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*

Andrew Gelman† and Eric Loken‡

14 Nov 2013

"I thought of a labyrinth of labyrinths, of one sinuous spreading labyrinth that would encompass the past and the future . . . I felt myself to be, for an unknown period of time, an abstract perceiver of the world." — Borges (1941)

Abstract

Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of *potential* comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.

1. Multiple comparisons doesn't have to feel like fishing

1.1. Background

There is a growing realization that statistically significant claims in scientific publications are routinely mistaken. A dataset can be analyzed in so many different ways (with the choices being not just what statistical test to perform but also decisions on what data to exclude or exclude, what measures to study, what interactions to consider, etc.), that very little information is provided by the statement that a study came up with a $p < .05$ result. The short version is that it's easy to find a $p < .05$ comparison even if nothing is going on, if you look hard enough—and good scientists are skilled at looking hard enough and subsequently coming up with good stories (plausible even to themselves, as well as to their colleagues and peer reviewers) to back up any statistically-significant comparisons they happen to come up with.

This problem is sometimes called "p-hacking" or "researcher degrees of freedom" (Simmons, Nelson, and Simonsohn, 2011). In a recent article, we spoke of "fishing expeditions, with a willingness to look hard for patterns and report any comparisons that happen to be statistically significant" (Gelman, 2013a).

But we are starting to feel that the term "fishing" was unfortunate, in that it invokes an image of a researcher trying out comparison after comparison, throwing the line into the lake repeatedly until a fish is snagged. We have no reason to think that researchers regularly do that. We think the real story is that researchers can perform a reasonable analysis given their assumptions and their data, but had the data turned out differently, they could have done other analyses that were just as reasonable in those circumstances.

*A slightly edited version of this paper has appeared in *American Scientist* under the title, "The statistical crisis in science." We thank Ed Vul, Howard Wainer, Macartan Humphreys, and E. J. Wagenmakers for helpful comments and the National Science Foundation for partial support of this work.

†Department of Statistics, Columbia University, New York

‡Department of Human Development and Family Studies, Penn State University

We regret the spread of the terms "fishing" and "p-hacking" (and even "researcher degrees of freedom") for two reasons: first, because when such terms are used to describe a study, there is the misleading implication that researchers were consciously trying out many different analyses on a single data set; and, second, because it can lead researchers who know they did not try out many different analyses to mistakenly think they are not so strongly subject to problems of researcher degrees of freedom. Just to be clear: we are not saying that Simmons et al. (2001), Vul et al. (2009), Francis (2013), and others who have written about researcher degrees of freedom were themselves saying that p-hacking implies that researchers are cheating or are even aware of performing data-dependent analyses. But the perception remains, hence the need for this paper.

Our key point here is that it is possible to have multiple *potential* comparisons, in the sense of a data analysis whose details are highly contingent on data, without the researcher performing any conscious procedure of fishing or examining multiple p-values.

1.2. Theoretical framework

The statistical framework of this paper is frequentist: we consider the statistical properties of hypothesis tests under hypothetical replications of the data. Consider the following testing procedures:

1. Simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$.

2. Classical test pre-chosen from a set of possible tests: thus, $T(y; \phi)$, with preregistered ϕ. For example, ϕ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as the decision of which main effect or interaction to focus on.

3. Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y; \phi(y))$, where the function $\phi(\cdot)$ is observed in the observed case.

4. "Fishing": computing $T(y; \phi_j)$ for $j = 1, \ldots, J$: that is, performing J tests and then reporting the best result given the data, thus $T(y; \phi^{\text{best}}(y))$.

Our claim is that researchers are doing #3, but the confusion is that, when we say this, researchers *think* we're accusing them of doing #4. To put it another way, researchers assert that they are not doing #4 and the implication is that they are doing #2. In the present paper we focus on possibility #3, arguing that, even without explicit fishing, a study can have a huge number of researcher degrees of freedom, following what de Groot (1956) refers to as "trying and selecting" of associations.

In this article we discuss several papers recently published in top psychology journals that are centered on p-values which we think have been obtained through process #3 above. Our goal is not to single out these particular projects but rather to use them to illustrate general points about researcher degrees of freedom.

It might seem unfair that we are criticizing published papers based on a claim about what they would have done had the data been different. But this is the (somewhat paradoxical) nature of frequentist reasoning: if you accept the concept of the p-value, you have to respect the legitimacy of modeling what would have been done under alternative data. Again, we are not saying that the papers under consideration have followed procedure #4 above, but we see no evidence that they are doing #2 or anything close to it. When criticisms of multiple comparisons have come up in

regards to some of the papers we discuss here, the researchers never respond that they had chosen all the details of their data processing and data analysis ahead of time; rather, they claim that they picked only one analysis *for the particular data they saw*. Intuitive as this defense may seem, it does not address the fundamental frequentist concern of multiple comparisons.

We illustrate with a very simple hypothetical example. A researcher is interested in differences between Democrats and Republicans in how they perform in a short mathematics test when it is expressed in two different contexts, either involving health care or the military. The research hypothesis is that context matters, and one would expect Democrats to do better in the health-care context and Republicans in the military context. Party identification measured on a standard 7-point scale and various demographic information also available. At this point there is a huge number of possible comparisons that can be performed—all consistent with the data. For example, the pattern could be found (with statistical significance) among men and not among women—explicable under the theory that men are more ideological than women. Or the pattern could be found among women but not among men—explicable under the theory that women are more sensitive to context, compared to men. Or the pattern could be statistically significant for neither group, but the difference could be significant (still fitting the theory, as described above). Or the effect might only appear among men who are being asked the questions by female interviewers. We might see a difference between sexes in the health-care context but not the military context; this would make sense given that health care is currently a highly politically salient issue and the military is not. There are degrees of freedom in the classification of respondents into Democrats and Republicans from a 7-point scale. And how are independents and nonpartisans handled? They could be excluded entirely. Or perhaps the key pattern is between partisans and nonpartisans? And so on. In our notation above, a single overarching research hypothesis—in this case, the idea that issue context interacts with political partisanship to affect mathematical problem-solving skills—corresponds to many different possible choices of the decision variable ϕ.

