

# Can we use Bayesian methods to resolve the current crisis of statistically-significant research findings that don't hold up?

Andrew Gelman

Department of Statistics and Department of Political Science  
Columbia University, New York

University of Amsterdam, 30 Oct 2013

# The crisis of non-reproducible research

- ▶ 10 stories
- ▶ 10 principles
- ▶ 3 steps toward a solution

# Story 1: The political attitudes of men with fat arms (Problems of measurement)

## Psychological SCIENCE

A Journal of the  
Association for  
Psychological Science

[Home](#)[OnlineFirst](#)[All Issues](#)[Subscribe](#)[RSS](#) [Email Alerts](#)

## The Ancestral Logic of Politics



### Upper-Body Strength Regulates Men's Assertion of Self-Interest Over Economic Redistribution

Over human evolutionary history, upper-body strength has been a major component of fighting ability. Evolutionary models of animal conflict predict that actors with greater fighting ability will more actively attempt to acquire or defend resources than less formidable contestants will. Here, we applied these models to political decision


## Story 2: ESP (Interactions)

### Journal's Paper on ESP Expected to Prompt Outrage

By **BENEDICT CAREY**

Published: January 5, 2011

One of psychology's most respected journals has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events.

 [Enlarge This Image](#)



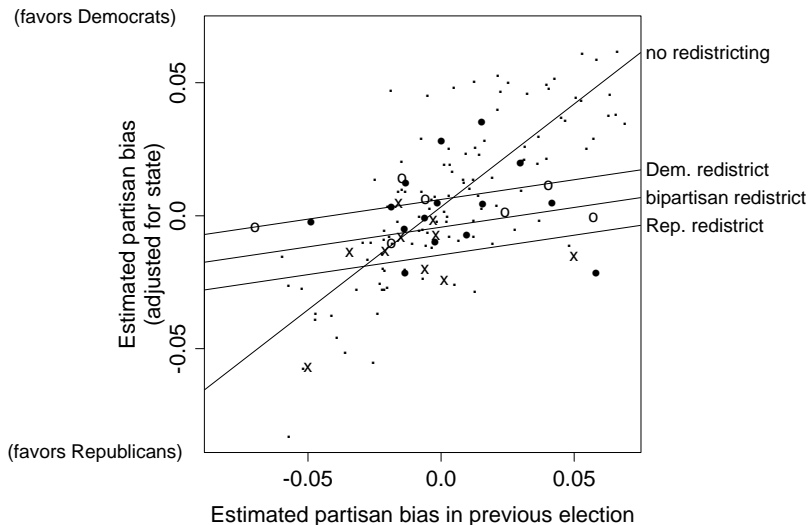
Heather Ainsworth for The New York Times

Work by Daryl J. Bem on extrasensory perception is scheduled to be published this year.

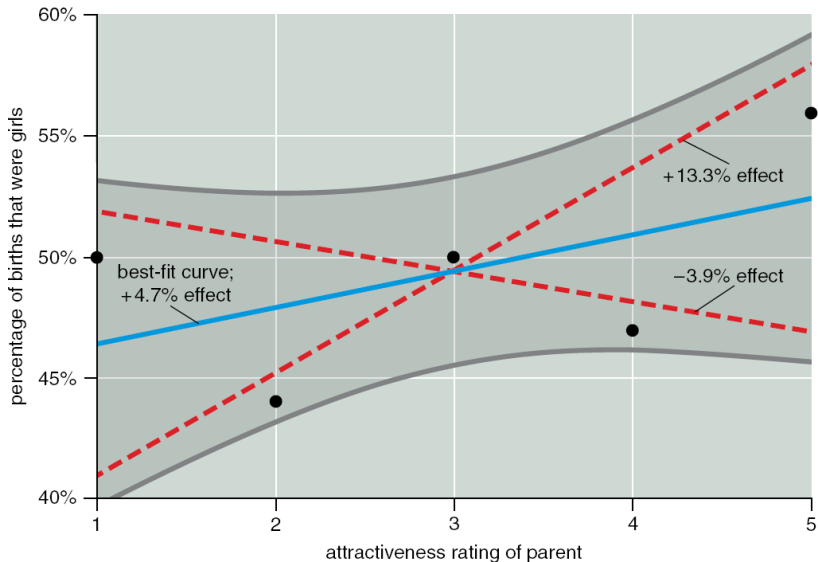
The decision may delight believers in so-called paranormal events, but it is already mortifying scientists. Advance copies of the [paper](#), to be published this year in The Journal of Personality and Social Psychology, have circulated widely among psychological researchers in recent weeks and have generated a mixture of amusement and scorn.

