View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by ScholarlyCommons@Penn

University of Pennsylvania
ScholarlyCommons

Operations, Information and Decisions Papers

Wharton Faculty Research

2014

P-curve: A Key to The File Drawer

Uri Simonsohn
University of Pennsylvania

Leif. D. Nelson

Joseph. P. Simmons
University of Pennsylvania

Follow this and additional works at: http://repository.upenn.edu/oid_papers

Part of the Other Psychology Commons, and the Other Statistics and Probability Commons

Recommended Citation

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A Key to The File Drawer. *Journal of Experimental Psychology: General, 143* (2), 534-547. http://dx.doi.org/10.1037/a0033242

This article may not exactly replicate the final version published in the APA journal. It is not the copy of record.

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/oid_papers/231
For more information, please contact repository@pobox.upenn.edu.

P-curve: A Key to The File Drawer

Abstract
Because scientists tend to report only studies (publication bias) or analyses (p-hacking) that "work," readers must ask, "Are these effects true, or do they merely reflect selective reporting?" We introduce p-curve as a way to answer this question. P-curve is the distribution of statistically significant p values for a set of studies (ps .05). Because only true effects are expected to generate right-skewed p-curves— containing more low (.01s) than high (.04s) significant p values— only right-skewed p-curves are diagnostic of evidential value. By telling us whether we can rule out selective reporting as the sole explanation for a set of findings, p-curve offers a solution to the age-old inferential problems caused by file-drawers of failed studies and analyses.

Keywords
publication bias, selective reporting, p-hacking, false-positive psychology, hypothesis testing

Disciplines
Other Psychology | Other Statistics and Probability

Comments
This article may not exactly replicate the final version published in the APA journal. It is not the copy of record.

This journal article is available at ScholarlyCommons: http://repository.upenn.edu/oid_papers/231

Journal of Experimental Psychology: General

This version: 2015 09 22

SUPPLEMENTARY MATERIALS

***P*-curve: A Key to The File Drawer**

Uri Simonsohn
University of Pennsylvania
The Wharton School
uws@wharton.upenn.edu

Leif D. Nelson
UC Berkeley
Haas School of Business
Leif_nelson@haas.berkeley,edu

Joseph P. Simmons
University of Pennsylvania
The Wharton School
jsimmo@wharton.upenn.edu

OUTLINE:

1) A technical framework for *p*-curve and computing *pp*-values for null of 33% power. (p.3-9)

2) Is *p*-curve uniform under the null if variables are not normally distributed? (p.10-11)

3) Modeling *p*-hacking. (p.12-13)

4) The problem with discrete tests, and, how bad is it to ignore it? (p.14-17)

5) Selection of JPSP studies and *p*-values for the demonstration. (p.18-21)

6) Other statistical tests applicable to *p*-curve. (p.22)

7) Selection bias as an alternative explanation to *p*-hacking for the ANCOVA example. (p.23-24)

1) A technical framework for *p*-curve and computing *pp*-values for null of 33% power.

1.1. P-curve and noncentral distributions

P-curve is closely related to statistical power. Power is the probability of a statistical test obtaining a *p*-value<α, where α is typically 5%. One can think of the distribution of significant *p*-values, *p*-curve, as computing power for every possible α between 0 and .05.[1]

Power calculations rely on "*noncentral"* distributions, so *p*-curve calculations rely on noncentral distributions also. Noncentral distributions are seldom covered by statistics textbooks for non-statisticians, so we provide a brief introduction to them here. For a more complete yet still accessible introduction see the article by Cumming and Finch (2001), beginning with the last paragraph of page 546.

Central vs. noncentral distributions

The central distribution captures how a test statistic is distributed when the null of no difference is true. The noncentral distribution captures how a test statistic is distributed when the null is *not* true. When we speak of "the" student distribution, then, we actually mean the *central* student distribution. The central student distribution is used, for example, to assess how likely a given difference of sample means would be if the true population means were the same. With the noncentral distribution, in contrast, we ask how likely a given difference would be if the true means were *not* the same.

For instance, when the results section of a paper reads, "`the means were significantly different, t(38)=2.024, p=.05,`" this indicates that, if the true means were identical, there is only a 5% chance that the two sample means would differ by the observed 2.024 standard errors or

[1] Though note that *p*-curve is the density rather than the *c.d.f.* and it only considers *p*<.05, so it is the probability of obtaining a given *p*-value conditional on it being <.05.

more. So unless otherwise stated, it implies that t(38) is evaluated for the central t distribution, so $t_{central}(38)=2.024$, $p=.05$. With the noncentral distribution, we instead ask how likely is $t(38)\geq 2.024$ if the true means differed not by 0, but by, say, one standard deviation from each other (d=1).

The noncentrality parameter (ncp)

While the shape of the central student distribution varies only as a function of the degrees of freedom (d.f.) of the test, that of the noncentral distribution is also a function of what's referred to as the "noncentrality parameter" (ncp), which in turn is a function of sample size and effect size. For the student distribution, ncp=$\sqrt{\frac{n}{2}}d$. This makes intuitive sense: If we want to know how likely a given observed difference of means is, if the population means differ by some amount *d*, we need to take into account what that amount is (*d*) and how big our sample is (*n*).

So for example, for a differences of means t-test performed on two samples of n=20 each, and (true) effect size of *d*=1, the resulting test statistic is distributed t, df=38 and ncp=$\sqrt{\frac{20}{2}}1$=3.16. We hence can answer the question: What is the probability that t(38) will be greater than 2.024 given a true effect size of d = 1, by evaluating the noncentral t(38), with ncp=3.16, at t=2.024. We do this the same way we find values for the central distribution, looking it up in a table, or running software that has access to the formulas behind those tables.

For example, to find the *p*-value associated with $t_{central}(38)=2.024$ we can look up a student table with 38 degrees of freedom, or rely on Excel's **tdist()** function, or rely on R's ***pt()*** function, etc. Because Excel does not have noncentral distributions built in (as of 2013) we will use R syntax (with detailed explanations) for the remainder of this supplement.

The *pt()* function in R gives the *c.d.f.* for a given t-statistic (that is, the probability of obtaining a value smaller than that t-statistic). Its syntax is *pt(q,df,ncp)*, where q is the value of the t-test we are

looking up. To find the probability of obtaining a value *larger* than that t-statistic, you simply subtract that formula from 1: *1 – pt(q,df,ncp).*

Thus, to find the (one-sided) *p*-value associated with $t_{central}(38)=2.024$, we would use the following r formula:

```
1-pt(q=2.024,df=38,ncp=0)
```

= .025. [2]

This is the probability of finding t>2.024 given ncp = 0, which is equivalent to the probability of finding t>2.024 when the null is true (i.e., using the central distribution). This is the one-tailed probability, and so the two-tailed probability can be obtained by multiplying by 2, which equals .05. This example shows that, when df = 38, the t = 2.024 represents the threshold for statistical significance (.05).

Now let's say that we are interested in knowing how likely we are to obtain a t-value greater than 2.024 if n = 20 and d = 1, and hence when ncp = 3.16. We would use the following formula:

```
1-pt(q=2.024,df=38,ncp=3.16)
```

= .869.

This indicates that there is an 86.9% chance of obtaining a t-value greater than 2.024. Because 2.024 is the threshold for statistical significance, we can say that, given d = 1 and n = 20, there is an 86.9% chance of obtaining a statistically significant result, and thus the "power" of this experiment is equal to 86.9%. This is precisely how power calculation software uses effect size estimates to generate recommended sample sizes or estimated power.

[2] This page includes two corrected typos identified by Ellen Evers. Corrections took place on 2013/12/12

If we are instead interested in knowing how likely our t-test is to result in $p \leq .04$ rather than $p \leq .05$, we would simply look up the value of $t_{central}(38)$ that produces p=.04. This value is 2.126. We now enter the formula:

```
1- pt(q=2.126,df=38,ncp=3.16)
```

= .8457. This indicates that there is an 84.57% chance of obtaining a t-value greater than 2.126, and thus an 84.57% chance of obtaining $p \leq .04$.

Now, if there is an 86.9% chance of $p \leq .05$, and there is a 84.6% chance of $p < .04$, then the chance of $.04 < p \leq .05$ is 86.9%-84.6%=2.3%. Hopefully, the relationship between ncp, power, noncentral distributions, and *p*-curve just became obvious.

