# STATS 604
# Reproducibility and replicability

Jonathan Terhorst

# Reproducibility

- Would a researcher obtain the same results if they used the same data, methods, and code?
- What could cause this to fail?
- What about same data and methods but not code?
- What about same data only?
- What about different data/methods/code?

# Replicability

- Would a researcher obtain the same results if they used **different** data/methods/code while attempting to answer the same scientific question?
    - Re-analyze the same data, but follow your own path of inquiry.
    - Or completely re-run the study, collecting a new data set, and analyze it.
- Two studies are said to replicate if they obtain consistent results in spite inherent uncertainty in the system under study.

# Researcher degrees of freedom

- Every statistical investigation requires you to make a sequence of choices about how to conduct, analyze and report the analysis.
  - Data collection: how much data to collect?
  - Analysis decisions: how are variables defined? What tests are used? What models are considered?
  - What hypotheses will be tested? What results will be reported? Will multiple testing be considered?
- Maliciously exploiting these degrees of freedom is known as p-hacking.
- But even well-intentioned studies can mistakenly fall prey.

# Researcher DOF

- Study 1: 30 UPenn undergrads were randomly assigned to listed to the control song ("Kalimba", an instrumental that comes with Windows 7) or a children's song ("Hot Potato," performed by The Wiggles.
    - Result: eople felt older after listening to "Hot Potato" vs. the control, p=0.033.
    - Conclusion: listening to a kid's song makes you feel older.
- Study 2: 20 ugrads listened to "When I'm 64" (Beatles) or Kalimba, and were also asked their bday and father's age.
    - Result: People were ~1.5 years younger when listening to "When I'm 64" (p=0.040).
    - Conclusion: listening to an oldie causes you to become older.



*General Article*

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

## False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

**Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]**
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

**Abstract**

In this article, we accomplish two things. First, we show that despite empirical psychologists' nominal endorsement of a low rate of false-positive findings (≤ .05), flexibility in data collection, analysis, and reporting dramatically increases actual false-positive rates. In many cases, a researcher is more likely to falsely find evidence that an effect exists than to correctly find evidence that it does not. We present computer simulations and a pair of actual experiments that demonstrate how unacceptably easy it is to accumulate (and report) statistically significant evidence for a false hypothesis. Second, we suggest a simple, low-cost, and straightforwardly effective disclosure-based solution to this problem. The solution involves six concrete requirements for authors and four guidelines for reviewers, all of which impose a minimal burden on the publication process.

**Keywords**

methodology, motivated reasoning, publication, disclosure

Received 3/17/11; Revision accepted 5/23/11

Our job as scientists is to discover truths about the world. We generate hypotheses, collect data, and examine whether or not the data are consistent with those hypotheses. Although we aspire to always be accurate, errors are inevitable.

Perhaps the most costly error is a *false positive*, the incorrect rejection of a null hypothesis. First, once they appear in the literature, false positives are particularly persistent. Because null results have many possible causes, failures to replicate previous findings are never conclusive. Furthermore, because it is uncommon for prestigious journals to publish null findings or exact replications, researchers have little incentive to even attempt them. Second, false positives waste resources: They inspire investment in fruitless research programs and can lead to ineffective policy changes. Finally, a field known for publishing false positives risks losing its credibility.

In this article, we show that despite the nominal endorsement of a maximum false-positive rate of 5% (i.e., $p \leq .05$), current standards for disclosing details of data collection and analyses make false positives vastly more likely. In fact, it is unacceptably easy to publish "statistically significant" evidence consistent with *any* hypothesis.

The culprit is a construct we refer to as *researcher degrees of freedom*. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?

It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields "statistical significance," and to then report only what "worked." The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.