The paper describes nine unusual lab experiments performed over the past decade by its author, [Daryl J. Bem](#), an emeritus professor at Cornell, testing the ability of college students to accurately sense random events,

# Story 3: Effects of redistricting (Interactions)



## Story 4: Beauty and sex ratio (Implausibly large claims)



## Story 5: Ovulation and the color of clothing (Researcher degrees of freedom)

Psychol Sci. 2013 Sep 1;24(9):1837-41. doi: 10.1177/0956797613476045. Epub 2013 Jul 10.

### **Women are more likely to wear red or pink at peak fertility.**

Beall AT, Tracy JL.

University of British Columbia.

#### **Abstract**

Although females of many species closely related to humans signal their fertile via a conspicuous manner, often involving red or pink coloration, no such display has been found for humans. Here, we provide evidence that men are sexually attracted to women wearing or surrounded by red, and that women show a behavioral tendency toward wearing reddish clothing when at peak fertility. In two experiments (samples (N = 124), women at high conception risk were more than 3 times more likely to wear a red or pink shirt than were women at low conception risk, and 77% of women who wore red or pink clothing were found to be at high, rather than low, risk. Conception risk had no effect on the prevalence of red or pink clothing color. Our results thus suggest that red and pink adornment in women is reliably associated with peak fertility, and that female ovulation, long assumed to be hidden, is associated with a salient color display.

## Story 6: Ovulation and voting (Implausibly large claims)

# The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle

**Kristina M. Durante<sup>1</sup>, Ashley Rae<sup>1</sup>, and  
Vladas Griskevicius<sup>2</sup>**

<sup>1</sup>College of Business, University of Texas, San Antonio, and <sup>2</sup>Carlson School of Management, University of Minnesota

### Abstract

Each month, many women experience an ovulatory cycle that regulates fertility. Although the cycle influences women's mating preferences, we proposed that it might also change women's political 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. We found that women with large and diverse families, who are ovulating, had distinctively different effects on political views and voting behavior compared to women who are not ovulating.



## Story 7: Monkeying around (Problems of measurement)

### Marc Hauser Resigns From Harvard



*By Tom Bartlett*

Marc D. Hauser, the Harvard psychologist found responsible for eight counts of misconduct by the university, has ended speculation about whether the embattled professor would return to Harvard this fall.

In a [letter](#) dated July 7, Mr. Hauser told Michael D. Smith, Harvard's dean of the

## Story 8: Sexy research (Fraud)

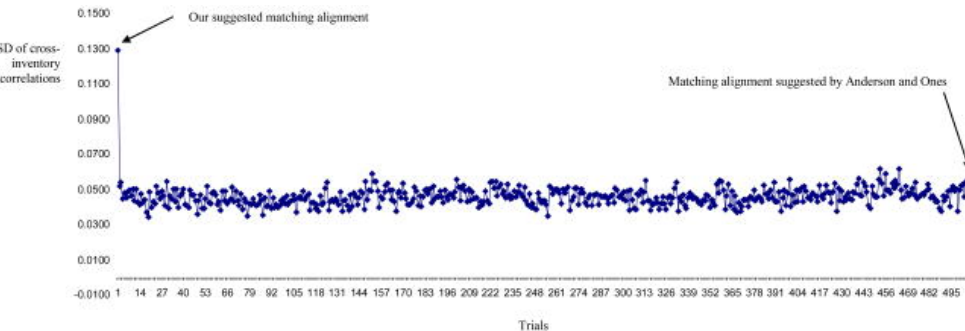
 [Enlarge This Image](#)



Joris Buijs/Pve

The psychologist, Diederik Stapel, committed academic fraud in “sexy” papers, many accepted in respect in the news media, according to a report Monday by the three Dutch institutions that worked: the University of Groningen, the University of Amsterdam, and Tilburg. The journal published one of Dr. Stapel’s papers as an “editorial expression of concern” online on Tuesday.

## Story 9: “No irrefutable proof” (Data processing errors)



## Story 10: Early childhood intervention (Small sample size)

Charles Murray: “To me, the experience of early childhood intervention programs follows the familiar, discouraging pattern . . . small-scale experimental efforts [ $N = 123$  and  $N = 111$ ] staffed by highly motivated people show effects. When they are subject to well-designed large-scale replications, those promising signs attenuate and often evaporate altogether.”

James Heckman: “The effects reported for the programs I discuss survive batteries of rigorous testing procedures. They are conducted by independent analysts who did not perform or design the original experiments. The fact that samples are small works *against* finding any effects for the programs, much less the statistically significant and substantial effects that have been found.