Because the distribution of *p*-values under the alternative is a function only of the noncentral distribution, which is itself a function only of sample size and effect size, *p*-curve is a function only of sample size and effect size. If we know the effect size and sample size, we know the expected *p*-curve; we know how likely each *p*-value is for any given effect size.

1.2 Computing pp-values under the null of 33% power

In the main text we introduce *pp*-values to test the significance of the deviation from an observed *p*-curve to a null *p*-curve. For the null of a uniform *p*-curve, *pp*-values are trivial to compute. They involve "stretching" the [0-.05] into [0-1] by multiplying *p*-values by 20. For example, among significant *p*-values, $p \leq .04$ is obtained 20*.04=80% of the time under the null of d=0, so *pp=.8*. In light of the previous discussion it is worth highlighting that we do not rely on noncentral distributions to test the uniform null, because we are still testing the null of no effect (*d* = 0) and that involves the central distribution.

For computing *pp*-values under the null that a test is powered to 33%, on the other hand, we do need to rely on a noncentral distribution. In particular, we rely on the noncentral distribution with a

noncentrality parameter (again: ncp) leading the observed test to have 33% power. This means that to compute *pp*-values for the null of 33% power for study *i*, arising from a given $t(df_i)=x_i$ test, we may follow these three steps:

Step 1. Find the critical value of the student distribution, X_i, for which $t_{central}(df_i)=X_i$, p=.05.

Step 2. Find the ncp_i for the $t_{ncp}(df_i)$ student distribution that has a 33% chance of obtaining $x \geq X_i$.

Step 3. Evaluate the observed x_i with the noncentral t with ncp_i.

Let's consider a concrete example. Imagine that a study's key test was t(38)=2.126, p=.04. To compute its *pp*-value under the null of 33%, we begin by finding the critical X_i for which it is true that t(38)=Xi, p=.05. We can find Xi using the qt(p,df,ncp) function in R. Using the formula, *qt(.975,df=38,ncp=0)* tells us that the critical t-value for one-tailed p = .025 (and hence two-tailed p = .05), the t-value that exceeds 97.5% of the values under the null, is 2.024.[3] We now need to find the ncp that would make t>2.024 have 33% chance. This function is not built into R but is easy to build it (in SAS it does exist, it is call TNONCT). We want to ask R something like: Hey R, why don't you go find the value of ncp such that: pt(x=2.024, df=38, ncp=???)=67%?

R can do this in three lines of code:[4]

```
f <- function(delta, pr, x, df) pt(x, df = df, ncp = delta) - pr
out <- uniroot(f, c(0, 37.62), pr =2/3, x = 2.024, df = 38)
out$root
```

The output this code produces is 1.568436. So R just told us that if one were to run

1-pt(2.024, df=38, ncp=1.568436),

one would obtain

.3333333.

[3] R's pt() function is like Excel's tdist(), and R's qt() function is like Excel's tinv()). The key difference is that R accommodates noncentral t's and Excel does not.

[4] That c(0,37.62) command is there because that's the range of the noncentral parameter which R is able to compute; ncp>37.62 just would not work, and ncp>37.62 are way too big for our purposes anyway.

This means that ncp=1.568436 is the noncentrality parameter that leads a test with df=38 to have 33% power. Now we ask R how likely it is to observe t<2.126 if ncp=1.568436 using the formula:

pt(2.126, df = 38, ncp = 1.568436)

= . 70127. That's the probability of t(38)<2.126 (and thus *p*>.04), but the *pp*-value is probability of obtaining *p*>.04 *given* that that we have observed a *p*-value less than .05. To get that value, we first subtract 2/3 from the above probability; because 2/3 is the probability of p>.05 (since power is 1/3, there is a 2/3 chance of p>.05), this subtracts out the probability of observing p>.05. We then divide by 1/3, or equivalently multiply by 3, because we are conditioning on being in one third of possible values. In short, the formula for computing a pp-value for 33% power given df = 38 and *p* = .04 is 3*(.70127-2/3)= .10. Thus, a two sample t-test with n=20 per cell has a 10% chance of obtaining p>.04 conditioning on the result being significant and the test being powered at 33%.[5]

[5] We thank Chad Danyluck for alerting us of some typos in this paragraph in an earlier version of this supplement. The typos have been corrected (August, 2013).

The code that follows creates a function in R that computes *pp*-values, for the 33% power null, for a t-test with degrees of freedom df_ and observed t=x_

```
##############################################
pp33 <-function(df_,x_) {
    #Find critical value of student (xc) that gives p=.05 when df=df_
        xc=qt(p=.975, df=df_)

    #Find noncentrality parameter (ncp) that leads 33% power to obtain xc
        f <- function(delta, pr, x, df) pt(x, df = df, ncp = delta) - pr
        out <- uniroot(f, c(0, 37.62), pr =2/3, x = xc, df = df_)
        ncp_=out$root

    #Find probability of getting x_ or larger given ncp
        p_larger=pt(x_,df=df_,ncp=ncp_)

    #Condition on p<.05 (i.e., get pp-value)
        pp=3*(p_larger-2/3)

    #Print results
     return(pp)
    }
 ##############################################
```

So for example, the last two pages of explanations looking for the *pp*-value of t(38)=2.126 can now be performed with the following invocation of the new function:

```
pp33(df=38,x=2.126)
```

resulting in

.1034

the *pp*-value of *p*=.04 for df=38 is *pp*=.1034.

(2) Is *p*-curve uniform if variables are not normally distributed?

In the paper we assume that the assumptions underlying the statistical tests of interest (e.g., the two-sample *t*-test) are met. We focus on the *t*-distribution (and hence the *F* distribution with $df_1=1$), which assumes that the underlying random variables are normally distributed. The literature contains several demonstrations of the robustness of the *t*-test to deviations from normality (Boneau, 1960; Pearson, 1931); nevertheless, we conducted simulations to verify *p*-curve's robustness to non-normality.

We created two small samples (n=15) drawn from the same population, conducted a t-test on them, and repeated this procedure several thousand times, tabulating how frequently we observed *p*-values in each of the five bins ($p<.01$, $.01<p<.02$, etc.). We simulated data using distributions that deviated from a normal distribution by an increasing amount: normal, uniform-continuous (0-1), uniform-discrete taking just four values (0.25, 0.5, 0.75 and 1), and a Poisson with $\lambda=2$ truncated at 1 and 4. The truncated Poisson leads to a distribution where *y* takes the values 1,2,3,4 with approximate probabilities .4, .3, .2 and .1, respectively. Figure S1 shows that, despite the severe deviations from normality and small sample sizes ($n=15$ per cell), *p*-curve is quite close to uniform for the four different distributions we simulated, with a very slight *right*-skew tilt.