At one level, these multiplicities are obvious. And it would take a highly unscrupulous researcher to perform test after test in a search for statistical significance (which could almost certainly be found at the 0.05 or even the 0.01 level, given all the options above and the many more that would be possible in a real study). We are not suggesting that researchers generally do such a search. What we are suggesting is that, given a particular data set, it is not so difficult to look at the data and construct completely reasonable rules for data exclusion, coding, and data analysis that can lead to statistical significance—thus, the researcher needs only perform one test, but that test is conditional on the data; hence, $T(y; \phi(y))$ as described in item #4 above. As Humphreys, Sanchez, and Windt (2013) write, a researcher when faced with multiple reasonable measures can reason (perhaps correctly) that the one that produces a significant result is more likely to be the least noisy measure, but then decide (incorrectly) to draw inferences based on that one only.

This is all happening in a context of small effect sizes, small sample sizes, large measurement errors, and high variation (which combine to give low power, hence less reliable results even when they happen to be statistically significant, as discussed by Button et al., 2013). Multiplicity would *not* be a huge problem in a setting of large real differences, large samples, small measurement errors, and low variation. This is the familiar Bayesian argument: any data-based claim is more plausible to the extent it is *a priori* more likely and less plausible to the extent that it is estimated with more error. That is the context; in the present paper, though, we focus on the very specific reasons that published p-values cannot be taken at face value, even if no explicit fishing has been done.

2. Choices of main effects or interactions

2.1. Example: fat arms and political attitudes

In a recent general-interest article (Gelman, 2013a) on researcher degrees of freedom, we wrote the following about a paper (Petersen et al., 2013) that claimed to find an association of men's upper-body strength, interacted with socioeconomic status, and their attitudes about economic redistribution:

> The authors wrote, "We showed that upper-body strength in modern adult men influences their willingness to bargain in their own self-interest over income and wealth redistribution. These effects were replicated across cultures and, as expected, found only among males." Actually, two of their three studies were of college students, and they did not actually measure anybody's upper-body strength; they just took measurements of arm circumference. It's a longstanding tradition to do research studies using proxy measures on students—but if it's OK to do so, you should be open about it; instead of writing about "upper-body strength" and "men," be direct and say "arm circumference" and "male students." Own your research choices!
>
> But, to return to the main theme here, these researchers had enough degrees of freedom for them to be able to find any number of apparent needles in the haystack of their data. Most obviously, the authors report a statistically significant interaction with no statistically significant main effect. That is, they did not find that men with bigger arm circumference had more conservative positions on economic redistribution. What they found was that the correlation of arm circumference with opposition to redistribution of wealth was higher among men of high socioeconomic status. But, had they seen the main effect (in either direction), we are sure they could have come up with a good story for that, too. And if there had been no main effect and no interaction, they could have looked for other interactions. Perhaps, for example, the correlations could have differed when comparing students with or without older siblings?

We regret the tone of the last paragraph above in that it might seem to imply that the researchers were fishing for statistical significance, looking at interaction after interaction until something jumped out. This is not what we meant to say. Rather, we think that it is completely reasonable for good scientists to refine their hypotheses in light of the data. When the desired pattern does not show up as a main effect, it makes sense to look at interactions. For example, our above mention of older siblings was no joke, as family relations are often taken to be crucial in evolutionary psychology explanations.

Our first reaction when seeing a analysis of young men's arm circumference is that this could be a proxy for age. And, indeed, for the analyses from the two countries where the samples were college students, when age is thrown into the model, the coefficient for arm circumference (or, as the authors put it, "upper-body strength") goes away.

But then comes the big problem. The authors draw their conclusions from their inferences on the interaction between arm circumference and socioeconomic status. But the analyses do not adjust for the interaction between age and socioeconomic status. The tricky part here is that it has been found that political attitudes and political commitments change around college age: people start voting, and their attitudes become more partisan (Mullainathan and Washington, 2009). It might be argued that all this is simply a product of changing upper-body strength, but to us such a claim would be more than a bit of a stretch.

There also appear to be some degrees of freedom involved in the measurement. From the supplementary material of Petersen et al. (2013):

> The interaction effect is not significant when the scale from the Danish study are used to gauge the US subjects' support for redistribution. This arises because two of the items are somewhat unreliable in a US context. Hence, for items 5 and 6, the inter-item correlations range from as low as .11 to .30. These two items are also those that express the idea of European-style market intervention most clearly and, hence, could sound odd and unfamiliar to the US subjects. When these two unreliable items are removed (α after removal = .72), the interaction effect becomes significant.
>
> The scale measuring support for redistribution in the Argentina sample has a low α-level and, hence, is affected by a high level of random noise. Hence, the consistency of the results across the samples is achieved in spite of this noise. A subscale with an acceptable $\alpha = .65$ can be formed from items 1 and 4.

Again, these decisions may make perfect sense but they are clearly contingent on data.