# Bonus story: This week in Psychological Science

October 22, 2013

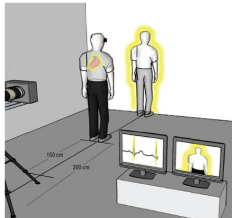


## This Week in *Psychological Science* (TWiPS)

The links below take you to the journal via the APS website. If not already logged in, you will be redirected to log-in using your last name (Gelman) and Member ID (8167).

### [Turning Body and Self Inside Out: Visualized Heartbeats Alter Bodily Self-Consciousness and Tactile Perception](#)

Jane Elizabeth Aspell, Lukas Heydrich, Guillaume Marillier, Tom Lavanchy, Bruno Herbelin, and Olaf Blanke



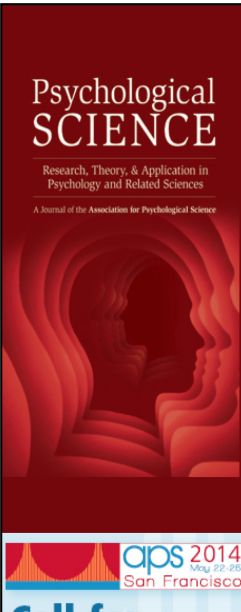
Studies of body perception have mostly focused on manipulations of exteroceptive cues (e.g., vision and touch); however, interoceptive cues (i.e., representations of internal bodily states) may be just as important for self-perception. Participants viewed a virtual body or a rectangle, each of which had a flashing outline that was synchronous or asynchronous with the participant's own heartbeat. Self-identification was stronger for people viewing the virtual body with the synchronous flashing outline than for those viewing the body with the asynchronous flashing outline or for those viewing the rectangles. This suggests that both interoceptive and exteroceptive cues play

important roles in bodily self-perception.

### [Aging 5 Years in 5 Minutes: The Effect of Taking a Memory Test on Older Adults' Subjective Age](#)

Matthew L. Hughes, Lisa Geraci, and Ross L. De Forrest

Subjective age – how old people feel – is related to psychological and physical well-being. In this study, the researchers examined whether common memory-testing procedures influence adults' subjective age. Older and younger adults rated their subjective age before and after taking a memory test. Older adults reported feeling older after taking the memory test, but younger adults did not. A follow-up study found that simply anticipating taking a memory test increased older adults' subjective age. These



# This week in Psychological Science

- ▶ “Turning Body and Self Inside Out: Visualized Heartbeats Alter Bodily Self-Consciousness and Tactile Perception”
- ▶ “Aging 5 Years in 5 Minutes: The Effect of Taking a Memory Test on Older Adults’ Subjective Age”
- ▶ “The Double-Edged Sword of Grandiose Narcissism: Implications for Successful and Unsuccessful Leadership Among U.S. Presidents”
- ▶ “On the Nature and Nurture of Intelligence and Specific Cognitive Abilities: The More Heritable, the More Culture Dependent”
- ▶ “Beauty at the Ballot Box: Disease Threats Predict Preferences for Physically Attractive Leaders”
- ▶ “Shaping Attention With Reward: Effects of Reward on Space- and Object-Based Selection”
- ▶ “It Pays to Be Herr Kaiser: Germans With Noble-Sounding Surnames More Often Work as Managers Than as Employees”

# This week in Psychological Science

- ▶  $N = 17$
- ▶  $N = 57$
- ▶  $N = 42$
- ▶  $N = 7,582$
- ▶  $N = 123 + 156 + 66$
- ▶  $N = 47$
- ▶  $N = 222,924$

# Principle 1: The difference between “significant” and “not significant” is not itself statistically significant

- ▶ Experiment 1:  $25 \pm 10$ : significant!
- ▶ Experiment 2:  $10 \pm 10$ : noise!
- ▶ Difference:  $15 \pm 14.1$ : ...



## Principle 2: Flat priors give inference we can't believe

- ▶ Experiment 2:  $10 \pm 10$ : noise!
- ▶ But, using flat prior,  $\Pr(\text{true effect} > 0) = 0.84!$
- ▶ Epidemiology studies with 95% conf interval  $[1.1, 8.5]$

# Principle 3: Research hypotheses and statistical “hypotheses”

In one case, you want to confirm; in the other, you want to reject.

- ▶ ESP example
- ▶ Fat arms example

# Principle 4: The statistical significance filter

Statistically significant results are overestimates.

- ▶ Beauty and sex ratio example
- ▶ Early childhood intervention example

# Principle 5: Researcher degrees of freedom

It's not just about the file drawer.