This exploratory behavior is not the by-product of malicious intent, but rather the result of two factors: (a) ambiguity in how best to make these decisions and (b) the researcher's desire to find a statistically significant result. A large literature documents that people are self-serving in their interpretation

**Corresponding Authors:**
Joseph P. Simmons, The Wharton School, University of Pennsylvania, 551 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104
E-mail: jsimmo@wharton.upenn.edu

Leif D. Nelson, Haas School of Business, University of California, Berkeley, Berkeley, CA 94720-1900
E-mail: leif_nelson@haas.berkeley.edu

Uri Simonsohn, The Wharton School, University of Pennsylvania, 548 Jon M. Huntsman Hall, 3730 Walnut St., Philadelphia, PA 19104
E-mail: uws@wharton.upenn.edu

# How it works

**Using the same method as in Study 1, we asked ~~20~~ 34 University of Pennsylvania undergraduates to listen** only **to either "When I'm Sixty-Four" by The Beatles or "Kalimba"** or "Hot Potato" by the Wiggles. We conducted our analyses after every session of approximately 10 participants; we did not decide in advance when to terminate data collection. **Then, in an ostensibly unrelated task, they indicated** only **their birth date (mm/dd/yyyy) and** how old they felt, how much they would enjoy eating at a diner, the square root of 100, their agreement with "computers are complicated machines," **their father's age**, their mother's age, whether they would take advantage of an early-bird special, their political orientation, which of four Canadian quarterbacks they believed won an award, how often they refer to the past as "the good old days," and their gender. **We used father's age to control for variation in baseline age across participants.**

# Durante et al. (2013)

- During the fertile phase of their menstrual cycle:
  - Single women were more likely to be politically liberal and less likely to be religious.
  - Women in relationships were more likely to be conservative.
  - Hypothesis: fertility might be influencing behavior to increase reproductive success.
- Thoughts?



*Research Article*

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
24(6) 1007–1016

DOI: 10.1177/0956797612466416
pss.sagepub.com
SAGE

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

**Kristina M. Durante[1], Ashley Rae[1], and Vladas Griskevicius[2]**
[1]College of Business, University of Texas, San Antonio, and [2]Carlson School of Management, University of Minnesota

**Abstract**
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.

**Keywords**
evolutionary psychology, fertility, relationships, political attitudes, religiosity, religious beliefs

Received 7/30/12; Revision accepted 10/5/12

Women are more likely to vote than men are (Negrin, 2012), which makes the female vote pivotal for anyone seeking political office. In the 2012 U.S. presidential election campaign, for example, both Republican Party candidate Mitt Romney and Democratic Party candidate Barack Obama went to great lengths to court female voters (Edwards-Levy, 2012). Each candidate appeared to be succeeding—sort of. Romney, the more conservative candidate, was strongly favored by married women, holding a 19% edge over Obama. But Obama, the more liberal candidate, was strongly favored by single women, holding a 33% edge over Romney (Knox, 2012). What might have been the source of this political divide?

We considered whether this difference might in part be related to a surprising biological factor—women's monthly ovulatory cycle. Building on the idea that reproductive goals might drive political and religious attitudes (Kurzban, Dukes, & Weeden, 2010; Li, Cohen, Weeden, & Kenrick, 2009; Weeden, Cohen, & Kenrick, 2008), we examined whether hormonal fluctuations associated with fertility influence women's politics, religiosity, and voting. In two studies using large and diverse samples of women, we tested how the ovulatory cycle influences religious and political orientation for single women and women in committed relationships. In addition, we tested whether changes in political ideology mediated women's voting preferences in the 2012 U.S. presidential race.

### Political and Religious Ideology

Political attitudes vary on a fundamental liberal-conservative (or left-right) dimension (Haidt, 2012; Jost, Glaser, Kruglanski, & Sulloway, 2003). Liberalism is characterized by advocacy for social change and a rejection of social inequality, whereas conservatism is characterized by a

**Corresponding Author:**
Kristina M. Durante, Department of Marketing, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249
E-mail: kristina.durante@utsa.edu