In our own work, we have often found interactions to be important (see Gelman and King, 1994, for an example and Gelman, 2004, for a longer discussion of the general importance of interactions in causal inference). So we are certainly not saying that researchers should avoid looking at interactions or measuring their statistical significance. And in our own work we have often thought of interactions only after looking at the data (see, for example, Gelman et al., 2007). The point is that claims of interactions need to make sense of their own accord. It is not enough to rely on statistical significance. In the particular example of arm circumference and political attitudes, there were many problems with the interpretation of the interaction as estimated from noisy data in a nonrepresentative sample, and our key point is that the statistically significant p-value cannot be taken at face value.

2.2. Example: Appearance or non-appearance of ESP

A much-discussed example of possibly spurious statistical significance is the claim of Bem (2011) to have found evidence for extrasensory perception in college students. This particular article got huge media attention after being published in a top journal despite the misgivings of many psychologists. After much criticism and some failed attempts at replications, the furor has mostly subsided, but this example remains of interest, partly because it is an example of how generally-accepted research practices can be used to find statistical significance anywhere (by analogy to the poster of Bennett et al., 2009, finding statistically significant patterns in an MRI of a dead salmon, illustrating the risks of multiple comparisons in brain imaging).

Here are some multiplicity problems we noted regarding Bem (2011):

> The paper included nine different experiments and many statistically significant results. Unfortunately . . . these experiments had multiple degrees of freedom that allowed Bem to keep looking until he could find what he was searching for. In his first experiment, in which 100 students participated in visualizations of images, he found a statistically significant result for erotic pictures but not for nonerotic pictures. But consider all the other possible comparisons: If the subjects had identified all images at a rate statistically significantly higher than chance, that certainly would have been reported. Or what if performance had been higher for the nonerotic pictures? One could easily argue that the erotic images were distracting and only the nonerotic images were a good test of the phenomenon. Or what if participants had performed statistically significantly better in

> the second half of the trial than in the first half? That would be evidence of learning. Or if they performed better on the first half? Evidence of fatigue. Bem reports, "There were no significant sex differences in the present experiment." If there had been (for example, if men had performed better with erotic images and women with romantic but nonerotic images), this certainly could have been presented as convincing evidence. And so on.

This criticism has been made by others (Wagenmakers et al., 2011), and Bem did reply to it. In a section entitled, "The Hypotheses Were Not Exploratory," Bem, Utts, and Johnson (2011) wrote:

> That experiment was designed to test the hypothesis that participants could identify the future left/right position of an erotic image on the computer screen significantly more frequently than chance. The results showed that they could. The specificity of this hypothesis derives from several earlier "presentiment" experiments (e.g., Radin, 1997) which had demonstrated that participants showed anomalous "precognitive" physiological arousal a few seconds before seeing an erotic image but not before seeing a calm or nonerotic image. Accordingly, Experiment 1 also included randomly interspersed trials with nonerotic images, leaving as an open question whether participants might also be able to anticipate the future left/right positions of these images. They could not, a finding consistent with the results of the presentiment experiments. The important point here is that the central psi hypothesis about erotic images was unambiguous, directional, based on previous research, not conditional on any findings about trials with nonerotic images, and was not formulated from a post hoc exploration of the data. In fact, there was no data exploration that required adjustment for multiple analyses in this or any other experiment.

We have no reason to disbelieve the above description of motivations; nonetheless it seems clear to us that the *scientific* hypotheses there described correspond to multiple *statistical* hypotheses. For example, consider the statement about "anomalous precognitive physiological arousal." Suppose that the experimental subjects had performed statistically significantly *worse* for the erotic pictures. This would fit right into the theory, in that the anomalous arousal could be interfering with otherwise effective precognitive processes.

The other point worth discussing is Bem's statement that his hypothesis "was not formulated from a post hoc exploration of the data." That may be so—but a data-dependent analysis will not necessarily look "post hoc." For example, if men had performed better with erotic images and women with romantic but nonerotic images, there is no reason why such a pattern would look like fishing or p-hacking. Rather, it would be a natural implication of the research hypothesis, as of course there is literature suggesting sex differences in response to visual erotic stimuli. There is a one-to-many mapping from scientific to statistical hypotheses.

2.3. Example: menstrual cycle and vote intentions

In a paper recently published in *Psychological Science*, Durante, Arsena, and Griskevicius (2013) wrote:

> 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 versus married women. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led married women to become more conservative, more religious, and more likely to vote for Mitt Romney. ... Overall, the ovulatory cycle not only influences women's politics, but appears to do so differently for single versus married women.

The causal language above is misleading as the study was entirely between-subject: that is, a sample of women completed a survey once, and correlations were found between their stated vote preference, marital status, and day in their menstrual cycle. There were no repeated measurements, hence individual women were not compared at different days in the cycle. We will set this issue aside here, however, and focus on the authors' decision to focus on an interaction as their central result. Given the authors' general theoretical perspective ("ovulation should lead women to prioritize securing genetic benefits from a mate possessing indicators of genetic fitness"), the interaction that they found makes sense. But various other main effects and interactions would also fit the theory. Indeed, as the authors note, their hypothesis "is consistent with the idea that women should support the more liberal candidate." Or suppose the data had followed the opposite pattern, with ovulation (as assessed by the researchers) being correlated with conservative attitudes among single women and correlated with liberal attitudes among married women. This would fit a story in which ovulation leads women's preferences away from party identification and toward more fundamental biological imperatives.

