null hypothesis testing. Better to start with the admission of variation in the effect and go from there.

Where to Go Next?

The problems of null hypothesis significance testing in social research are well known; see, for example, the review by Krantz (1999). The question is, what is gained by moving to other statistical methods? As Greenland and Poole (2013a, 2013b) point out, two frequently proposed alternatives—confidence intervals and noninformative Bayesian analysis—can be viewed as mere re-expressions of *p*-value information in different forms, and the resulting inferences can be problematic if data are weak (Gelman, 2013b).

We argue that the key problem with significance testing is not with the *p* value as data summary but rather with the underlying model of constant effects. There is a long literature on varying treatment effects, with statisticians and applied researchers rebelling in frustration against the default model (Berrington & Cox, 2007; Bloom, Raudenbush, & Weiss, 2011; Bryk & Raudenbush, 1988).

The challenge in fitting interaction models is that they require additional data (or prior information). Consider the following quick calculation: The simplest interaction is the difference between effects in two equally sized groups with equal within-group variation. The standard error of estimation of this difference is a factor of 2 higher than the standard error of the main effect (with a factor of the square root of 2 coming from the half-sized sample within each group, and another square root of 2 coming from the differencing). If the study is powered so that the main effect is likely to be just barely statistically significant, and if the true size of the interaction is somewhat smaller than the main effect, this suggests that it will be difficult to estimate interactions to the level of certainty typically demanded in applied research.

One thus typically has the uncomfortable choice between a no-interaction model that is misleading or an interaction model that cannot be estimated. A Bayesian approach may provide a resolution. Here are some ideas:

- *Prior information on the parameter of interest*. There is a fair amount of prior information that is essentially generically available in causal or predictive models. In logistic regression with binary or standardized predictors, it is rare to see coefficients that are much greater than 2, and a cross-validation study has found improved predictive performance arising from a gentle regularization by assigning Cauchy(0,1) prior distributions to the coefficients (Gelman, Jakulin, Pittau, & Su, 2008). Including this basic and uncontroversial knowledge serves two useful functions: First, in sparse-data settings, these priors stabilize coefficient estimates, keeping them away from unreasonably high levels; second, the resulting "regularized" model fit gives, on average, better predictions for new cases.
- *Multilevel modeling and regularization with large numbers of coefficients*. There is a range of problems that can be expressed as regression models with many predictors. These problems include pure prediction problems (which are often placed in the category of "machine learning"), adjustments in causal inference (for example, propensity score modeling), small-area estimation (used, for example, in political science and public health to map opinions or risk factors across states or counties), and models of contextual effects (Raudenbush & Bryk, 1986). In all these settings, multilevel regression allows the fitting of models in a stable way, even as

the number of coefficients gets large. Such ideas have become popular in non-Bayesian frameworks as well (Tibshirani, 1996). The practical result is that one can include more variables, making models more realistic, with less worry of overfitting.

- *Bayesian interaction models*. Paradoxically, Bayesian modeling can be more difficult in simpler problems with fewer predictors. When researchers are performing inference for 64 demographic categories within 50 states, their models have many coefficients and they can easily fit hierarchical regressions. But in a simple before-after study with two treatment levels, the simplest interaction model has only four coefficients (main effect, treatment effect, coefficient for "before," and the interaction), and the usual hierarchical model that assumes effects are exchangeable (that is, come from a common probability distribution) will not make sense. A further challenge is that a treatment can be reasonably expected to alter the variance as well as the mean, leading to a set of models involving an additional variance component that is only weakly identified (Gelman, 2004). In our current understanding, fitting such a model requires a large amount of subject-matter information (see Gelman & Huang, 2008, for an example). A payoff of interaction models is that the results can map more directly to the ultimate decisions of interest, which typically involve not just deciding among some set of options but also deciding which options to apply in which settings.

One challenge here is to distinguish large variation in a posterior distribution due to reliable variability in effects across scenarios, as compared to large uncertainty due to noninformativeness of data. In a hierarchical Bayesian analysis, high uncertainty about the level of variation should express itself in a high posterior variance for the variance parameters themselves, which indicates a need for new data, or possibly the incorporation of old data into the model via a meta-analysis.

When one is considering outcomes under replications, Bayesian inference serves some of the roles played by classical power analysis in providing a way to use subject-matter knowledge. But we think it is important for this knowledge to come in the analysis as well as in the design stage, even in a classical setting (Gelman & Carlin, 2013).

