This version: 2014 12 16

SUPPLEMENTARY MATERIALS FOR:

"Small Telescopes: Detectability and the Evaluation of Replication Results"

Paper: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2259879

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

OUTLINE.

R programs used to generate all figures in the paper and in this supplement:
https://osf.io/adweh/files/

Supplemental Materials - Introduction

In these supplementary materials I provide step-by-step calculations with detailed explanations for all results reported in the paper. They are written so that readers who have never been exposed to power calculations or noncentral distributions can follow them without much difficulty and without needing to consult additional sources. They are therefore longer than strictly necessary.

The following two equations are used throughout:

Equation #	**Formula**	**Brief Description**
(Eq. 1)	$d=\frac{2t}{\sqrt{2n}}$	Links the effect size (Cohen-d) with the t-statistic for a difference of means test, for given per-cell sample size n.
(Eq. 2)	$ncp=\sqrt{\frac{n}{2}}\,d$	For a two-sample t-test, the noncentrality parameter (ncp) of the student distribution is obtained by multiplying effect size by the square root of half the sample size of each cell.

R programs used to generate all figures in the paper and in this supplement:

https://osf.io/adweh/files/

Supplement 1 – Calculations for Example 1 – replication of Zhong and Liljenquist (2006)

Zhong and Liljenquist (2006) do not report an effect size for Study 1. The reported means and standard deviations in their Table 1 (ethical: M=.9, SD=1.88, unethical: M=1.43, SD=1.77) are incompatible with the reported F(1,58)=4.26 for assessing the statistical significance of such difference of means. Through email communications the first author indicated that the SDs were erroneously reported in the paper.

The *d*=.53 used here is based off the reported F-test, presumed to be correct. If F(1,58)=4.26, then we have t(58)=2.064 (recall that if t(*df*) is distributed student, t^2 is distributed F(1,*df*)). Using (Eq.1) we get from t(58)=2.064 to d=.533.

Confidence intervals

The confidence intervals for this *d*, plotted in Figure 1, are obtained by following the method discussed by Cumming and Finch (2001, pp. 544-545) which relies on noncentral distributions (see also Smithson, 2003). In particular, for the 90% confidence interval one identifies the noncentral parameter (ncp) of the student distribution that would lead 95% of observed *d*s be smaller than that observed, d=.54, (i.e., $prob(d_{observed}<.53|ncp_{high})=.95$) and the ncp leading to 95% of them being larger (i.e., $prob(d_{observed}>.53|ncp_{low})=.95$). Those ncps are associated with the ends of the confidence intervals, and are converted into ds using (Eq.2). For the 95% confidence intervals one proceeds analogously but setting those probabilities right of the equal sign to .975.

Below I provide SAS and R code, side to side, that compute the confidence intervals this way.

SAS CODE

```
data z1;
*Get the ncp low for 90 and 95 confidence intervals;
ncp_low90=tnonct(2.064,58,.95);
ncp_low95=tnonct(2.064,58,.975);

*Get the ncp high;
ncp_high90=tnonct(2.064,58,.05);
ncp_high95=tnonct(2.064,58,.025);

*Go from ncp to d;
dhigh95=ncp_high95/sqrt(30/2);
dhigh90=ncp_high90/sqrt(30/2);
dlow95=ncp_low95/sqrt(30/2);
dlow90=ncp_low90/sqrt(30/2);

run;

proc print;
run;
```

R CODE

```
#Create tnonct function, like SAS
tnonct = function(delta,pr,x,df) pt(x,df = df, ncp = delta)-pr


#Get the ncp high
ncp_low90=uniroot(tnonct,c(-10, 10),pr=.95, x=2.064, df=58)$root
ncp_low95=uniroot(tnonct,c(-10, 10),pr=.975,x=2.064, df=58)$root

#Get the ncp low
ncp_high90=uniroot(tnonct,c(-10,10),pr=.05,x=2.064,df=58)$root
ncp_high95=uniroot(tnonct,c(-10,10),pr=.025,x=2.064,df=58)$root

#Go from ncp to d
dhigh95=ncp_high95/sqrt(30/2)
dhigh90=ncp_high95/sqrt(30/2)
dlow95=ncp_low95/sqrt(30/2)
dlow90=ncp_low90/sqrt(30/2)

#print result
c(dhigh95,dhigh90,dlow90,dlow95)
```