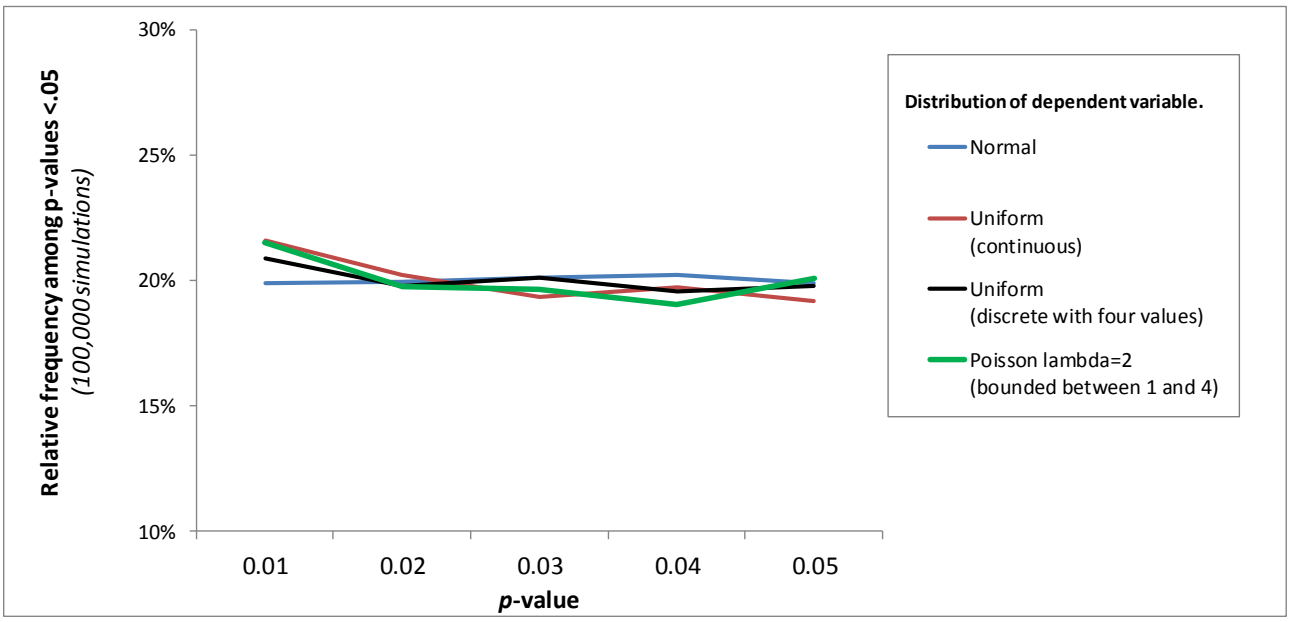


Fig S1. Observed *p*-curves from 500,000 simulated t-tests on two small samples (n = 15) drawn from the same population. Each *p*-curve represents a different underlying distribution, ranging from normal to increasingly non-normal. Because 5% of the 500,000 simulations are expected to be $p<.05$, each *p*-curve is based on roughly 25,000 *p*-values. The normally distributed dependent variable is N(0,1); the uniform-continuous is (0,1); the uniform-discrete can be y = 0.25, 0.5, 0.75, or 1; the Poisson can take values 1, 2, 3, 4 with probabilities roughly equal to .4, .3, .2 and .1, respectively.

(3) Modeling *p*-hacking

Let the $\{p_n\}$, $n \in \mathbb{N}$, be a sequence of *p*-values obtained by a researcher. p_1 is the first *p*-value that is obtained, presumably from the most straightforward test of the prediction of interest, using the entirety of the data that were collected. If $p_1<.05$ the sequence ends. If $p_1>.05$ a researcher may engage in file-drawering or *p*-hacking, which generates p_2. If $p_2<.05$ the sequence ends; otherwise, p_3 is generated, etc. In the discussion that follows we refer to the correlation between consecutive *p*-values p_i, p_{i-1} (for $i>1$) as r_i.

Uniform p-curve with file-drawering

Because file-drawering involves obtaining new data and entirely disposing of the previous data, file drawering leads to $r_i=0\ \forall\ i$, and hence $E(p_i|p_{i-1})=E(p_i)$. This means that the expected *p*-curve is the same in the presence or absence of file-drawering. For example, it is uniform under the null of no effect.

r>0 with (most forms of) p-hacking.

P-hacking can take a variety of forms, many of which lead to $r>0$. For example, adding a covariate, data peeking (adding new observations), data exclusions (e.g., dropping "outliers"), choosing among several correlated dependent variables, and choosing among several experimental conditions all produce sequences of *p*-values generated from statistical tests performed on datasets with overlapping observations. Thus, these forms of *p*-hacking generate sequential *p*-values that are positively correlated ($r>0$). A few forms of *p*-hacking may lead to $r=0$, such as choosing among uncorrelated dependent variables, choosing between reporting an interaction or a main effect, or choosing non-overlapping subsets of data (e.g., comparing treatment 1 with control, and then treatment 2 with placebo).

Note that r is not constant for a given sequence of p-values, even if they all arose from the same form of p-hacking within the same study. For example, when a researcher collects 20 observations and then adds new ones in sets of 10, r_i will be increasing in i; as more and more observations are added, the percentage of data that remains unchanged from test to test increases, and, thus, so does the correlation between resulting p-values (e.g., a t-test between two samples of 300,000 observations each, and one between the same 300,000 plus a new 10 per cell will lead to virtually the same result).

Left-skewed p-curves with p-hacking

Considering that p_i is obtained only if $p_{i-1}>.05$, it follows that if $r_i>0$ and $p_i < .05$, then p_i will be "close" to p_{i-1} and hence not too far from .05. More formally:

$$E(p_i \mid p_{i-1}>.05, r_i>0, p_i<.05) > E(p_i \mid p_i<.05)=.025$$

P-curve will be more skewed the higher r is, and moreover,

$$\lim_{r \to 1} E(p_i \mid p_{i-1}>.05, p_i<.05)=.05.$$

When the correlation between consecutive p-values is arbitrarily close to 1, then if a p-value in the sequence is significant (<.05) but the previous one is not, then it must be very close to .05. This means that p-curves of sets of p-hacked studies will be more left-skewed for p-hacking techniques that produce more correlated p-values.

4) The problem with discrete tests, and, how bad is it to ignore it?

4.1 The problem

While the paper focuses on the student distribution, it is straightforward to generalize it to others like the χ^2, the F with $df_1>1$, the normal, etc. This is not so for discretely distributed test statistics, including χ^2 approximations for examining contingency tables (e.g., difference of proportions tests). Note that while the χ^2 distribution is continuous, the distribution of possible χ^2 values for any given contingency table is not. A difference of proportions test is only *approximately* $\sim\chi^2$ (keep in mind that the normal test for the difference of proportions is *identical* to the χ^2 one).

The discrete nature of a statistical test imposes two challenges for applying *p*-curve. The first is that Fisher's method can no longer be used to aggregate *pp*-values. This challenge is easy to overcome as there are well-known methods to integrate discretely distributed *p*-values (Kincaid, 1962). The second challenge is that *pp*-values for discrete tests, or at least those based on contingency tables, depend on a nuisance parameter: the underlying proportions. For example, the *pp*-value for a difference of proportions test depends not only on the observed proportions, but on the true and unknown actual proportions.

The challenge is closely related to a long-standing controversy regarding tests of contingency tables condition (for extremely interesting reviews see Little, 1989; Yates, 1984). In a nutshell the controversy arises because when testing if two proportions, say $prop_1$ and $prop_2$ are equal, the result of a difference of proportions test depends on what we test those proportions being equal to, such that if we test $prop_1=prop_2=50\%$, we get a different result that if we test $prop_1=prop_2=30\%$. What they are assumed to be equal to is a nuisance parameter, in that it affects the result but we do not observe it.

The controversy arises in part also because using Fisher's "exact" test, which conditions on both margins (that is, conditions not only on sample sizes but also on the true overall proportion being exactly the same as that observed across both samples) the resulting *p*-value does not occur with its nominal frequency, so Fisher's exact test results in $p<.05$ less than 5% of the time. Fisher's exact test is *conservative*.

This, it turns out, reflects a difference in how Fisher and Neyman Pearson interpret *p*-values. Fisher did not think it was relevant if 5% of test are $p<.05$ under the null, Neyman Pearson thought that was the whole point (for a contrast of both schools of thought on p-values see Lehmann, 1993).

In any case, this nuisance parameter carries over to *pp*-value calculations and is amplified, such that the *pp*-value of obtaining $p=.04$ in a difference of proportions test between two samples of a given size, depends on what the two proportions are equal to, and can vary quite substantially depending on that parameter. In ongoing research we are considering an alternative that defines *pp*-values slightly differently for discrete tests, in a way that seems to eliminate this nuisance parameter. Another alternative is to Monte Carlo / bootstrap false-positive rates for the sample sizes one is *p*-curving, see section 4.3

4.2 How bad is it to ignore it

A pertinent question is just how bad is it to blindly compute *pp*-values on difference of proportions tests ignoring their discrete nature. We tried to address this question through simulations. We simulated sets of five difference of proportions test drawn under the null (the proportions are identical across any two samples being tested), computed the *p*-value using a $\chi^2(1)$ test, computed *pp*-values ignoring the discrete nature of the distribution, and aggregated the 5 studies to arrive at overall $\chi^2(10)$ tests for right-skew.

If the test were 'valid,' then x% of simulations would obtain an overall right-skew $p<x$, e.g., 5% of them would be $p<.05$, the p-value would correspond to the false-positive rate. It turned out that how accurate the p-value captured the false-positive rate depended on sample sizes and underlying proportions in non-monotonic ways. For example, if n=20 in each of the two samples in all five simulated studies, and in all of them the underlying proportion is 50% (so we expect 10 of the 20 observations to be 1s, and the other 10 to be 0s), then the nominal right-skew test for the five studies combined arrives at a p-value that is off on average off by 3 percentage points, e.g., there is an 8.3% chance of $p<.05$. If n=22 then the nominal rate is within 0.006 percentage points of the actual false-positive rate, but that is not thanks to the "larger sample," consider that if n=24 it is off by 2 percentage points again.