Bayesian methods allow researchers to fold more information into their inferences and decisions. So the most useful Bayesian advice for researchers in management and social science might well be to make analyses larger. Do not shy away from including more data; hierarchical modeling will allow you to handle it.

Recommendations

To illustrate our last point on the relevance of fully using one's data, we point to a recent article by Durante, Arsena, and Griskevicius (2013), which began,

> Each month, many women experience an ovulatory cycle that regulates fertility. Although research has found that this cycle influences women's mating preferences, we proposed that it might also change women's political and religious views. Building on theory suggesting that political and religious orientation are linked to reproductive goals, we tested how fertility influenced women's politics, religiosity, and voting in the 2012 U.S. presidential election. In two studies with large and diverse samples, ovulation had drastically different effects on single women and women in committed relationships. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led women in committed relationships to become more conservative, more religious, and more likely to vote for Mitt Romney. In addition, ovulation-induced changes in political orientation

> mediated women's voting behavior. Overall, the ovulatory cycle not only influences women's politics but also appears to do so differently for single women than for women in relationships. (p. 1007)

The paper in question features a bunch of comparisons and *p* values, some of which were statistically significant, and then lots of stories. The problem is that there are so many different things that could be compared, and all one sees is some subset of the comparisons. Many of the reported effects seem much too large to be plausible. And there is a casual use of causal language (for example, the words *influenced, effects*, and *induced*) to describe correlations.

Beyond all that, the claimed effects seem implausibly large relative to our understanding of the vast literature on voting behavior. For example, the authors report that among women in relationships, 40% in the ovulation period supported Romney, compared to 23% in the nonfertile part of their cycle. Given that opinion polls find very few people switching their vote preferences during the campaign for any reason, these numbers seem unrealistic. The authors might respond that they do not care about the magnitude of the difference, just the sign, but (a) with a magnitude of this size, one is talking noise, noise, noise (not just sampling error but also errors in measurement), and (b) one could just as easily explain this as a differential nonresponse pattern: Maybe liberal or conservative women in different parts of their cycle are more or less likely to participate in a survey. It would be easy enough to come up with a story about that!

In criticizing this study, we are not saying that its claims (regarding the effects of ovulation) are necessarily wrong. We are just saying that the evidence is not nearly as strong as the paper makes it out to be. To researchers in this field, our message is, yes, this topic is important; what is needed is to avoid the trap of considering this sort of small study as providing definitive evidence—even if certain comparisons happen to be statistically significant.

The relevance for the present discussion is that this paper was published in *Psychological Science*, which advertises itself as "the flagship journal of the Association for Psychological Science . . . the highest ranked empirical journal in psychology." So if general statistical principles can improve that particular piece of research, it is reasonable to suppose that these principles could be useful for researchers more generally.

Here is some advice, presented in the context of the study of ovulation and political attitudes but applicable to any study requiring statistical analysis, following the principle noted above that the most important aspect of a statistical method is its ability to incorporate more information into the analysis:

- Analyze *all* your data. For most of their analyses, the authors threw out all the data from participants who were experiencing PMS or having their period. ("We also did not include women at the beginning of the ovulatory cycle [cycle days 1–6] or at the very end of the ovulatory cycle [cycle days 26–28] to avoid potential confounds due to premenstrual or menstrual symptoms.") That was a mistake. Instead of discarding one third of their data, they should have just included that other category in their analysis. This is true of any study in management or elsewhere: Use all of your data to provide yourself and your readers with all the relevant information. Better to anticipate potential criticisms than to hide your data and fear for the eventual exposure.
- Present *all* your comparisons. The paper leads us through a hopscotch of comparisons and *p* values. Better just to present everything. We have no idea if the researchers combed through everything and selected the best results or if they simply made a bunch of somewhat arbitrary decisions throughout of what to look for.

For example, it would be good to see a comparison of respondents in different parts of their cycle on variables such as birth year, party identification, and marital status, and it would be good to see the distribution of reported days of the menstrual cycle. Just a big table (or, better still, a graph) showing these differences for every possible variable.

Instead, what do we get? Several pages full of averages, percentages, F tests, χ^2 tests, and p values, all presented in paragraph form. It would have been better to have all possible comparisons in one convenient display, which if constructed carefully should not take much more space than the traditional prose presentation of a series of estimates and significance tests. These comparisons could then be modeled using hierarchical regression, but even just presenting them would be a start. The point here is not to get an improved p value via a multiple-comparisons correction but rather to see the big picture of the data (Gelman, Hill, & Yajima, 2012). We recognize that, compared to the usual deterministically framed summary, this might represent a larger burden of effort for the consumer of the research as well as the author of the paper.