To find $d_{33\%}$ we identify the effect size that gives a two-sample difference of means test, with n=30, 33% power, which is d=.401.[1]

Calculations for replication by Gámez, Díaz, and Marrero (2011)

Gámez et al. report a critical F test comparing the two condition means of F(1,45)=.08, implying a t(45)=sqrt(.08)=.283.[2]

Using Eq. 1 this is equivalent to *d*=.083, indicated by the marker in Figure 1.

For the confidence interval one proceeds analogously to what was described above.

To compute the *p*-value for the test where the null hypothesis is that $t_{rue}= d_{33\%}$ we assess how likely it would be to observe a t-value at least as extreme as that which we observe, t≤.28, if the true effect size were $d_{33\%}$. This calculation relies again on the

[1] E.g., in R we type:
library(pwr)
pwr.t.test(n=30,power=1/3)$d

[2] The paper reports F(1,*46*), but given the sample size, df_2=45.

noncentral t-distribution. In particular, we set the *ncp* (Eq.2) evaluated at the *n* of the replication, n=47/2=23.5, and the *d* of the $d_{33\%}$ computed earlier, $d_{33\%}$=.401, so that:

ncp=$\sqrt{\frac{n}{2}}$ d=$\sqrt{\frac{23.5}{2}}$. 401 =1.375 (that is the ncp under the null of $d_{33\%}$ for a sample of the size of the replication) and we now evaluate the probability of observing a t-test with an t≤.283 for that ncp.

In R syntax:
pt(.283,df=45,ncp=1.375)
[1] .13717

In words, if the true effect size were $d_{33\%}$ (d=.401), there is a 13.7% chance that a difference of means test with samples n=23 and n=24 would result in as small or smaller *d* than that *d* reported by Gámez et al. We do *not* reject the null of a detectable ($d_{33\%}$) effect, *p*=.137. [3]

Calculations for replication by Siev (2012)

Jedidiah Siev contacted me via email and informed me of his replication of Studies 1 and 2 in Zhong and Liljenquist (2006). I have strongly encouraged Jedidiah to make the details of his study publicly available (e.g., posting it on PsychFileDrawer.org).

The study by Siev (2012) was a web based survey completed by Penn undergraduates for course credit. The conditions of interest had participants imagine an ethical act (n=170) or unethical one (n=165) and then completed 8 word-stems that could have cleansing related meanings. The ethical and unethical primes lead to similar number

[3] The ncp for a difference of means test is actually ncp=$\sqrt{\tilde{n}}\delta$ where $\tilde{n}=\frac{n_1 n_2}{n_1+n_2}$. When both samples have the same n, this becomes Eq.2: ncp=$\sqrt{\frac{n}{2}}\delta$. For this study where the samples are different, rather than dividing the average n, 23.5, by 2, 11.75, one should compute $\frac{23*24}{23+24}$=11.74468. Rather similar.

of words completed ($M_{Unethical}$=.8667, $SD_{Unethical}$=1.0153, vs. $M_{Ethical}$=.8765, $SD_{Ethical}$=.94339). These results, t(333)=-.092, imply a trivial effect in the opposite direction, d=-.01. The confidence interval and *p*-value for the null of $d_{33\%}$ are computed analogously to what was done for the Gámez et al. paper.