Basically what's happening is that the continuous approximation to the discrete distribution will undulate around the true value, and as sample size changes one can be in the peak or trough of that undulation (or right in the middle and get it just right). We did not find combinations of parameters that led to results worse than being off by more than 3 percentage points on average (for nominal $p<.1$).

When the set of studies is heterogeneous, e.g., some n=20, some n=22, the gap between the false-positive rate and the nominal p-value will be in between the extremes of 0 and 3 percentage points, which is encouraging because in the real world there will be heterogeneity.

We are led to tentatively conclude that until a better approach to p-curving discrete tests is available, it is reasonable to blindly treat the χ^2 as continuously distributed but be aware that the result is not as precise as it is for truly continuous statistics. It would be best practice to combine this approximate calculations with Monte Carlo simulations, see section 4.3.

4.3 Monte Carlo / bootstrap for discrete tests

When *p*-curves include difference of proportions test we recommend doing Monte Carlo simulations for studies of those exact characteristics to assess how accurate the resulting overall tests for skew are for that specific combination of parameters.

So for example, if a *p*-curve includes three difference of proportion tests studies with samples pairs of (n_1=21,n_2=24), (n_3=41,n_4=41) and (n_5=40, n_6=40) then we propose simulating studies with those exact sample sizes, under the null that the proportions are the same within each pair, and assess how close the nominal *p*-value for the overall test is to the false-positive rate, this allows making an informed guess as to how accurate the continuity approximation is for those specific parameters. One could go a step further and treat the percentage of simulated samples obtaining a nominal *p*-value below that observed in the real sample as the bootstrapped *p*-value for skew.

If a *p*-curve combines discrete and continuous test statistics one could bootstrap just the discrete ones and combine the result with that arising from the continuous ones.

This is a tentative solution; its performance ought to be assessed by future research.

(5) Selection of JPSP studies for the demonstration.

In the paper we plot *p*-curves for sets of studies published in the *Journal of Personality and Social Psychology (JPSP)* that we expected to have been intensely *p*-hacked and that we expected not to have been intensely *p*-hacked. Here we provide details of how the papers and *p*-values were selected and provide robustness results for their *p*-curves.

5.1) Set of studies reporting statistical results only with a covariate.

In a recent paper (Simmons, Nelson, & Simonsohn, 2011) we simulated false-positive rates obtained by researchers who *p*-hack by exploiting four specific researchers' degrees-of-freedom: (1) data-peeking (deciding whether to continue collecting data based on the statistical significance of existing data), (2) dropping a dependent variable, (3) dropping a condition (e.g., reporting only two cells of a three cell design), and (4) controlling for a covariate, especially under conditions of random assignment.

The first three of these are hard to detect in published research that does not follow our recommended disclosure rules (Simmons et al., 2011). The fourth, in contrast, is typically straightforward as authors do routinely disclose if their analyses control or do not control for covariates. With this in mind we decided to identify experiments using covariates as ones that might have been *p*-hacked.

We explored the feasibility of this approach by searching the archives of *JPSP*, using this interface: http://psycnet.apa.org/index.cfm?fa=search.defaultSearchForm. After browsing JPSP articles published in 2011 and 2012 that included the word "covariate," we defined three rules for selecting studies and applied those rules to studies published before 2011 (to ensure our rule

selection was not being influenced by its consequences, as in choosing rules that would favor our predicted *p*-curve shape).

Our rules for selecting articles had two main motivations. One was to focus on usage of covariates that would ex-ante be expected to be associated with *p*-hacking. The second was to minimize the subjectivity involved in the selection of *p*-values. These considerations led us to select only articles that satisfied <u>all</u> of the following criteria:

1) All independent variables of interest to the researchers were randomly assigned. This rule excluded, for example, studies examining correlates of personality scales, and those comparing people of different genders, races, or personality types. Note that this only applies to the independent variable of interest, not the covariate. Studies that randomly assigned participants to conditions and merely controlled for gender could be included.
2) The statistical results without the covariate are not reported. This rule excluded, for example, mediation analyses and robustness checks (e.g., authors examining if their results *also* hold when controlling for gender differences).
3) The covariate may not be causally affected by the manipulation. We applied this rule because when a covariate is correlated with the manipulation, collinearity may result in one observing a flatter *p*-curve even in the absence of *p*-hacking. This rule excluded, for example, studies that control for mood differences across conditions, if mood was measured after the manipulation.

We applied these rules to JPSP articles containing the words "covariate" and "experiment" in the full text, and published before 2011. We sorted the results by descending date and proceeded to examine papers one-by-one. If none of the exclusion rules applied to the first study using the word "covariate" in the text, we selected the key result, using the guidelines from Table 3. If any of the

rules were not met, we made a note of it and moved on to the next article. For simplicity we did not consider the next study in the same paper. This broad search led to articles from areas of psychology we were unfamiliar with; on a few occasions we excluded studies because we could not understand the hypothesis being tested. We registered those instances on a spreadsheet available upon request. We decided beforehand to collect significant *p*-values from 20 articles.

5.2) Set of studies expected not to have been intensely p-hacked.

After a similar exploratory process with articles published in 2011 and 2012, we conducted a search for pre-2011 JPSP articles that included the phrase "Experiment 2" and none of the following terms: *exclude, excluded, suspicion* (sometimes participants who express suspicion are dropped from experiments but the decision to exclude them can be made ex-post), *transform* (as when dependent variables are log or arcsine transformed), *log, covariate*. We included "Experiment 2" because we found it to be useful to help identify experiments in which all variables were manipulated.

We then proceeded to inspect articles one-by-one, and coded the *p*-values of articles that, in addition to the three rules from section 5.1 above, did not make any explicit allusion to the elimination of data or transformation of variables. This broad search led to articles from areas of psychology we were unfamiliar with; on a few occasions we excluded studies because we could not understand the hypothesis being tested. We registered those instances on a spreadsheet available upon request. We decided beforehand to collect significant *p*-values from 20 articles.

5.3) *Robustness tests for the demonstrations*.

As described in the paper, it is important that *p*-curve include only *p*-values that both directly test the stated hypothesis and that are statistically independent form each other. When more than one *p*-value directly tests the stated hypothesis but is not independent from another (e.g., when

t-tests on two correlated measures constitute equally appropriate tests of the stated hypothesis), then the researcher *p*-curving the study must decide which *p*-value to select initially, and then conduct robustness tests, where the selected *p*-value is replaced by one that was not chosen initially.

When we encountered this situation in our demonstration, we used the following rule. We initially selected the first *p*-value if two were equally relevant and we selected the median *p*-value if three were equally relevant. The *p*-curves and analyses depicted in Figure 3 feature those initial selections.

For the set with the covariate, Figure 3a, robustness involved a single instance where authors reported three tests of the hypothesis of interest. In the main text we reported *p*-curve results using the median of the three. Replacing the median with the lowest *p*-value reported in the triad barely affected the results; the test for lack of evidential value remained highly significant, $\chi^2(40)=80.5$, $p=<.0001$ (down from $\chi^2(40)=82.5$).

For the set without keywords associated with *p*-hacking, Figure 3b, robustness involved five instances when the authors reported two tests of the key hypothesis. In the main text we reported *p*-curve results always including the one appearing first in the text; for robustness we reran the analyses including only the one appearing second. The overall test for right-skew remained highly significant, $\chi^2(44)=93.6$, $p<.0001$ (down from $\chi^2(44)=94.2$).

6) Other statistical tests applicable to *p*-curve.

In the paper we propose two methods for conducting statistical inference with *p*-curve: binomial test of high vs low *p*-values, and computing *pp*-values. An alternative worth considering in the future is the Kolmogorov-Smirnov (KS) test. While it is known to have low power for small samples (in this case, few *p*-values), its one-tail version has the great advantage of allowing simultaneously testing for left-skew and right-skew. So a pair of one-tail KS tests could reject the uniform null and suggest some studies do have evidential value, and other studies within that same set were intensely *p*-hacked. Given our interest in applying *p*-curve to small sets of *p*-values we have not considered it in much detail but it may be useful for meta-analytical contexts.

Future research may also consider central tendency tests on *p*-curve, contrasting, for example, the mean or median *p*-value to those expected under different nulls through parametric (e.g., t-test) or nonparametric (e.g., Wilcoxon) tests.