One might feel that these comments of ours are a form of nitpicking; after all, these researchers found a large effect that was consistent with their theory. So what if the significance level was somewhat overstated because of multiple comparisons problems? How important is this, really? Are we just trying to take down a high-quality study based on a technicality? Our answer is no, for two reasons. First, the claimed effect size, in the range of a 20 percentage point difference in vote intention at different phases of the menstrual cycle, is substantively implausible given all the evidence from polling that very few people change their vote intentions during presidential general election campaigns. The study at hand thus has very low power (even more so given the inexact measurement of fertility based on a survey response), hence any observed difference is likely attributable to noise (Button et al., 2013). Second, the statistical significance of the published comparisons is a central part of the argument of Durante et al.—certainly the paper would not have been published in a top journal without $p < 0.05$ results, and the high multiplicity of all the potential interactions is relevant to this point.

In addition to the choice of main effects or interactions, the authors of this study also had several different political questions to work with (attitudes as well as vote intention) along with other demographic variables (age, ethnicity, and parenthood status), and flexibility in characterizing relationship status. (In the abstract of the paper, the authors refer to "single women" and "married women," but later in the article, they write, "participants who indicated being engaged, living with a partner, or married were classified as In a Committed Relationship" and "all others (e.g., not dating, dating) were classified as Single." Considering all these possible choices represents a combinatorial explosion of possible data analyses. To stay with the theme of the present paper, we emphasize that, from the researchers' perspective with their single data set, these might not seem like choices at all: it can be perfectly possible to take the data and take a series of steps that flow naturally from theory—not realizing that, with a different data set, other choices could be made (for example, comparing women with children to women without children, or comparing whites to nonwhites, or dividing up the relationship categories in a different way, or keeping rather than discarding women in the last days of their menstrual cycle) that would make just as much

theoretical sense, and would seem like the only possible analysis decisions given the data that appeared.

3. Choices in data processing and data analysis

In the previous section we discussed several prominent research papers in which statistical significance was attained via a sort of invisible multiplicity: data-dependent analysis choices that did not appear to be degrees of freedom because researchers analyze only one dataset at a time.

Here we focus on a single published study, also in a top psychology journal, and consider several other forms of multiplicity. Beall and Tracy (2013) reported that women who were at peak fertility were three times more likely to wear red or pink shirts, compared to women at other points in their menstrual cycles. In a published criticism (Gelman, 2013), we wrote that there were many different comparisons that could have been reported in the data, so there was nothing special about a particular comparison being statistically significant. We likened their paper to other work which we considered flawed for multiple comparisons (too many researcher degrees of freedom), including the papers on arm circumference and ESP discussed above.

In their response, Tracy and Beall (2013) wrote:

> Gelman's concern here seems to be that we could have performed these tests prior to making any hypothesis, then come up with a hypothesis post-hoc that best fit the data. While this is a reasonable concern for studies testing hypotheses that are not well formulated, or not based on prior work, it simply does not make sense in the present case. We conducted these studies with the sole purpose of testing one specific hypothesis: that conception risk would increase women's tendency to dress in red or pink. This hypothesis emerges quite clearly from the large body of work mentioned above, which includes a prior paper we co-authored ... The existence of this prior published article provides clear evidence that we set out to test a specific theory, not to conduct a fishing expedition....

But, whether or not the authors of the study were "mining the data," it seems clear that their analysis was contingent on the data. They had many data-analytic choices, including rules for which cases to include or exclude and which comparisons to make, as well as what colors to study. Their protocol and analysis were not pre-registered. Even though Beall and Tracy did an analysis that was consistent with their general research hypothesis—and we take them at their word that they were not conducting a fishing expedition—many degrees of freedom remain in their specific decisions.

3.1. Data exclusion and coding rules

Beall and Tracy had several inconsistencies in their rules for which data to use in their analysis. For their second sample, 9 of the 24 women didn't meet the inclusion criteria of being more than five days away from onset of menses, but they were included anyway. 22% of their first sample also didn't meet that criterion but were included anyway. And even though the first sample was supposed to be restricted to women younger than 40, the ages of the wome included ranged up to 47. Out of all the women who participated across the two samples, 31% were excluded for not providing sufficient precision and confidence in their answers (but sufficient precision will vary with time since menses; it is easier to be certain ± 1 day for something that occurred 5 days ago as opposed to 22 days ago).

In addition, the authors found a statistically significant pattern after combining red and pink, but had they found it only for red, or only for pink, this would have fit their theories too. In their words: "The theory we were testing is based on the idea that red and shades of red (such as the pinkish swellings seen in ovulating chimpanzees, or the pinkish skin tone observed in attractive and healthy human faces) are associated with sexual interest and attractiveness." Had their data popped out with a statistically significant difference on pink and not on red, that would have been news too. And suppose that white and gray had come up as the more frequent colors? One could easily argue that more bland colors serve to highlight the pink colors of a (European-colored) face.

Finally, there was ambiguity about the choice of days to compare during the menstrual cycle. Beall and Tracy defined "peak fertility" as days 6–14 of the menstrual cycle (which they confusingly defined as a 28-day period going from day 0 through day 28). But according to womenshealth.gov, the most fertile days are between days 10 and 17. Babycenter.com says days 12 to 17. When we pointed this out, Beall and Tracy said that they chose days 6–14 a priori as it represents "the standard practice in our field." Maybe, but that's not necessarily the whole story. What is standard practice appears a bit flexible in practice. Consider another paper published in *Psychological Science* the same year, that of Durante et al. (2013) which we discussed in Section 2.3. That paper has a lot of similarities to that of Beall and Tracy (it is a between-subjects study on a small sample of internet participants), but in that case they compared days 7–14 of the menstrual cycle to days 17–25, completely excluding days 1–6, 15–16, and days 26–28 (and, unlike Beall and Tracy, not defining a day 0 at all). So, even if you accept the days 6–14 period (rather than Durante et al.'s choice of days 7–14) as high fertility, there's still the choice of what to compare to. And, yes, it may be that Beall and Tracy planned ahead of time to compare 6–14 to all other days. But suppose they had seen a lot of red during days 0–6? Perhaps they would have decided to follow the literature and just use 17–25 as a comparison. And that would not seem like an arbitrary choice; indeed, it could well be seen as the only legitimate data-analytic choice in that it would have been following the previous literature and excluding data which, from their theory, could be considered irrelevant to the comparison.