Supplement 2. Calculations for Example 3 – Schwarz and Clore (1983)

2.1 Computing effect size in the original paper

Schwarz & Clore do not report effect size, do not report SDs, and do not report the precise test statistic for the contrast of interest (sunny vs raining in the no "prime" column). Schwarz & Clore do report the means (for three dependent variables) across the six conditions of their experiment 2 in the table reprinted below:

Table 3
Mean Ratings of General Happiness, Desire to Change, and Life Satisfaction: Experiment 2

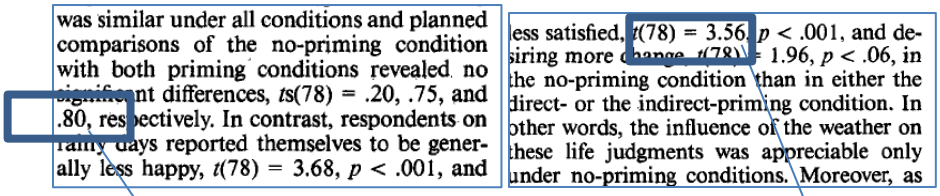
Dependent variable	Priming		
	None	Indirect	Direct
General happiness			
Sunny	7.43_a	7.29_a	7.79_a
Rainy	5.00_b	7.00_a	6.93_a
Desire to change			
Sunny	3.93_a	3.43_a	3.57_a
Rainy	5.79_b	$4.57_{a,b}$	$4.93_{a,b}$
Life satisfaction			
Sunny	6.57_a	6.79_a	7.21_a
Rainy	4.86_b	6.71_a	7.07_a

Note. $n = 14$ per cell. Means that do not share a common subscript differ at $p < .05$ (Newman–Keuls test).

Because the replications I examine use a life-satisfaction measure, I focus only on that variable going forward (last two rows in the table). The two key cells of interest are the 6.57 vs 4.86 for sunny vs. rainy weather when respondents are not reminded of the weather.

Schwarz & Clore report statistical results comparing, within a row, the left most cell to the next two collapsed, through a planned contrast based off an ANOVA using all 6 cells. This allows computing the pooled SD for all 6 cells.

In particular, from these two paragraphs in page 520:



we learn that a planned contrast between life-satisfaction on sunny day with no prime (6.57), vs. the average of sunny day with the two prime manipulations (6.79 & 7.21) leads to a t-test of t(78)=.8. This implies SD_{pool}=1.64 across all six cells.[4]

Proceeding analogously for the rainy condition we find a t-test of 3.56 for the planned comparison between 4.86 vs (6.71 & 7.07), SD_{pool}=1.74 across all six cells. These two numbers, 1.64 and 1.74, ought to be identical as they both represent the pooled SD across the same six cells, but they probably differ due to rounding errors (e.g., the t-test reported as .80 may have been .796 or .804).

The simple average of these two 'estimates' is 1.69 which I employ to compute the effect size of the comparison of interest. In particular, in the no-prime column, sunny (6.57) vs. rainy (4.86) is a 1.71 difference of means, dividing by SD=1.69, we arrive at the estimated effect size: d=1.71/1.69=1.01.

2.2 How big was the sunny vs. rainy day manipulation by Schwarz and Clore?

The unstandardized effect size in Study 2 by Schwarz and Clore (1983) is 1.7 difference in the likert scale for a discrete "sunny" vs. "rainy" day comparison. The effect

[4] The formula for the t-test for a planned contrast involving three means is $= \frac{M1*w1+M2*w2+M3*w3}{\sqrt{MSE}*(\frac{w1^2}{n1}+\frac{w2^2}{n2}+\frac{w3^2}{n3})}$, where w_i, M_i and n_i are the weight, mean and sample size of cell i, and MSE is the overall mean square error. For the values reported in Schwarz & Clore this becomes $.8 = \frac{6.57*1-6.79*(-.5)+7.21*(-.5)}{\sqrt{MSE}*(\frac{1^2}{14}+\frac{.5^2}{14}+\frac{.5^2}{14})}$, solving for $\sqrt{MSE}$, a pooled SD, we get SD~1.64l

size reported by Feddersen, Metcalfe, and Wooden (2012), in contrast, is for one standard deviation change in sunshine.