7) Selection bias as an alternative to p-hacking to explain the ANCOVA example

A referee proposed an alternative explanation for our demonstration with studies reporting only ANCOVA and not ANOVA results (Figure 3a). If some researchers, call them "choosers," who obtain $p<.05$ with both ANOVA and ANCOVA *choose* to report the ANOVA result, then our sample which only includes ANCOVAs, will not include them, obviously.

Importantly, those missing studies are likely to have had low (ANCOVA) *p*-values, because the corresponding ANOVA ones were $p<.05$, and ANCOVAs tend to lead to lower *p*-values than ANOVAs. This type of selection bias would lead a set of studies reporting only ANCOVA results to have fewer than expected low *p*-values and hence to be less right skewed than otherwise expected.

We conducted simulations to assess if this type of selection bias could result not only in a less right-skewed *p*-curve, but also in a left-skewed one. That is, we conducted simulations to assess if this hypothetical form of selection bias was a plausible alternative explanation for Figure 3a. Our simulations involved two-cell studies with n=20 participants per cell and a covariate. We varied three parameters:

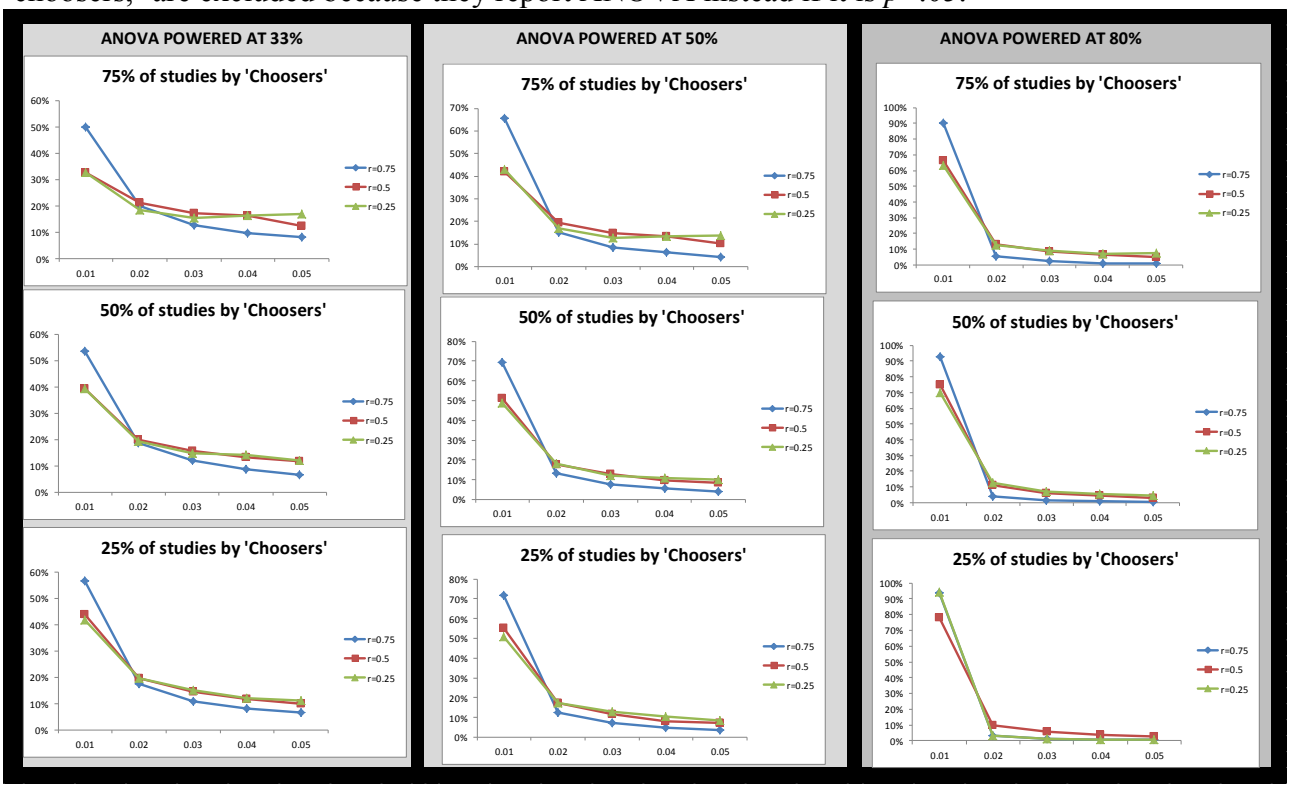
1. Percentage of researchers who are *choosers*: 25%,50% or 75%
2. Power of ANOVA test: 33%,50% or 80%.
3. Correlation between covariate and dependent variable r(y,z): r = .25,.5 or .75.

Those values of parameters can be combined in 27 different ways. We report all of them in Figure 2S below. We found that the expected *p*-curve is markedly right-skew for all of them. Even under the most extreme assumptions – most researchers are "choosers", the studies are severely underpowered, and the covariate is hardly related to the dependent variable – we still expect a right-skewed p-curve (the green line of the top-left panel in Figure S2). The p-curve we

actually observe for studies reporting experimental results only with a covariate, in sharp contrast, is distinctly, and significantly, left-skewed (Figure 3a).

Note that if researchers are ex-ante deciding on employing a covariate (z) to use for a dependent variable (y), such covariate is likely to have a high r(y,z), and for such situations *p*-curve is strongly right skewed .

Figure S2. Expected *p*-curves from sets of studies reporting ANCOVA where some researchers, 'choosers,' are excluded because they report ANOVA instead if it is *p*<.05.



Notes: Each panel reports three expected *p*-curves, obtained from 20,000 simulations each, for studies reported with ANCOVA, when some percentage of authors, 'choosers', choose to report ANOVA results if they are *p*<.05 and are hence excluded from an ANCOVA sample. ANCOVA samples exclude 'choosers' when their ANOVAs are *p*<.05, but include them otherwise. Each column contains simulations for a given level of statistical power for the ANOVA (the power for the ANCOVA is always higher of course and increases more the higher r(y,z) is), each row for a given percentage of researchers being 'choosers', and within each panel we consider three possible correlations, ex-ante, between a covariate and the dependent variable. Note that the percentage of choosers is not the ex-post share of people exiting, but the percentage that *would* do so if their ANOVA is *p*<.05. For example, the top left panel shows that if the ANOVA is powered to 33%, 25% of researchers would choose to report ANOVA instead of ANCOVA if it came up *p*<.05, and the covariate is correlated .25 with the dependent variable, then 33% of *p*-values for studies reporting ANCOVA would be *p*<.01, and 17% would be p>.04.

Table S1. *P*-curve disclosure table (PDT) for JPSP demonstration (part 1 of 3), also available as an Excel file from www.p-curve.com