- Make your data public. If the topic is worth studying, you should want others to be able to make rapid progress. If there are confidentiality restrictions, remove the respondents' identifying information. Then post the data online.

Conclusion

In summary, there are several reasons that Bayesian ideas are relevant to the current crisis of unreplicable findings in social science. First are the familiar benefits of prior information and hierarchical models that allow partial pooling of different data sources. Second, Bayesian approaches are compatible with large uncertainties, which in practice are inevitable when studying interactions. Interactions, in turn, are important because statistically significant but unreplicable results can be seen as arising from varying treatment effects and situation-dependent phenomena (consider, just for one example, the wide variation in estimates of the effects of stereotype threat under different experimental conditions; Fryer, Levitt, & List, 2008; Ganley et al., 2013). Finally, hierarchical Bayesian analysis can handle structured data and multiple comparisons, allowing researchers to escape from the paradigm of the single data comparison or single p value being conclusive.

But Bayesian methods alone will not solve the problem. Most obviously, good analysis is no substitute for good data collection. In small-sample studies of small effects, often all that a good Bayesian analysis will do is reveal the inability to learn much from the data at hand. In addition, little is gained from switching to Bayes if you remain within a traditional hypothesis testing framework. We must move beyond the idea that effects are "there" or not and the idea that the goal of a study is to reject a null hypothesis. As many observers have noted, these attitudes lead to trouble because they deny the variation inherent in real social phenomena, and they deny the uncertainty inherent in statistical inference. Incorporating Bayesian methods in a discipline provides a great opportunity for methods introspection that is unusual in many areas of social research. And, like all statistical methods, Bayesian inference is only as good as its assumptions. We feel it is a strength of Bayesian methods that they force the user to explicitly model the effect size as well as the measurement process, but prior assumptions must then themselves be defensible.

Presenting data more fully and increasing data availability should help all analyses, Bayesian or otherwise. To the extent that there is a norm of presenting and analyzing all comparisons rather than a select few, this should motivate the use of statistical methods that incorporate more data. Indeed, prior information on effect sizes can be used for classical

design analyses, before or after seeing new data (Gelman & Carlin, 2013). More generally, we favor the Bayesian approach for its ability to directly handle multiple levels of uncertainty, but non-Bayesian multilevel models (for example, Skrondal & Rabe-Hesketh) are also available for researchers who are more comfortable with that approach. In a non-Bayesian framework, prior information can be included by performing the analysis with a range of different assumed values for key parameters or by directly modeling the external data that are the basis for the prior knowledge.

It is fine to design a narrow study to isolate some particular features of the world, but you should think about variation when generalizing your findings to other situations. Does $p < .05$ represent eternal truth or even a local truth? Quite possibly not, for two reasons. First is uncertainty: When studying small effects, it is very possible for a large proportion of statistically significant findings to be in the wrong direction as well as to be gross overestimates of the magnitude of the underlying effect (Gelman & Tuerlinckx, 2000). Second is variation: Even if a finding is "real" in the sense of having the same sign as the corresponding comparison in the population, things can easily be different in other populations and other scenarios. In short, an estimated large effect size is typically too good to be true, whereas a small effect could disappear in the noise.

Does this mean that social science is hopeless? Not at all. We can study large differences, we can gather large samples, and we can design studies to isolate real and persistent effects. In such settings, Bayesian inference can help us estimate interactions and make predictions that more fully account for uncertainty. In settings with weaker data and smaller samples that may be required to study rare but important phenomena, Bayesian methods can reduce the now-common pattern of researchers getting jerked around by noise patterns that happen to exceed the statistical significance threshold. We can move forward in social research by accepting uncertainty and embracing variation.

References

Anderson, N., & Ones, D. S. (2008). Rejoinder to Goldberg, Lee and Ashton: Explaining counterintuitive findings. *European Journal of Personality*, 22: 157-162.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100: 407-425.

Berrington, A., & Cox, D. R. (2007). Interpretation of interaction: A review. *Annals of Applied Statistics*, 1: 371-385.

Bloom, H. S., Raudenbush, S. W., & Weiss, M. (2011). *Estimating variation in program impacts: Theory, practice, and applications*. Technical report, MDRC, New York, NY.

Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, 104: 396-404.

Dehejia, R. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, 125: 141-173.

Durante, K. M., Arsena, A. R., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, 24: 1007-1016.

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods*, 13: 668-689.

Emirbayer, M. (1997). Manifesto for a relational sociology. *American Journal of Sociology*, 103: 281-317.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7: 555-561.