To more directly compare effect sizes between both studies we need to know how many standard deviations of sunshine separated those "sunny" vs. "rainy" days examined by Schwarz and Clore (1983). They do not give enough details to answer this in that paper, but, piecing together additional details they reported in their 2003 paper with data I obtained from WeatherUnderground.com for Champaign, IL, where the study was conducted, I set out to estimate it.[5]

In particular, from these three quotes in Schwarz & Clore 2003, pp. 298:

(i) "We conducted the experiments […] in 1980–1981",

(ii) "The sunny days we used were the first two sunny spring days after a long period of gray overcast days" and

(iii) "the rainy days we used were several days into a new period of low-hanging clouds"

I determined that March 12th and 13th of 1981 met the Sunny day descriptions, and April 10th and 11th the rainy day ones. The difference in cloudcover between them was equivalent to 2.5 standard deviations (of daily cloudcover for Champaign for 1981). The effect in Schwarz and Clore (1.7) divided by 2.5 is hence more directly comparable to Feddersen et al.'s; it (1.7/2.5=.68) is 60 times the size obtained in the replication (.012). One cannot be certain those were the days when the experiment was run, but even if it was run on other days, given how well these days match the description, weather conditions were presumably extremely similar on those days.

[5] I have posted these weather data with the Excel spreadsheet here:https://osf.io/adweh/files/ go to the tab "Suppl 2.Weather in Champagin 81"

2.3 Calculations behind Figure 2

Confidence intervals for Schwarz and Clore (1983)

Effect size calculations for Schwarz and Clore are covered in Supplement 2.1. The point estimate is an effect of $\hat{d}$=1.01. Relying on (Eq. 1) we find that this is equivalent to a t-test for the sunny vs. rainy comparison of

t=d*sqrt(df)/2=1.01*sqrt(28)/2=2.67.

With calculations analogous to those used for Example 1, we find the noncentrality parameters (ncps) that give this t=2.67, or more extreme, a 95%,97.5%, and 5% and 2.5% chance of being observed, obtaining a 95% confidence interval for d of (.212,1.79), and a 90% confidence interval of (.339,1.664). Both are plotted in Figure 2.

To find $d_{33\%}$ we identify the effect size that gives a two-sample difference of means test 33% power, which is d=.5998.[6]

Feddersen et al.

Feddersen et al. report regression results which I convert into units comparable to those of Schwarz and Clore in a few steps. Let's begin with the point estimate. Feddersen et al's preferred specification is Model 4 in their Table A1. It shows a point estimate for sunshine on life-satisfaction of β=.00191. That says that a unit increase of sunshine leads to an increase of .00191 units of life satisfaction. Multiplying by the SD of sunshine (SD=6.43, their Table A10) we arrive at .0122, one SD increase in sunshine increases life satisfaction by .0122 points. Dividing that by the SD of life satisfaction (SD=1.52, their Table A10), we obtain .008, so one SD of sunshine increases life satisfaction by .008 SD.

[6] E.g., in R we type:
library(pwr)
pwr.t.test(n=14,power=1/3)$d

As mentioned in Supplement 2.2, the manipulation by Schwarz and Clore is equivalent to 2.5 SDs of sunshine, so we multiply that by 2.5 to arrive at Cohen's d effect size for a comparably sized manipulation, $d_{Feddersen}$=2.5*.008=.0202, the point estimate plotted in Figure 2 for the Feddersen study.

The sample in Feddersen et al is large enough that we can obtain a confidence interval for this d using the standard error (SE) of the point estimate (i.e., ignoring that σ is estimated). Their Table A1 model 4, from where the β from above was obtained, reports a t-test of 2.11. Considering that t=β/SE, we then have that 2.11=.00191/SE, so SE=.000905. That's the SE of the impact of one point increase in sunshine on life-satisfaction. Multiplying by the SD of sunshine (SD=6.43, their Table A10) we obtain the SE for the effect of a 1 SD increase: .005821. Dividing by the SD of life satisfaction (SD=1.52, their Table A10), we obtain .003829; the SE of a one SD increase in sunshine measured in SD of satisfaction. The manipulation of interest is equivalent to 2.5 SD of sunshine, so multiplying by 2.5 we arrive at the SE(d)=.009573.