Again, all the above could well have occurred *without* it looking like "p-hacking" or "fishing." It's not that the researchers performed hundreds of different comparisons and picked ones that were statistically significant. Rather, they start with a somewhat-formed idea in their mind of what comparison to perform, and they refine that idea in light of the data. They saw a pattern in red and pink, and they combined the colors. Had they seen a pattern only in pink, that would have worked too, and we could be having this very same discussion except that the researchers would be explaining why they only considered pink. Similarly with the days of fertility: had the data for dates 0–5, 15–16, and 26–28 not fit the story, they easily could have followed the choice of Durante et al. and, again, they could be insisting (perfectly truthfully) that they were following the literature and going with their pre-chosen research plan. Or if they had found a pattern among the middle-aged adult women in their internet sample but not in their sample of college students, this would make perfect sense too, and they could have been perfectly comfortable reporting that (just as Durante et al. had no problem reporting an interaction as their central claim, with opposite patterns for single and for married women).

In each of these cases, the data analysis would not feel like "fishing" because it would all seem so reasonable. Whatever data-cleaning and analysis choices were made, contingent on the data, would seem to the researchers as the single choice derived from their substantive research hypotheses. They would feel no sense of choice or "fishing" or "p-hacking"—even though different data could have led to different choices, each step of the way.

This example is particularly stark because Beall and Tracy on one hand, and Durante et al. on the other, published two studies inspired by very similar stories, using similar research methods,

in the same journal in the same year. But in the details they made different data-analytic choices, each time finding statistical significance with the comparisons they chose to focus on. Both studies compared women in ovulation and elsewhere in their self-reported menstrual cycles, but they used different rules for excluding data and different days for their comparisons. Both studies examined women of childbearing age, but one study reported a main effect whereas the other reported a difference between single and married women. In neither case were the data inclusion rules and data analysis choices pre-registered.

In this garden of forking paths, whatever route you take seems predetermined, but that's because the choices are done implicitly. The researchers are not trying multiple tests to see which has the best p-value; rather, they are using their scientific common sense to formulate their hypotheses in reasonable way, given the data they have. The mistake is in thinking that, if the particular path that was chosen yields statistical significance, that this is strong evidence in favor of the hypothesis.

3.2. Choices in combining information from separate studies

Beall and Tracy note that their pattern appears and is statistically significant in two samples, one from adults on the internet and one of college students. However, a pattern in just one group and not the other could also have been notable and explainable under the larger theory, given the different ages of the two groups of participants. And, had a striking but not-quite statistically significant pattern been observed in each sample, it would have seemed reasonable to combine the results from the two samples or to gather a third sample to reach significance. Again, their data-analysis choices seem clear, conditional on the data they saw, but other choices would have been just as reasonable given other data, allowing many different possible roads to statistical significance.

When considering women's clothing choices, Beall and Tracy (2013) find consistent data on college students and adult women. That is, they analyze main effects. But, as noted in Section 2.3, in an otherwise very similar study on women's vote intentions, performed using similar methodology, Durante et al. (2013) contrast the attitudes of married and unmarried women. They find no main effect but report a statistically significant interaction. The option to pool data from different studies, or to isolate individual studies as statistically significant, or to contrast studies or subgroups within a study (that is, to analyze interactions) provides a huge number of potential options and comparisons to report, and it is unsurprising that some of these will be both statistically significant and consistent with pre-existing theories. For example, as noted above, it would seem completely consistent with Beall and Tracy's model of sexual signaling if college women, at the height of their fertility, were to be dressing in a way consistent with their ovulation cycle, while older women (many of whom would be married) would not feel that subconscious need to attract the male gaze. The opposite story would also fit: it would make sense that college women—who are often sexually active and want to avoid getting pregnant—would subconsciously *avoid* sending sexual messages during their most fertile times—whereas older women would send these signals more urgently. It would be the role of the data (in a large, well-powered study with low measurement error) to distinguish among such theories—but it is not particularly informative that one of many possible main effects and interactions happens to have reached the level of statistical significance.

4. Discussion

4.1. Summary

When we say an analysis was subject to multiple comparisons or "researcher degrees of freedom," this does not require that the people who did the analysis were actively trying out different tests

in a search for statistical significance. Rather, they can be doing an analysis which at each step is contingent on the data. The researcher degrees of freedom do not *feel* like degrees of freedom because, conditional on the data, each choice appears to be deterministic. But if we average over all possible data that could have occurred, we need to look at the entire garden of forking paths and recognize how each path can lead to statistical significance in its own way. Averaging over all paths is the fundamental principle underlying p-values and statistical significance and has an analogy in path diagrams developed by Feynman to express the indeterminacy in quantum physics.