# Criticisms

## Causality

**Abstract**

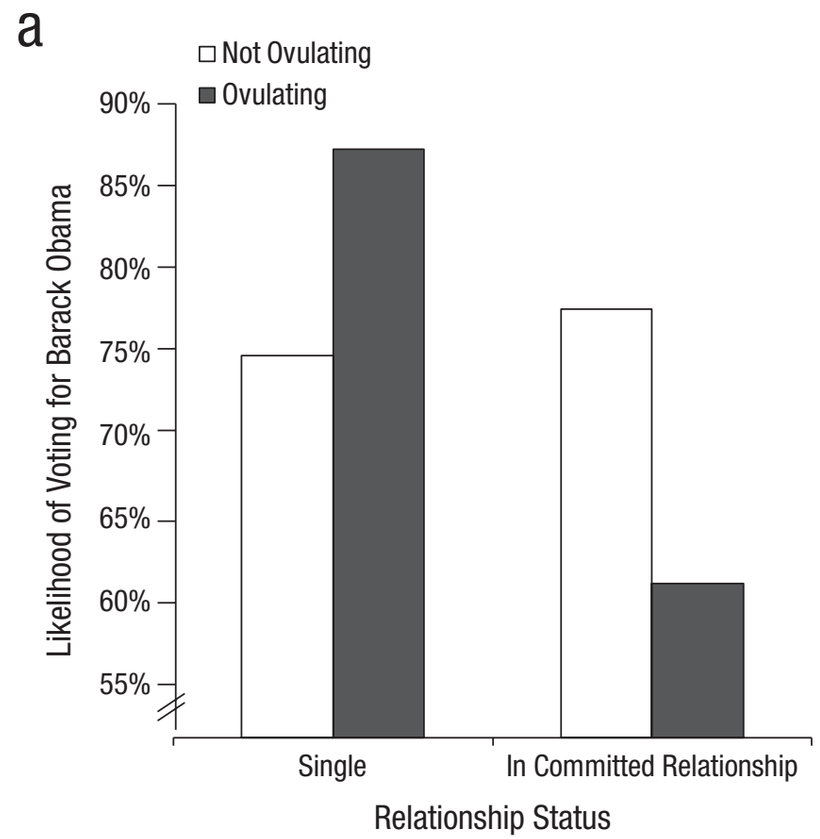
Each month, many women experience an ovulatory cycle that regulates fertility. Although research has found that this cycle influences women's mating preferences, we proposed that it might also change women's political and religious views. Building on theory suggesting that political and religious orientation 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.

- Language strongly suggests of causality.
- Does the paper establish that?
- Does the paper compare the same women at different stages of their cycle?