The 90% and 95% confidence interval in Figure 2 for the Feddersen study multiply this .009573 SE by 1.64 and 1.96 respectively. So Figure 2 plots a point estimate of d=.0202, with a 90% confidence interval between .0202+1.64*.009573, and a 95% confidence interval of .0202+1.96*.009573.

The null that the effect is $d_{33\%}$=.599 (see previous supplemental subsection for calculation of $d_{33\%}$) is hence quite comfortably rejected. The t-test for that null is t=(.598-.0202)/.009573=60.4. The point estimate is 60 standard errors away from the null, p<.0001. The calculations above must suffer from rounding errors here and there, but, a conclusion based on an effect 60 standard errors from the null is not dependent on them.

Lucas & Lawless

Lucas & Lawless do not find an effect of sunshine that is significant, and all their results are extremely close to 0 with great precision. It does not seem necessary to conduct detailed calculations but we cannot just plot a 0, so some calculations are conducted anyway.

Their tables 2 & 3 contain their key results for rain and sunshine, each contains a few specifications. All t-test are close to 0 (despite the ~1 million observations). For Figure 2 I selected the largest estimate reported in the direction predicted by Schwarz & Clore (many are in the opposite direction).

In particular, the point estimate obtained in a model with extreme rain as the predictor with a point estimate of β=-.002, and SE(β)=.003. With such close to 0 estimates there is little to gain by converting into more comparable units, but nevertheless, we can divide those results by the SD of the life-satisfaction measure, SD=.63, to obtain something measured in the same unit as the other two results in the same figure (SD of life satisfaction). Figure 2 plots the resulting β=.0032, with a 95% confidence interval of +-1.96*.00476 and a 90% confidence interval +-1.64*.00476

Supplement 3.Calculating probability that same sample-size replication of a false-positive finding obtains statistically significantly smaller effect size.

Here I provide details of calculations reported in the text that a replication of the same sample size as an original false-positive finding (so d=0), obtained with *p*=.025, has a ≤34% chance of being statistically significantly different from it. I first introduce the formulas used and then carry out the calculations.

Formulas used.

To compare two effect size estimates, the original study's, $\hat{d}_1$, and replication's, $\hat{d}_2$, I employed the meta-analysis formula for examining heterogeneity among *k* different effect sizes, setting k=2 in this case. The formulas are discussed in detail in chapters 5 and 6 of the textbook by Hedges and Olkin (1985). The equation numbers below match those in that book.

Let $\bar{d}$ be the average of both effect sizes, original and replication, weighted by the inverse of the variances, such that

(6.6) $\bar{d} = \frac{\hat{d}_1/\sigma_1^2+\hat{d}_2/\sigma_2^2}{1/\sigma_1^2+1/\sigma_2^2}$

Where the variance of d_i, σ_i^2, is

(5.15) $\sigma_i^2=\frac{n_i^e+n_i^c}{n_i^e n_i^c}+\frac{\hat{d}_i^2}{2(n_i^e+n_i^c)}$

Where n_i^c and n_i^e, in turn, are the sample sizes in the two conditions of study i.

The null that both effects are identical can be assessed via this test:

(6.25) $Q=\frac{(\hat{d}_1-\bar{d})^2}{\sigma_1^2}+\frac{(\hat{d}_2-\bar{d})^2}{\sigma_2^2}$

With $Q \sim \chi^2(1)$, such that a high enough Q leads to rejecting the null that $\delta_1=\delta_2$

Calculations

We begin by computing, for a given sample size, the effect size of the original study that would lead to *p*=.025. Call it $\hat{d}_1$. Then we find the effect size of the replication, $\hat{d}_2$, that would be just significantly (*p*=.05) different from $\hat{d}_1$ given the same sample size. We then convert that $\hat{d}_2$ into a t-test value, and the *c.d.f.* of the student distribution for that value is the likelihood of obtaining that low or lower an effect size in the replication under the null.