link	Quoted text from paper stating hypothesis Comments by Simonsohn, Nelson Simmons, in purple	Brief description			Quoted text from paper describing results Selected result in bold	(6) results	(7) Robustness results
http://psycnet.apa.org/journals/psp/99/6/883.html	The first experiment was designed to be an initial examination of our prediction that **stereotype threat can reduce** learning by interfering with encoding processes. We proposed that stereotype threat reduces women's ability to learn mathematical rules and operations by reducing their ability to encode mathematical information into memory, not by inhibiting the ability to retrieve mathematical information from memory. [this is tested by stereotype threat influencing only math rules learn after the threat. They report first results on remembering rules, then on performance. We include remembering of rules in p-curve, and report robustness for the other]	**2 (Stereotype threat: control vs. threat) x** **2 (learning time; before vs after instructions)** *(attenuated interaction)*	**Interaction**	**Yes.** (note: this column is always "yes" when assessing evidential value of the finding of interest to researchers, it may only be "No" when engaging in a meta-analysis of findings that were not the primary interest of original researchers.	The results for mathematical learning showed the expected two-way interaction, **F(1, 57) = 7.25, p < .01, ηp2 = .11** (see Table 1). As predicted, the stereotype threat manipulation did not affect women's learning of mathematical rules presented before the instructions, F(1, 57) = 0.68, p = .41, ηp2 = .01; however, women in the stereotype threat condition learned fewer mathematical rules presented after the instructions than did women in the control condition, F(1, 57) = 3.96, p = .05, ηp2 = .07. Also as predicted, learning time did not impact the number of mathematical rules women learned in the control condition, F(1, 28) = 0.61, p = .44, ηp2 = .02, but women in the stereotype threat condition learned more of the mathematical rules presented before the instructions than the mathematical rules presented after the instructions, F(1, 29) = 15.83, p < .001, ηp2 = .35. The results for math performance also showed a two-way interaction, ***F* (1, 57) = 4.02, *p* = .05**, ηp2 = .07 (see Table 1). The stereotype threat manipulation did not affect women's performance on problems that used mathematical rules presented before the instructions, F(1, 57) = 0.16, p = .69, ηp2 = .00; however, women in the stereotype threat condition solved fewer mathematical problems based on the rules presented after the instructions than did women in the control condition, F(1, 57) = 8.12, p = .01, ηp2 = .13. [again, we p-curved F(1,57)=7.25 and report robustness for F(1,57)=4.02]	F(1,57)=7.25, p=0.0093	F(1,57)=4.02, p=0.0497
http://psycnet.apa.org/journals/psp/99/4/573.html	Dissonance theory only predicts spreading in the RCR condition because the spreading of preferences is considered an effect of choosing. A preference-driven choice theory, in contrast, predicts positive spreading for both. [authors want to know if RRC also leads to spreading, the key test is the simple effect of spreading in that condition]	**two-cell**	**Simple difference between cells**	Yes.	Critically, however, as predicted by a preference-driven model of choice, we also found positive chosen spread for participants in the RRC condition (*M* = 1.75, *SD* = 2.66, *n* = 40). In other words, on average, from Rank 1 to Rank 2, the item that participants eventually chose also moved further apart in participants' rankings. This spreading was also significantly different from 0, ***t* (39) = 4.16, *p* < .001.**	t(39)=4.16, p=0.0002	
http://psycnet.apa.org/journals/psp/98/6/872.html	We predicted that people would perceive their embarrassing moment as less psychologically distant when described emotionally.	**two-cell**	**Simple difference between cells**	Yes.	As predicted, participants perceived their previous embarrassing moment to be less psychologically distant after describing it emotionally (*M* = 4.90, *SD* = 2.30) than after describing it neutrally (*M* = 6.66, *SD* = 1.83), ***t* (38) = 2.67, *p* < .025** (see Table 1).	t(38)=2.67, p=0.0111	
http://psycnet.apa.org/journals/psp/98/5/761.html#S-6	We measured fixed-pie perceptions prior to the negotiation and then again after participants had received information about their counterpart's issue chart (or not, in the control condition). We predicted **stronger revisions of fixed-pie perceptions** when negotiators had a high rather than a low level of construal. Participants would be confirmed in their fixed-pie assumption when they either did not receive or process information on their counterpart or focused on issues only. They would revise their initial fixed-pie perception only when they would receive information on their counterpart and process the information on underlying interests rather than issues. Note: The authors are predicting that the effect of construal on revision (which is manipulated as time of measurement) will be greater when information is available than when it is not. This is a 3-way interaction: a moderated interaction - revision will be greater when construal is high vs. low - will be stronger under one condition than another (information availability).	**2 (Construal: high vs. low) x** **2 (Info on other: present vs. absent) x** **2 (time: time 1 vs. time 2)** (attenuation of an attenuated interaction)	**3-way interaction**	Yes.	We submitted fixed-pie perceptions assessed at Time 1 and Time 2 to a 2 × 2 × 2 (Construal Level × Information × Time of Measurement) ANOVA, with construal level and information as between-subjects factors and time of measurement as a within-subjects factor. Results revealed a main effect for time of measurement, F(1, 77) = 34.62, p < .001, η2 = .31; an interaction between time and construal level, F(1, 77) = 5.91, p < .05, η2 = .07; and an interaction between time and information, F(1, 77) = 30.52, p < .001, η2 = .28. These effects were all qualified by the predicted three-way interaction among time, construal level, and information, **F(1, 77) = 5.86, p < .05**, η2 = .07. Fig 1 shows that in the no-information condition, construal level had no effects on perceptual accuracy at Time 1 or at Time 2. In the information available condition, participants were more accurate at Time 2 than at Time 1, but this effect was considerably stronger in the high-construal level condition, F(1, 77) = 72.2, p < .001, η2 = .48, than in the low-construal level condition, F(1, 77) = 9.61, p < .01, η2 = .11. As a result, participants in the high-construal level condition were more accurate at Time 2 than those in the low-construal level condition, F(1, 77) = 3.77, p < .06 (approached significance).	F(1,77)=5.86, p=0.0178	
http://psycnet.apa.org/journals/psp/98/5/721.html	We predicted that the classical effect by Jacoby, Kelley, et al., namely the **misattribution** of increased fluency to fame, **would vanish** under the oral motor task but would still be detected under a manual motor task. [technically one cannot test that an effect vanishes, but one can that it gets significantly smaller, so we code the attenuated interaction]	**2 (exposure: old vs. new) x** **2 (motor task: oral vs manual)** *(attenuated interaction)*	**2-way interaction**	Yes.	Over the fame ratings in the test phase, a 2 (exposure: old items, new items) × 2 (concurrent motor task: manual, oral) analysis of variance (ANOVA) was run with motor task as a between-subjects factor. A main effect of exposure, *F* (1, 48) = 5.54, *p* < .023, ηp2 = .10, surfaced, as well as an interaction between exposure and motor task, ***F* (1, 48) = 4.12, *p* < .05, ηp2 = .08.** The conditional means are displayed in Table 1.	F(1,48)=4.12, p=0.0479	
http://psycnet.apa.org/journals/psp/98/4/605.html	Thus, we predicted that participants who were not mimicked would consume more of the snack than participants who were mimicked. Accordingly, we predicted that participants who were not mimicked would consume more of the snack than would control participants.	**three-cell (mimicking confederate, nonmimicking confederate, confederate absent)** *(one treatment, two controls)*	**Treatment vs control 1** *(confederate w/o mimicking)*	Yes.	Planned comparisons revealed that participants who were not mimicked consumed more grams of cookies than did participants who were mimicked, ***F* (1, 27) = 4.21, *p* < .05**, η2 = .13, and more grams of cookies than did control participants, ***F* (1, 27) = 5.51, *p* < .05**, η2 = .25.	F(1,27)=4.21, p=0.0500002	F(1,27)=5.51, p=0.0265
http://psycnet.apa.org/journals/psp/98/1/29.html	We expected that, consistent with the resource attribution hypothesis, the feedback would affect individuals in the low and high depletion states differently. Specifically, participants in the low depletion condition were expected to use our feedback to interpret their amount of available mental resources and, consequently, to persist longer on our problem-solving task when given the replenished (vs. depleted) feedback. Conversely, participants in the high depletion condition were expected to use our feedback to explain their amount of available mental resources and, consequently, to persist longer on our problem-solving task when given the depleted (vs. replenished) feedback.	**2 (depletion: high vs low) x** **2 (feedback: depleted vs. replenished)** *(reversing interaction)*	**Two simple effects**	Yes.	In the low depletion condition, participants persisted significantly longer when given the replenished, as opposed to depleted, feedback, ***t* (30) = −2.52, *p* < .02.** In the high depletion condition, participants persisted significantly longer when given the depleted, as opposed to replenished, feedback, ***t* (30) = 2.50, *p* < .02.**	t(30)=2.52, p=0.0173 t(30)=2.5, p=0.0181	
http://psycnet.apa.org/journals/psp/97/6/946.html	Type of gesture (gestures of approval vs. gestures of disapproval) was the manipulation. Reported attitudes served as the dependent measure. Role (participant vs. observer) was the predicted moderator variable. The central prediction called for a type of Gesture × Role interaction, in which perceivers would report more positive attitudes after seeing gestures of approval than disapproval made by someone else toward an attitude object, but observers who saw the same gestures made by someone else would not, because observers would not have the same visual illusion and inferential cues to agency as would perceivers.	**2 (Role: perceiver vs. observer)x** **2 (Gesture: aproval vs. dissaproval)** *(attenuated interaction)* *(note: there is a third role condition over which no strong prediction is made: hand helper)*	**two-way interaction**	Yes.	A 2 (type of gesture) × 3 (role) ANOVA was performed on participants' answers to the question about their current attitudes toward gay men, after participating in the experiment. That ANOVA yielded the predicted type of Gesture × Role interaction, *F* (2, 62) = 3.53, *p* < .05. In addition, an ANOVA that included just perceivers and observers yielded the same significant interaction, ***F* (1, 31) = 6.73, *p* < .05**	F(1,31)=6.73, p=0.0143	