# Criticisms

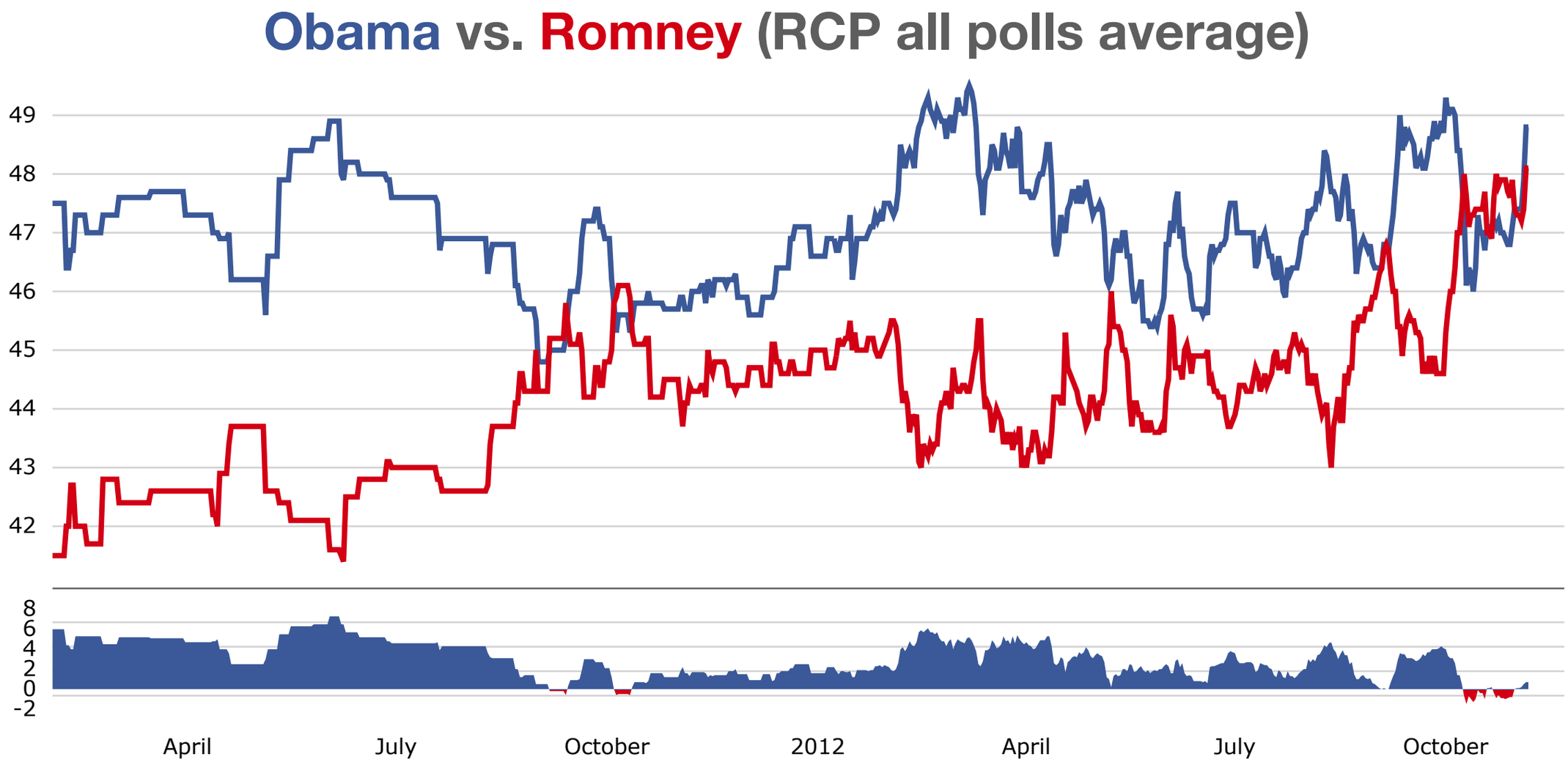
## Effect sizes

- Raw effect sizes of ~17% were reported.
- (Can you think of an intervention that would make a person 17% more likely to vote for the other side?)
- The std. effect size here is $\text{qt}(1 - .035/2, 134) = 2.1$.
- $s = 17\,\%/2.1 = 8.1\,\%$.




***Voting preferences.*** Single women were more likely to vote for Barack Obama (79.3%) than were women in relationships (69.4%), $\chi^2(1, N = 502) = 3.88$, $p = .049$. However, a logistic regression revealed that this main effect was qualified by a Fertility × Relationship Status interaction, $b = -1.62$, Wald(1) = 8.35, $p = .004$. Single women were more likely to vote for Obama if they were in the high-fertility group (86.5%) than if they were in the low-fertility group (73.7%), $\chi^2(1, N = 169) = 4.15$, $p = .042$, $d = 0.32$. Women in relationships, however, were more likely to vote for Romney at if they were in the high-fertility group ($M = 40.4\%$) than if they were in the low-fertility group (23.4%), $\chi^2(1, N = 134) = 4.44$, $p = .035$, $d = 0.37$ (see Fig. 4a).

# Sanity check




https://www.realclearpolitics.com/epolls/2012/president/us/general_election_romney_vs_obama-1171.html

# Criticisms

## Researcher DOF

- The study makes numerous seemingly arbitrary decisions about variable definitions, hypothesis tests, etc.
- How to define/operationalize complex concepts like "religiosity", "political beliefs", or even "fertility"?
- What other hypothesis were tested but not reported?

# Study 2

## Ovulation, Political Attitudes, and Voting Attitudes

- 502 women with a mean age of 27.3 years (SD= 6.14, range = 18–42 years)

- Regular monthly menstrual cycles (25–35 days) and were not using hormonal contraception.

- 54.6% were single (not dating or casually dating), and 45.4% were in a committed relationship (engaged, living with a partner, or married).

# Assessing fertility

- Classified in a high or low fertility group based on their cycle day:
  - High fertility: 7-14 days; low fertility: 17-25; everyone else excluded.
  - Cycle day = days before next onset.
  - Cycle length = difference between the start of last menstrual period and previous one.
- How does binning into groups affect the analysis?
- How does filtering affect the analysis?
- Why compute cycle length instead of using reported cycle length?

***Assessing fertility.*** We obtained from participants (a) the start date of their last menstrual period and the previous menstrual period, (b) the expected start date of their next menstrual period, and (c) the typical length of their menstrual cycle. We then used the established reverse-cycle-day method to predict the day of ovulation for each participant (DeBruine, Jones, & Perrett, 2005; Durante et al., 2011; Haselton & Gangestad, 2006).