We have focused on several high-profile, small-sample studies in social psychology. We expect that similar problems arise in other fields such as medical research, but we stick with psychology examples here, partly because they can be addressed with minimal technical knowledge regarding drug mechanisms, imaging, genes, and so forth. In all the cases we have discussed, the published analysis has a story that is consistent with the scientific hypotheses that motivated the work, but other data patterns (which, given the sample sizes, could easily have occurred by chance) would naturally have led to different data analyses (for example, a focus on main effects rather than interactions, or a different choice of data subsets to compare) which equally could have been used to support the research hypotheses. The result remains, as we have written elsewhere, a sort of machine for producing and publicizing random patterns.

We are not saying the scientific claims in these papers are necessarily wrong. Maybe there really do exist large ESP effects. Maybe there really are consistent differences in the colors that women wear during different parts of their cycle. And so on. What we are saying is that the evidence in these research papers is not as strong as stated. We certainly do not believe that the patterns published in these articles (for example, women being three times more likely to wear red or pink during certain days in their menstrual cycle) would hold in the general population, nor do we believe that even the direction of these effects would be likely to reappear in pre-registered replications (for reasons explained by Button et al., 2013, and Nosek, Spies, and Motyl, 2013). The scientific hypotheses in the papers at hand are general enough that they could have some validity even if the particular claims in the published claims do not hold.

To put it another way, we view these papers—despite their statistically significant p-values—as exploratory, and when we look at exploratory results we must be aware of their uncertainty and fragility. It does not seem to us to be good scientific practice to make strong general claims based on noisy data, and one problem with much current scientific practice is the inadvertent multiplicity of analysis that allows statistical significance to be so easy to come by, with researcher degrees of freedom hidden because researcher only sees one data set at a time.

We want our scientists to be creative, but we have to watch out for a system that allows any hunch to be ratcheted up to a level of statistical significance that is then taken as scientific proof. And we need to be aware of these concerns in our own research, not just in that of others.

Meanwhile, in understanding these statistical problems, we want to make clear that multiple comparisons can be a big problem, without any implication that the researchers in question are cheating or stupid or trying to manipulate the system. When we described these sorts of research as "fishing expeditions," we were making a mistake. "Fishing" and "p-hacking" imply an active pursuit of statistical significance, whereas what might well be going on here is a set of data-analysis choices that might well be reasonable, were it not for problems of small sample size and measurement error that make the results noisier than people realize.

4.2. Possible reactions

How might the researchers in the above-cited papers (and similar studies) react to our remarks about data-dependent analyses? We can think of three possible responses:

1. Concession: Yes, the skeptics are correct. Published p-values cannot generally be taken at face value because the analyses are contingent on data. As a result, the authors of such studies would have to recognize that the evidence in favor of their research hypotheses is much weaker than they had presumed. This is the response that we hope for, and one of the motivations for writing this paper is to emphasize that, by saying that researchers did data-dependent analyses, we are not implying any "fishing" or unethical individual behavior. Rather, we just think it is natural that researchers focus on what they did with the data at hand, without putting much thought into how the details of their decisions could have changed under different realizations of the data.

2. Insistence: No, the published analyses are the only operations that would have been done to the data. Had the observed data come out differently, the exact same rules would have been used to keep or discard data and the exact same comparisons would have been done. We would find such insistence implausible, for at least three reasons. First, if the data analysis protocol really were decided ahead of time, why not preregister it? Second, different data-analysis choices appear even in studies on the same topic in the same subfield (as we saw above regarding the different data exclusion rules and choices of dates in the menstrual cycle, comparing the Beall and Tracy study to that of Durante et al.). Third, when researchers insist that their analysis choices were not data-dependent, we generally see a gap between the stated scientific hypotheses and the tested statistical hypotheses. For one of many examples, often an analysis is performed on two groups or under two conditions. When a result is statistically significant over all, that can be reported. When it is significant in only one of the groups but not the other, that can be reported as evidence too. Or if the difference (the interaction) is statistically significant, that also works. All of these will seem fully consistent with a research hypothesis (see, for example, Bem et al.'s discussion of the nonerotic images, data which can be taken as a control condition or as a separate confirmatory study) but they represent different analyses, different paths that can be taken.

3. Contingency: A final, and in a way most interesting, response would be for a researcher to say, Yes, had the data gone differently the published analysis would have been different, but that's exactly the point: we found what we found. For example, suppose that a statistically significant difference had occurred only with pink clothing and not with red. Beall and Tracy's paper then would indeed be different, all the way to its title, and they could argue that this is as it should be: a scientific paper should report what has been found, not merely what has been expected.

   We are sympathetic to this last argument, but to truly follow its implications, we must think carefully about exactly what information is contained in a published statistically-significant comparison. Once we recognize that analysis is contingent on data, the p-value argument disappears—one can no longer argue that, if nothing were going on, that something as extreme as what was observed would occur less than 5% of the time.

Once the claim of statistical significance is abandoned, one can take a Bayesian, or approximately Bayesian, view and consider the observed comparison as evidence of the underlying effect. For example, from Bem (2011): "Across all 100 sessions, participants correctly identified the future position of the erotic pictures significantly more frequently than the 50% hit rate expected by chance: 53.1%, $t(99) = 2.51$, $p = .01$, $d = 0.25$." Setting the p-value aside, we can still take this as an observation that 53.1% of these guesses were correct, and if we combine this with a flat prior distribution (that is, the assumption that the true average probability of a correct guess under these conditions is equally likely to be anywhere between 0 and 1) or, more generally, a locally-flat

prior distribution, we get a posterior probability of over 99% that the true probability is higher than 0.5; this is one interpretation of the one-sided p-value of 0.01 (Greenland and Poole, 2013). The trouble is that a uniform or even a locally uniform prior distribution is *not* appropriate in this setting. The Bem study is typical of scientific research in that a new idea is being evaluated, in a context where various earlier tries were not so successful. There was no track record of conditions under which subjects could consistently achieve 53% success in a precognition experiment (and, indeed, Bem's own experiments failed to replicate in independent trials).