Table S2 continues (part 2 of 3)

link	Quoted text from paper stating hypothesis Comments by Simonsohn, Nelson Simmons, in purple	Brief description			Quoted text from paper describing results Selected result in bold	(6) results	(7) Robustness results
http://psycnet.apa.org/journals/psp/97/5/823.html	The prediction that **only** high identifiers would respond to a heightened identity threat by giving more help [note: the inro and setup of this experiment give slightly different predictions, one involves an unreported comparison of one cell against all three others, the other just the simple effect, but given the design, and this stated prediction, we are selecting the p-value for the most natural test: the interaction]	**2 (group identification: low vs. high) x** **2 (threat level: low vs high)** *(attenuated interaction)*	**two-way interaction**	Yes.	A 2 (high- vs. low ingroup identification) × 2 (high- vs. low threat) analysis of variance (ANOVA) on amount of help-giving revealed a Threat × Ingroup Identification interaction, ***F* (1, 92) = 4.45, *p* < .05**, η2 = .05. To pursue our findings for the prediction that only high identifiers would respond to a heightened identity threat by giving more help, we used orthogonal contrasts to compare helping under high- and low-threat conditions for high- and low identifiers separately. This analysis indicated that participants in the high-identification condition gave more help to the outgroup when it posed a relatively high than low threat to social identity (*M* = 4.30, *SD* = 2.40 and *M* = 2.57, *SD* = 2.10, respectively), *F* (1, 45) = 7.80, *p* < .01, η2 = .13. The amount of help given in the high-threat cell did not differ from the amount of help given in the low-threat cell in the low-identification condition (*M* = 3.36, *SD* = 1.91 and *M* = 3.46, *SD* = 1.86, respectively), *F* (1, 47) < 1 (see Table 1).	F(1,92)=4.45, p=0.0376	
http://psycnet.apa.org/journals/psp/95/2/319.html	Thus, we hypothesize that when a sociability goal is activated via priming, the accessibility of various friends should be guided by their instrumentality for this goal. When no such goal has been activated, we hypothesize, friend accessibility should not be affected by instrumentality.	**2 (prime: goal vs. control) x** **2 (instrumentality; yes vs no)** *(attenuated interaction)*	**two-way interaction**	Yes.	As predicted, a significant two-way interaction emerged, ***F* (1, 32) = 4.46, *p* < .05.** As illustrated in Figure 1,the active goal influenced the cognitive accessibility of instrumental versus noninstrumental friends. For participants primed with a sociability goal, instrumental friends were significantly more accessible than were noninstrumental friends, F(1, 16) = 19.63, p < .05. Participants in the control condition showed no such effect; no significant difference in the accessibility of instrumental and noninstrumental targets emerged (F < 1, ns).	F(1,32)=4.46, p=0.0426	
http://psycnet.apa.org/journals/psp/94/6/988.html	We predicted that Jews would perceive Palestinians as being more responsible for the conflict and legitimize Israeli actions more when reminded of the Holocaust than when not reminded, and doing so would lessen feelings of collective guilt.	**two-cells**	**Difference of means** *(on collective guilt)*	Yes.	As predicted, participants in the Holocaust reminder condition (*M* = 2.92, *SD* = 1.67) reported less collective guilt than participants in the no-reminder condition (*M* = 4.14, *SD* = 2.24), ***F* (1, 52) = 5.08, *p* = .03**, *d* = .62.	F(1,52)=5.08, p=0.0284	
http://psycnet.apa.org/journals/psp/94/4/547.html	We predicted that presenting these items together in one image would increase the value of unhealthy (temptation) items, whereas presenting them apart, in separate images, would increase the value of healthy (goal) items.	**3 (presentation format: together, single (control) , apart)x** **2 (food: healthy vs unhealthy)** *(rreversing trends)*	**Two simple effects of high vs low** *(because trends including control cell not reported)*	Yes.	A Presentation Format × Food Type ANOVA of these composite value scores yielded a main effect of food type, *F* (1, 62) = 5.64, *p* < .05, indicating that the healthy foods were more appealing than the unhealthy foods. It also yielded the predicted Presentation Format × Food Type interaction, *F* (2, 62) = 12.31, *p* < .001 (see Figure 2) A contrast analysis revealed that in the single (control) presentation format, participants provided similar ratings to healthy food items (M = 4.56, SD = 1.00) and unhealthy food items (M = 4.18, SD = 1.13), t(23) = 1.17, ns. Thus, we were successful in choosing healthy and unhealthy food items with a priori similar value. Moreover, when the items were presented together, participants provided higher value ratings to unhealthy food items (M = 5.02, SD = 0.76) compared with healthy items (M = 4.30, SD = 1.04), **t(20) = 3.36, p < .01.** In contrast, when the items were presented apart, participants provided higher value ratings to healthy food items (M = 5.39, SD = 1.27) compared with unhealthy items (M = 3.65, SD = 1.33), **t(19) = 3.79, p < .01.**	t(20)=3.36, p=0.0031 t(30)=2.5, p=0.0181	
http://psycnet.apa.org/journals/psp/93/4/515.html#S-2	In Experiment 1, we tested the proposition that a disappointing choice would be regretted more if it were made from a larger decision set than from a smaller decision set.	**Two-cells**	**Difference of means** *(one of the two cells consists of two counterabalanced ones that were collapsed)*	Yes.	Contrasts between the conditions showed that there was more regret when there were two alternatives to going to the movie (*M* = 5.59, *SD* = 1.44) than when there was just one (*M* = 3.82, *SD* = 1.73), ***F* (1, 72) = 22.56, *p* < .001,**	F(1,72)=22.56, p=0.00001	
http://psycnet.apa.org/journals/psp/93/2/143.html	It was predicted that more unrequested (negative) cognitions would be reported in the difficult than in the easy condition.	**two-cells**	**Difference of means**	Yes.	As predicted, the number of positive thoughts manipulation had a significant effect on participants' self-reported unrequested cognitions, ***t* (26) = -4.98, *p* < .001.** Participants indicated that more negative thoughts came to mind when they had been asked to list 10 (*M* = 5.00, *SD* = 2.04) rather than 2 (*M* = 2.07, *SD* = 0.83) positive thoughts.	t(26)=4.98, p=0.00003543	
http://psycnet.apa.org/journals/psp/91/6/1009.html#S-6	Experiment 1 tested whether answering questions correctly before attempting to answer them randomly would result in successful random answers.	**two-cells**	**Difference of means**	Yes.	Participants who were allowed to answer only once, randomly, exhibited a significantly higher mean proportion of correct responses (*M* = .58, *SD* = .15) than did correct-random participants (*M* = .49, *SD* = .12), ***t* (46) = 2.07, *p* < .05**, η2 = .09 (see Figure 1).	t(46)=2.07, p=0.0441	
http://psycnet.apa.org/journals/psp/91/1/97.html#S-5	Our main predictions rest on our theorizing that when **status relations are perceived as relatively unstab**le, dependence on the high-status outgroup is inconsistent with group members' quest for equality and results in a threat to social identity. This threat should be expressed in relatively low affect, drive group members to positively distinguish the ingroup by discriminating against and devaluing the outgroup, and perceive the ingroup and the outgroup as more homogeneous (Studies 1 and 2). (prediction is cleaer in discussion of results): When the status hierarchy was perceived as relatively stable, the receipt of help from the high-status outgroup did not influence recipients' affect, ingroup favoritism, and perceptions of the outgroup. Yet when the status hierarchy was perceived as unstable, being helped by a member of the high-status outgroup led recipients to feel worse.	**2 (Help: yes vs no) x** **2 (relations status: stable vs. unstable)** *(attenuated interaction)*	**two-way interaction** *(for three d.v.s)*	Yes.	(1)A 2 (help vs. no help) × 2 (stable vs. unstable status) ANOVA on the measure of ingroup favoritism revealed no significant effects. Although the predicted Help × Stability interaction was not significant, **F(1, 63) < 1,** (2) A 2 (help vs. no help) × 2 (stable vs. unstable status) ANOVA on the general evaluation score.... ... The Stability × Help interaction was not significant, F(1, 63) < 1. (3) A 2 (help vs. no help) × 2 (stable vs. unstable status) ANOVA on perceived aggressiveness of the outgroup revealed a significant interaction, **F(1, 63) = 3.70**, p < .05 (4) A similar ANOVA on the perceived homogeneity of the outgroup revealed a significant Status Stability × Help interaction, **F(1, 63) = 8.27, p < .005**	F(1,63)=3.7, p=0.0589	F(1,63)=8.27, p=0.0055