- ▶ Fat arms example
- ▶ Redistricting example

# Principle 6: The garden of forking paths

Researcher degrees of freedom can be a problem even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time.

- ▶ ESP example
- ▶ Ovulation examples

# Principle 7: The “That which does not destroy my statistical significance makes it stronger” fallacy

A deterministic intuition that fails when variation is large.

- ▶ Ovulation and clothing example
- ▶ Early childhood intervention example

# Principle 8: The quest for certainty

What do usual research practice, fringe science, unethical scholarship, and fraud have in common?

- ▶ Psychological desire for certainty
- ▶ Incentives for appearing certain
- ▶ The never-back-down attitude
  
- ▶ Psychological Science examples
- ▶ ESP example
- ▶ “No irrefutable proof” example
- ▶ Hauser and Stapel examples

## Principle 9: Type S and Type M errors

- ▶ I've never made a type 1 error in my life
- ▶ I've never made a type 2 error in my life
- ▶ I make Type S (sign) errors
- ▶ I make Type M (magnitude) errors



# Principle 10: Variation and interactions

Interactions are substantively important and surely exist but are difficult to estimate with precision.

- ▶ Monkey example
- ▶ Psychological Science examples

## Solution 0: Open science

- ▶ Public data (including measurement protocols, survey forms, information about data processing and analysis)
- ▶ Publish successful and unsuccessful studies
- ▶ Prominent publication of retractions, criticisms, and replications
- ▶ Replication with preregistered protocols in psychology, political science, etc.

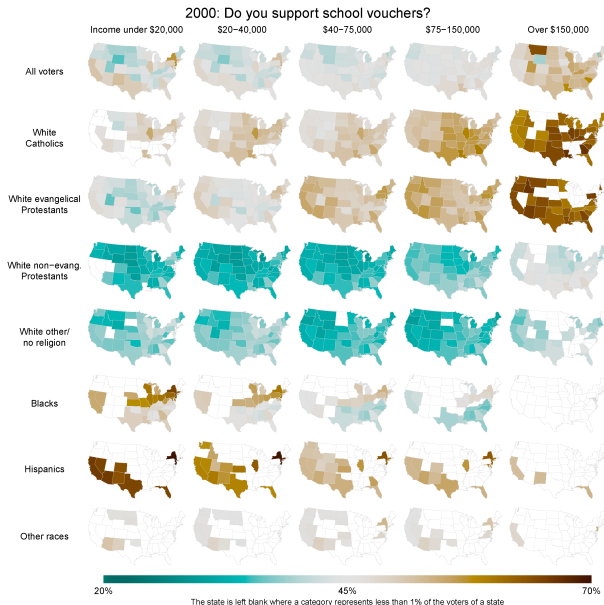
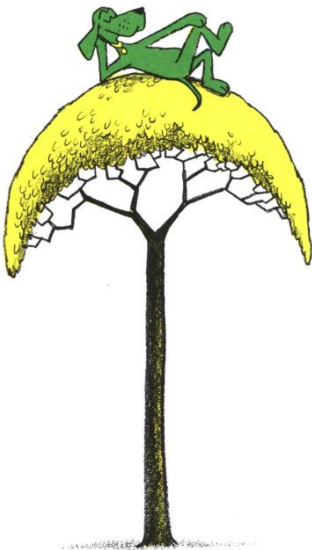
# Solution 1: Design calculations

- ▶ Generalizing the concept of “power analysis”
- ▶ Estimate: beautiful parents are 4.7 percentage points more likely to have girls (with standard error of 4.3):
- ▶ Suppose the true effect was 0.3%
- ▶ Retrospective design calculation:
  - ▶ 3% probability of a statistically-significant positive result
  - ▶ 2% probability of a statistically-significant *negative* result
  - ▶ Type S error rate is 40%
  - ▶ Type M inflation factor is at least  $\frac{1.96*4.3\%}{0.3\%} = 28$

## Solution 2: Informative priors

- ▶ Can implement using Bayes or design calculation
- ▶ Sex ratio example: effect in the range  $(-0.3\%, +0.3\%)$
- ▶ Ovulation and voting example: effect in the range  $(-2\%, 2\%)$
- ▶ Three sorts of prior belief:
  - ▶ Effect is near 0 (most things don't work, attenuation due to measurement error, etc.)
  - ▶ Effect is positive (researcher's belief)
  - ▶ Effect is negative (bias toward pessimism)

# Solution 3: Hierarchical models for interactions



# Data don't always “speak for themselves”