At this point, Bem or others might argue that we would being unfair as Bayesians to assigning such a pessimistic prior. One response (beyond a statement that this is a prior distribution we judge to be reasonable given available information) is that it would be possible to perform a convincing experiment via a p-value obtained from a preregistered study. But without preregistration or the equivalent, it is difficult to take $p = 0.01$ as evidence for much of anything. The correct Bayesian result also should be conditional on all the data, not just on the single comparison that is highlighted. It is not appropriate to condition on a single comparison while ignoring all the other non-significant comparisons that were not done, hence an appropriate Bayesian analysis can require additional effort in the form of hierarchical modeling.

4.3. The way forward

There are many roads to statistical significance, and if data are gathered with no preconceptions at all, it is obvious that statistical significance can be obtained from pure noise, just by repeatedly performing comparisons, excluding data in different ways, examining different interactions and controlling for different predictors, and so forth. Realistically, though, a researcher will come into a study with strong substantive hypotheses, to the extent that, for any given dataset, the appropriate analysis can seem evidently clear. But, even if the chosen data analysis is a deterministic function of the observed data, this does not eliminate the multiple comparisons problem. P-values are based on what would have happened under other possible datasets.

What, then, can be done? Humphreys, Sanchez, and Windt (2013) and Monogan (2013) recommend preregistration: defining the entire data-collection and data-analysis protocol ahead of time. For most of our own research projects this strategy hardly seems possible: in our many applied research projects, we have learned so much by looking at the data. Our most important hypotheses could never have been formulated ahead of time. In addition, as applied social science researchers we are often analyzing public data on elections, the economy, and public opinion that have already been studied by others many times before, and it would be close to meaningless to consider preregistration for data with which we are already so familiar. In fields such as psychology where it is typically not so difficult to get more data, preregistration might make sense. In two of the examples discussed in the present paper, Bem and also Beall and Tracy implied that their data analysis choices were not affected by the data. If they feel strongly about this, perhaps in the future they and others will preregister their research protocols so that their future p-values will not be questioned.

At the same time, we do not want demands of statistical purity to strait-jacket our science. The most valuable statistical analyses often arise only after an iterative process involving the data (see, e.g., Tukey, 1980, and Box, 1997). One message of the present article is that researchers can and should be more aware of the choices involved in their data analysis, partly to recognize the problems with published p-values but, ultimately, with the goal of recognizing the actual open-ended aspect of their projects (in the notation of Section 1.2, a large space of decision variables ϕ corresponding to a single research hypothesis) and then analyzing their data with this generality in mind.

One can follow up an open-ended analysis with pre-publication replication, which is related

to the idea of external validation which is popular in statistics and computer science. The idea is to perform two experiments, the first being exploratory but still theory-based, and the second being purely confirmatory with its own preregistered protocol. All three of the papers we have discussed here could have followed this strategy, as all were relatively inexpensive and noninvasive psychology experiments performed on volunteers. In each of these cases, we personally doubt that a replication would have been successful—we would guess that, in each of the examples, the probability is something 5% that the result under a preset protocol would be been statistically significant—but the data could (probabilistically) prove us wrong. One would still have to contend with the file-drawer problem, but one would hope that unsuccessful replications would still be published in some outlet such as Plos-One.

Nosek, Spies, and Motyl (2013) recently presented an appealing example of pre-publication replication in one of their own studies, in which they performed an experiment on perceptual judgment and political attitudes, motivated and supported by substantive theory. They found a large and statistically significant relationship—but rather than stopping there and publishing these results, they gathered a large new sample and performed a replication with predetermined protocols and data analysis. According to their estimates based on their initial, statistically-significant result, the replication had over 99% power. Nonetheless, the attempted replication was unsuccessful, with a p-value of .59. We suspect this sort of thing would happen if the authors of the papers under discussion were to attempt to replicate their studies in an environment in which all data decisions were specified in advance.

In (largely) observational fields such as political science, economics, and sociology, replication is more difficult. We cannot easily gather data on additional wars, or additional financial crises, or additional countries. In such settings our only recommendation can be to more fully analyze existing data. A starting point would be to analyze all relevant comparions, not just focusing on whatever happens to be statistically significant. We have elsewhere argued that multilevel modeling can resolve multiple-comparisons issues (Gelman, Hill, and Yajima, 2012) but the practical difficulties of such an approach are not trivial. When the number of comparisons is small or the problem is highly structured, inference can be sensitive to model assumptions. Ultimately this should not be a problem (after all, researchers are necessarily relying on model assumptions all the time when deciding which comparisons are worth looking at), but developing general approaches to analyze multiple potential comparisons is an area for future research. Until then, we think it is important to recognize that, even if only a single analysis is performed on a dataset, this does not mean that p-values can be taken literally. A multiple comparisons problem does not have to come from "fishing" but can arise more generally from reasonable processing and analysis decisions that are contingent on data.