Table S2 continues (part 3 of 3)

	Quoted text from paper stating hypothesis Comments by Simonsohn, Nelson Simmons, in purple	Brief description			Quoted text from paper describing results Selected result in bold	(6) results	(7) Robustness results
http://psycnet.apa.org/journals/psp/88/2/288.html#S-5	We expected that Jews would be more willing to **forgive** Germans for the past when they categorized at the human identity level and that the **guilt** assigned to contemporary Germans would be lower in the human identity condition compared with the social identity condition.	**two-cell**	**Difference of means** *(for two d.v.s)*	Yes.	Participants assigned significantly less collective guilt to Germans when the more inclusive human-level categorization was salient (M = 5.47, SD = 2.06) than they did when categorization was at the social identity level (M = 6.75, SD = 0.74), **F(1, 45) = 7.62, p <.01**, d = 0.83. Participants were more willing to **forgive** Germans when the human level of identity was salient (M = 5.84, SD = 1.25) than they were when categorization was at the social identity level (M = 4.52, SD = 0.92), **F(1, 45) = 16.55, p <.01, d = 1.20.**	F(1,45)=7.62, p=0.0083	F(1,45)=16.55, p=0.0002
http://psycnet.apa.org/journals/psp/75/2/347.html	Hypothesis 1.1 : When target persons belonging to two different categories are presented in a mixed presentation mode, participants will infer that their primary task is to differentiate between the categories. Consequently, **assimilation effects** will occur. Hypothesis 1.2 : When target persons belonging to two different categories are presented in blocks, participants will infer that their primary task is to differentiate within the categories. Consequently, **contrast effects** will occur.	**2 (categorical info: nurse vs. stockbroker) x** **2 (presentation: mixed vs. blockwise) x** **2 (individuating information: mildly unhelpful vs. helpful)** *(reversing interaction for first 2x2)* *note: third factor is not explicitly made a prediction for in the motivation of the stuy)*	**Both simple effects**	Yes.	Supporting our hypotheses, nurses were rated as more helpful than stockbrokers in the mixed presentation condition, ***t* (9) = 2.03, *p* < .04 (one-tailed)**, which indicates an assimilation effect; nurses were rated as slightly less helpful than stockbrokers in the blockwise presentation condition, ***t* (9) = 1.69, *p* < .07 (one-tailed), d**emonstrating a contrast effect	t(9)=2.03, p=0.0729 t(9)=1.69, p=0.1253	
http://psycnet.apa.org/journals/psp/66/1/48.html#S-4	We created a situation in which some subjects were to think that they possessed individuating information about an introverted or extraverted target without having actually been confronted with such information. Contrary to other subjects who were not told that they had been informed, those who believed they had been informed were expected to display more **confidence** in their judgments. Also, their ratings should be more **polarized** in the direction of the activated stereotype.	**two-cell** *(two measures)*	**Difference of means** *(for two d.v.s)*	Yes.	As far as the **confidence** measure was concerned, the analysis of the *don't know* answers revealed a highly significant information status main effect, ***F*(1, 55) = 9.96, *p* < .003**. Subjects who thought that individuating information had been given to them avoided the questions less often than the other subjects (*M* s = 5.07 and 10.13, respectively). **[polarized:]**prediction was supported by the presence of a highly significant information status main effect for the congruence scores, **F(1, 55) = 8.26, p < .006**. In other words, our subjects judged the archivist to be more introverted and the comedian more extraverted when they supposedly had received individuating information (M = 9.97) than when no such induction had taken place (M = 6.30)	F(1,55)=9.96, p=0.0026	F(1,55)=8.26, p=0.0058
http://psycnet.apa.org/journals/psp/62/4/699.html#S-5	The first study was planned as a simple demonstration of the hypothesis that the arousal of any mood state, whether positive or negative, leads to self-directed attention	**three-cell** (treatment 1, treatment 2, control)	**Quadratic trend** *(note: this is an unusual prediction and the quadratic trend seems a natural way to test it)*	Yes.	Subjects in the happy mood condition (*M* = 25.75, *SD* = 5.42) and the sad mood condition (*M* = 24.92, *SD* = 5.80) scored higher on the Linguistic Implications Form than did subjects in the neutral mood condition (*M* = 22.86, *SD* = 4.85). The U-shaped pattern of means is congruent with the hypothesis that both happy and sad moods produce more self-focus than do neutral moods. To confirm this hypothesis, a one-way ANOVA was conducted in which a quadratic trend was specified using contrast weights of 1, −2, and 1 for the happy, neutral, and sad mood conditions, respectively. This analysis revealed that the data fit this hypothesized pattern of results, ***F* (1, 104) = 5.04, *p* < .05**. The contrast residual, however, was not significant, indicating that a U-shaped pattern of means, as predicted by Hypothesis 3, fits the data well.	F(1,104)=5.04, p=0.0269	
http://psycnet.apa.org/journals/psp/89/4/504.html	We expected participants to relate the stereotype-consistent behaviors more abstractly than the stereotype-inconsistent behaviors. Moreover, we expected this effect to be more pronounced when the category label was presented before participants heard the story than when the category label was presented after participants heard the story.	**2 (category label: chess master vs hairdresser)x** **2 (label: before vs. after)x** **2 (behavior: intelligent vs. sociable)** (attenuation of attenuating interaction)	**Three-way interaction**	Yes.	The only significant effect was the expected three-way interaction between category label, label presentation, and behavior, ***F* (1, 44) = 8.38, *p* < .01**, η2 = .14 (see Table 1). The two-way interaction between category label and behavior was significant only when the category label was presented before the story (an LEB effect), F(2, 44) = 5.05, p = .01, η2 = .18, but not when the category label was presented after the story, F(2, 44) = 1.63, p = .21. When the category label was presented before the story, sociable behavior was described more abstractly than intelligent behavior for a hairdresser, t(9) = 2.80, p = .02, d = 0.91, and intelligent behavior was described somewhat more abstractly than sociable behavior for a chess master, t(10) = 1.85, p = .09, d = 0.54.	F(1,44)=8.38, p=0.0059	
http://psycnet.apa.org/journals/psp/86/2/219.html	Our primary hypothesis was that although failed counterarguing would lead to attitudes that were equivalent in valence to those that followed undirected thinking, the former attitudes would be held with **greater certainty.**	**Four-cell design** (two expected to show effect, other two are controls)	**Difference of means** *(for focal two conditions)*	Yes.	There was a significant effect of treatment on attitude certainty, *F* (2, 54) = 4.51, *p* =.02. Individuals instructed to generate negative thoughts (*M* = 7.64, *SD* = 1.20) were more certain of their attitudes than individuals who attempted to generate either thoughts (*M* = 6.72, *SD* = 1.20), ***t* (36) = 2.35, *p* =.02**, or positive thoughts (*M* = 6.50, *SD* = 1.26), *t* (35) = 2.82, *p* <.01. [note: there is also a prediction of same valence, that's considered supported in the paper by lack of statistical significance]	t(36)=2.35, p=0.0244	

References for Supplementary Materials

Boneau, C. A. (1960). The Effects of Violations of Assumptions Underlying the T Test. *Psychological bulletin, 57*(1), 49.

Cumming, G., & Finch, S. (2001). A Primer on the Understanding, Use, and Calculation of Confidence Intervals That Are Based on Central and Noncentral Distributions. *Educational and Psychological Measurement, 61*(4), 532-574.

Kincaid, W. (1962). The Combination of Tests Based on Discrete Distributions. *Journal of the American statistical association*, 10-19.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *Journal of the American statistical association, 88*(424), 1242-1249.

Little, R. J. A. (1989). Testing the Equality of Two Independent Binomial Proportions. *American Statistician*, 283-288.

Pearson, E. S. (1931). The Analysis of Variance in Cases of Non-Normal Variation. *Biometrika, 23*(1-2), 114-133.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science, 22*(11), 1359-1366.

Yates, F. (1984). Tests of Significance for 2× 2 Contingency Tables. *Journal of the Royal Statistical Society*. *Series A (General)*, 426-463.