On the basis of this established method, we created a high-fertility group (cycle days 7–14, $n$ = 78) and a low-fertility group (cycle days 17–25, $n$ = 85). For our main analyses, we did not include women on cycle days 15 and 16 because of the difficulty of determining fertility status on these days via counting estimates (DeBruine et al., 2005; Haselton & Gangestad, 2006). We also did not include women at the beginning of the ovulatory cycle (cycle days 1–6) or at the end of the ovulatory cycle (cycle days 26–28) to avoid potential confounds due to premenstrual or menstrual symptoms.

# Relationship status

- Relationship options:
  1. Not dating;
  2. Dating one partner;
  3. Engaged and/or cohabiting;
  4. Married.
- #1 and #2 grouped to single; #3/#4 grouped into "committed relationship".
- Why this choice of groupings?
- What effect would other choices have?

***Relationship status.*** Participants indicated their current relationship status by selecting one of the following five descriptions: "not currently dating or romantically involved with anyone" (26.3%), "dating" (26.9%), "engaged or living with my partner" (14.7%), "married" (30.7%), or "other" (1.4%). Participants who indicated that they were engaged, living with a partner, or married were classified as being in a committed relationship ($n$ = 228), and all others were classified as being single ($n$ = 274). As in Study 1, if a participant selected "other," she was prompted to provide a descriptor for her current relationships status so that we could accurately assign her to a relationship category.

```
> library(multiverse)
> data(durante)
> ?durante
```

All questions were preceded by the prompt "Please indicate how much you agree with the following statements".

The following items were responses to religiosity items (on a scale of 1 - 9): *Rel1, Rel2, Rel3*.

The following items were responses to fiscal political attitudes items (on a scale of 1 - 7): *RichTax*, *TooMuchProfit*, *StandardLiving*, *FreeMarket*, *PrivSocialSec*

The following items were responses to social political attitudes items (on a scale of 1 - 7): *Abortion*, *Marriage*, *StemCell*, *Marijuana*, *RestrictAbortion*

# Effect size and power

# Exercise

Let $X \sim N(\mu, 1)$ and suppose you are testing

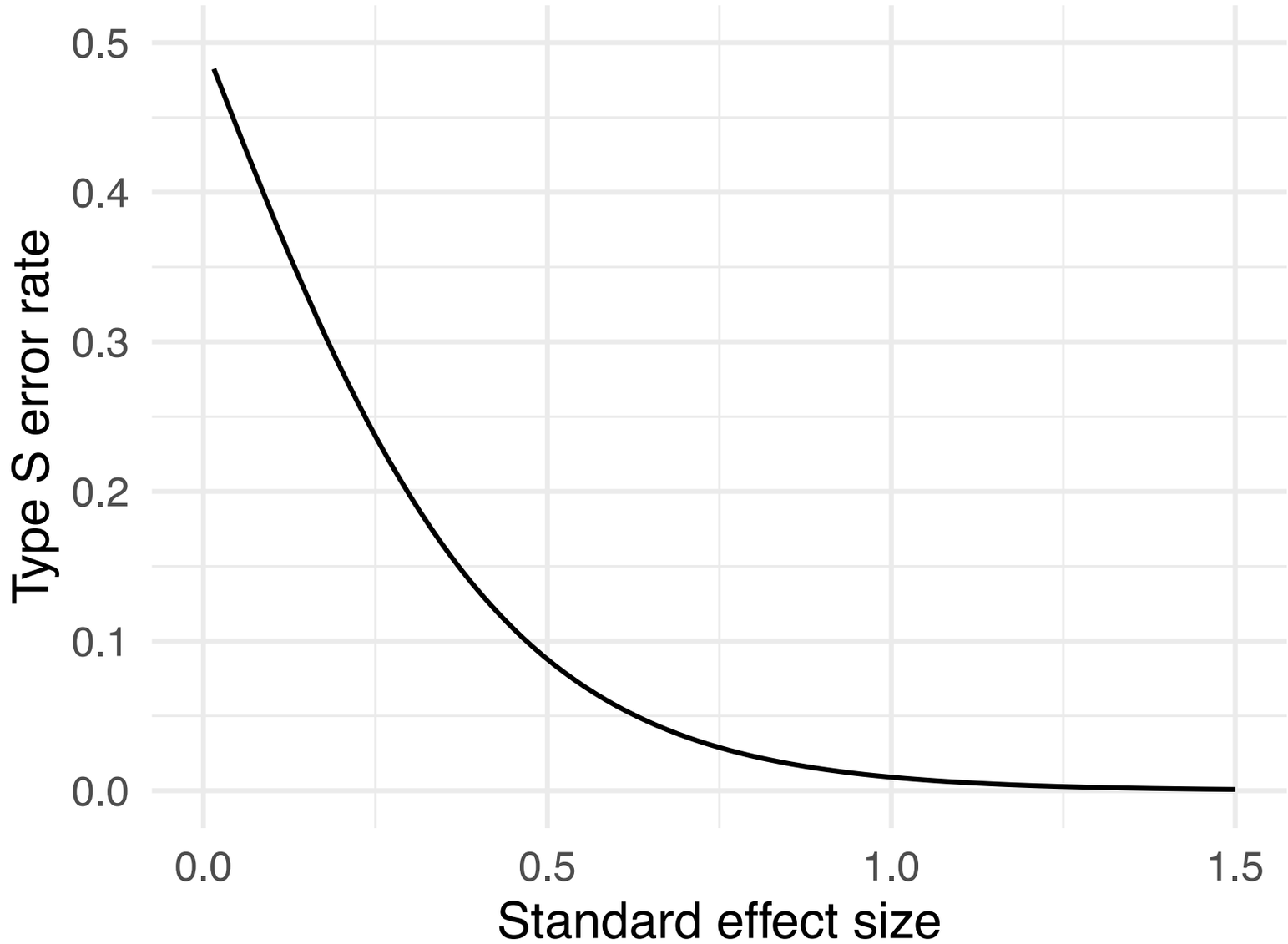
$$H_0 : \mu = 0 \text{ vs. } H_A : \mu \neq 0.$$