4.4. Toward a positive message

Criticism is easy, doing research is hard. Flaws can be found in any research design if you look hard enough. Our own applied work is full of analyses that are contingent on data, yet we and our colleagues have been happy to report uncertainty intervals (and thus, implicitly, claims of statistical significance) without concern for selection bias or multiple comparisons. So we would like to put a positive spin on the message of this paper, to avoid playing the role of statistician as scold (Gelman, 2013b). P-values are a method of protecting researchers from declaring truth based on patterns in noise, and so it is ironic that, by way of data-dependent analyses, p-values are often used to lend credence to noisy claims based on small samples. To put it another way: *without* modern statistics, we find it unlikely that people would take seriously a claim about the general population of women, based on two survey questions asked to 100 volunteers on the internet and 24 college students. But

with the p-value, a result can be declared significant and deemed worth publishing in a leading journal in psychology.

Our positive message is related to our strong feeling that scientists are interested in getting closer to the truth. In the words of the great statistical educator Frederick Mosteller, it is easy to lie with statistics, but easier without them. In our experience, it is good scientific practice to refine one's research hypotheses in light of the data. Working scientists are also keenly aware of the risks of data dredging, and they use confidence intervals and p-values as a tool to avoid getting fooled by noise. Unfortunately, a byproduct of all this struggle and care is that, when a statistically significant pattern *does* show up, it is natural to get excited and believe it. It is the very fact that scientists generally *don't* cheat, generally *don't* go fishing for statistical significance, that they are inclined to draw strong conclusions when they do encounter a pattern that is strong enough to cross the $p < .05$ threshold.

We must realize, though, that absent pre-registration, our data analysis choices will be data-dependent, even when they are motivated directly from theoretical concerns. When pre-registered replication is difficult or impossible (as in much research in social science and public health), we believe the best strategy is to move toward an analysis of all the data rather than a focus on a single comparison or small set of comparisons. If necessary, one must step back to a sharper distinction between exploratory and confirmatory data analysis (de Groot, 1956, Tukey, 1977), recognizing the benefits and limitations of each.

In fields where new data can readily be gathered (such as in all four of the examples discussed above), perhaps the two-part structure of Nosek et al. (2013) will be a standard for future research. Instead of the current norm in which several different studies are performed, each with statistical significance but each with analyses that are contingent on data, perhaps researchers can perform half as many original experiments in each paper and just pair each new experiment with a pre-registered replication.

References

Beall, A. T., and Tracy, J. L. (2013). Women are more likely to wear red or pink at peak fertility. *Psychological Science*.

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.

Bem, D. J., Utts, J., and Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology* **101**, 716–719.

Bennett, C. M., Baird, A. A., Miller, M. B., and Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. Poster presented at Human Brain Mapping conference.

Borges, J. L. (1941). El jardin de senderos que se bifurcan. Buenos Aires: Sur. Translated by D. A. Yates (1964), in *Labyrinths: Selected Stories & Other Writings*, 19–29. New York: New Directions.

Box, G. E. P. (1997). Scientific method: The generation of knowledge and quality. *Quality Progress* (January), 47–50.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**, 365–376.

de Groot, A. D. (1956). The meaning of the concept of significance in studies of various types.

Nederlands Tijdschrift voor de Psychologie en Haar Grensgebieden **11**, 398–409. Translated 2013 by E. J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, D. Matzke, D. Mellenbergh, and H. L. J. van der Maas. https://dl.dropboxusercontent.com/u/1018886/Temp/DeGroot_v3.pdf

Durante, K., Arsena, A., Griskevicius, V. (2013). The fluctuating female vote: politics, religion, and the ovulatory cycle. *Psychological Science* **24**, 1007–1016.

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

Gelman, A. (2004). Treatment effects in before-after data. In *Applied Bayesian Modeling and Causal Inference from an Incomplete Data Perspective*, ed. A. Gelman and X. L. Meng, chapter 18. London: Wiley.

Gelman, A. (2013a). Too good to be true. *Slate*, 24 Jul. http://www.slate.com/articles/health_and_science/science/2013/07/statistics_and_psychology_multiple_comparisons_give_spurious_results.html

Gelman, A. (2013b). Is it possible to be an ethicist without being mean to people? *Chance*.

Gelman, A. (2014). The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective. *Journal of Management*.

Gelman, A., Hill, J., and 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., and King, G. (1994). Enhancing democracy through legislative redistricting. *American Political Science Review* **88**, 541–559.

Gelman, A., Shor, B., Bafumi, J., and Park, D. (2007). Rich state, poor state, red state, blue state: What's the matter with Connecticut? *Quarterly Journal of Political Science* **2**, 345–367.

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

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

Greenland, S., and Poole, C. (2013). Living with P-values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology* **24**, 62–68.

Humphreys, M., Sanchez, R., and Windt, P. (2013). Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis* **21**, 1–20.

Monogan, J. E. (2013). A case for registering studies of political outcomes: An application in the 2010 House elections. *Political Analysis* **21**, 21–37.

Mullainathan, S., and Washington, E. (2009). Sticking with your vote: cognitive dissonance and political attitudes. *American Economic Journal: Applied Economics* **1**, 86–111.

Nosek, B. A., Spies, J. R., and Motyl, M. (2013). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science* **7**, 615–631.

Petersen, M. B., Sznycer, D., Sell, A., Cosmides, L., and Tooby, J. (2013). The ancestral logic of politics: Upper-body strength regulates men's assertion of self-interest over economic redistribution. *Psychological Science*.

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

Tracy, J. L.., and Beall, A. T. (2013). Too good does not always mean not true. 30 Jul. University of British Columbia Emotion and Self Lab. http://ubc-emotionlab.ca/2013/07/too-good-does-not-always-mean-not-true/

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *American Statistician* **34**, 23–25.

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

Wagenmakers, E. J., Wetzels, R., Borsboom, D., and van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology* **100**, 426–432.