Francis, G. (2013). Replication, statistical consistency, and publication bias (with discussion). *Journal of Mathematical Psychology*, 57 (5): 153-169.

Freese, J. (2007). The problem of predictive promiscuity in deductive applications of evolutionary reasoning to intergenerational transfers: Three cautionary tales. In A. Booth, et al. (Eds.), *Intergenerational Caregiving*: 145-178. Washington, D.C.: Urban Institute Press.

Fryer, R. G., Levitt, S. D., & List, J. A. (2008). Exploring the impact of financial incentives on stereotype threat: Evidence from a pilot study. *American Economic Review*, 98: 370-375.

Ganley, C. M., Mingle, L. A., Ryan, A. M., Ryan, K., Vasilyeva, M., & Perry, M. (2013). An examination of stereotype threat effects on girls' mathematics performance. *Developmental Psychology*, 49: 1886-1897.

Gelman, A. (2004). Treatment effects in before-after data. In A. Gelman & X. L. Meng (Eds.), *Applied Bayesian modeling and causal inference from an incomplete data perspective* (pp. 195-202). New York, NY: Wiley.

Gelman, A. (2013a). Correction for "Should the Democrats Move to the Left on Economic Policy?" *Annals of Applied Statistics*, 7: 1248.

Gelman, A. (2013b). *P* values and statistical practice. *Epidemiology*, 24: 69-72.

Gelman, A., & Carlin, J. B. (2013). *Design analysis, prospective or retrospective, using external information*. Technical report, Department of Statistics, Columbia University, New York, NY.

Gelman, A., Carlin, J. B., Stern, H. S., Vehtari, A., Dunson, D. B., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). London, UK: Chapman and Hall.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5: 189-211.

Gelman, A., & Huang, Z. (2008). Estimating incumbency advantage and its variation, as an example of a before/after study (with discussion). *Journal of the American Statistical Association*, 103: 437-451.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2: 1360-1383.

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Technical report, Department of Statistics, Columbia University, New York, NY.

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 15: 373-390.

Gelman, A., & Weakliem, D. (2009). Of beauty, sex, and power: Statistical challenges in estimating small effects. *American Scientist*, 97: 310-316.

Gelman, S. (2005). *The essential child*. Oxford, UK: Oxford University Press.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17: 295-314.

Goldberg, L. R., Lee, K., & Ashton, M. C. (2008a). Comment on Anderson and Ones. *European Journal of Personality*, 22: 151-156.

Goldberg, L. R., Lee, K., & Ashton, M. C. (2008b). *Response to Anderson and Ones (2008)*. Unpublished. Retrieved from http://people.ucalgary.ca/~kibeom/Anderson\%20Ones/AndersonOnes.htm

Greenland, S., & Poole, C. (2013a). Living with *p*-values: Resurrecting a Bayesian perspective on frequentist statistics (with discussion). *Epidemiology*, 24: 62-68.

Greenland, S., & Poole, C. (2013b). Living with statistics in observational research. *Epidemiology*, 24: 73-78.

Ioannidis, J. (2005). Why most published research findings are false. *PLOS Medicine*, 2: e124.

Kanazawa, S. (2005). Big and tall parents have more sons: Further generalizations of the Trivers-Willard hypothesis. *Journal of Theoretical Biology*, 233: 583-590.

Kanazawa, S. (2007). Beautiful parents have more daughters: A further implication of the generalized Trivers-Willard hypothesis. *Journal of Theoretical Biology*, 244: 133-140.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 44: 1372-1381.

Leung, A. K., Kim, S., Polman, E., Ong, L. S., Qui, L., Goncalo, J. A., & Sanchez-Burks, J. (2012). Embodied metaphors and creative "acts." *Psychological Science*, 23: 502-509.

Mischel, W. (1968). *Personality and assessment*. New York, NY: Wiley.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7: 528-530.

Raudenbush, S., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59: 1-17.

Rubin, D. B. (1989). A new perspective on meta-analysis. In K. W. Wachter & M. L. Straf (Ed.), *The future of meta-analysis* (pp. 155-165). New York, NY: Russell Sage Foundation.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science*, 22: 1359-1366.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. London, UK: Chapman and Hall.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58: 267-288.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition (with discussion). *Perspectives on Psychological Science*, 4: 274-324.

Yaari, G. (2012, January 12). *"Hot hand" debate is warming up [Web log post]*. Retrieved from Brainstorm Private Consulting Blog: http://blog.gostorm.net/2012/01/12/hot-hand-debate-is-warming-up/