Suppose the true $\mu > 0$.

As a function of $\mu$, what is the probability that you reject $H_0$ at the 95% level *and* that $X < 0$?

(Make a plot, for example.)

```r
p.type.S <- function(mu) {
   a <- pnorm(-1.96, mean=mu)
   b <- 1 - pnorm(1.96, mean=mu)
   b / (a+b)
}
```



# Type S error

- In the preceding plot, $\mu$ was the standardized effect size.
- For effect sizes in the $.1\sigma - .2\sigma$ range, there is a 30-40% chance that a statistically significant finding **does not even get the sign correct**.
- This is referred to as *Type S* error.

Gelman, Andrew, and John Carlin. 2014. “Beyond Power Calculations: Assessing Type S (sign) and Type M (magnitude) Errors.” *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 9 (6): 641–51.

# Exercise

As a function of $\mu$, what is $\mathbb{E}_\mu(\,|X/\mu|\,\mid \text{reject } H_0)$?

(Same setup as before.)

```r
E.type.M <- function(mu) {
  # E_mu(|X/mu| | |X| > 1.96) = E_mu(sqrt(X^2) | X^2 > 1.96^2) / mu
  # where X^2 follows noncentral chi-squared distribution.
  lam <- mu^2
  m <- 1 - pchisq(1.96^2, df = 1, ncp = lam)
  mean(qchisq(runif(n = 10000, min = m), df = 1, ncp = lam)) / abs(mu)
}
```

# Type M error

- Low-powered experiments will tend to drastically inflate effect sizes.
- Gelman & Carlin refer to this as the “exaggeration ratio”.

# Typical effect sizes

- What sort of effect sizes do we expect to see in practice?
- Much depends on your field, but generally, the effect size is often smaller than you (the researcher) ~~hope~~ expect.
- The literature can offer a guide.

One can reasonably conclude that any average differences among women at different parts of their menstrual cycle would be small… We would consider an **effect size of 2 percentage points** to be on the upper end of plausible differences in voting preferences.

—**Gelman and Carlin (2015)**

# Exercise

Suppose the true effect size is 2%, with experimental standard error as reported by Durante *et al*. (and calculated three slides ago).

1. What is the power?
2. What is the probability of a type S error?
3. What is the expected exaggeration ratio?

# Solution

- The true effect size is $2\,\%/8.1\,\% = .25$ standard errors.
- Power = $1 - \Phi(1.96 - .25) + \Phi(-1.96 - .25) = .057$.
- p.type.S(.25) = 0.236
- E.type.M(.25) = 3.5

🤔 I get a different expected exaggeration ratio than the article…