For example, if n=20. We first find the value of a t-test, X, such that t(df)=X leads to a two-sided test to have *p*=.025. That number is X=2.334, because t(38)=2.334, *p*=.025.

We now convert that into a Cohen-d to identify the effect size that leads to *p*=.025 with n=20. From (Eq.1) we find d by replacing t and n for those values, $\hat{d}_1 = \frac{2*2.334}{\sqrt{40}}$ =.738. This means that an original study with n=20 per cell and effect size estimate $\hat{d}_1$=.738 obtains *p*=.025. We now find $\hat{d}_2$, the effect size for the replication, that, if its sample size is also *n*=20, leads Q from equation (6.25) to obtain *p*=.05. In words, we find the effect size for the replication that would be (just) significantly different from the original $\hat{d}_1$=.738.

We do this by solving for $\hat{d}_2$ in (6.25) which in turn requires equations (6.6) and (5.15). obtaining $\hat{d}_2$ = -.154. If one verifies the comparison of $\hat{d}_1$=.738 vs. $\hat{d}_2$=-.154 with those equations one obtains $\chi^2(1)=3.84$, *p*=.05.

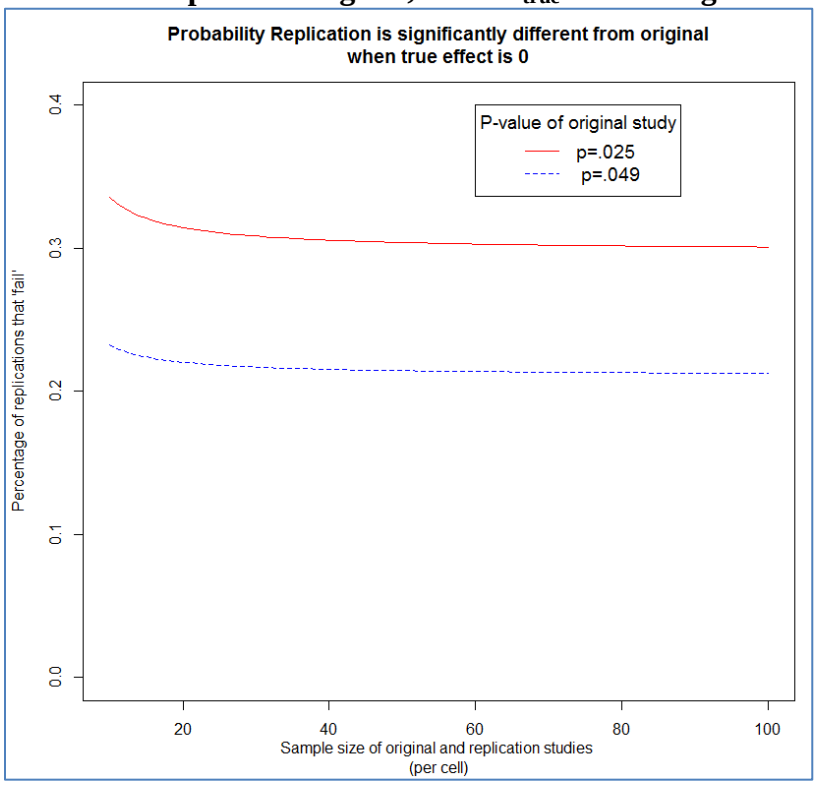
So the replication needs to obtain that point estimate, or lower, for it to be significantly different from $\hat{d}_1$=.738. Under the null that true d=0 we now find out how likely $\hat{d}_2 \leq$-.154 is by again relying on (Eq. 1), now solving for t. This leads to

t(38)=-.475,which is associated with *p*=.32. So the replication obtains a sufficiently low value of $\hat{d}_2$, under the null that the true effect size is d=0, 32% of the time.

This result is a function of sample size (recall that we are assuming the replication is of the exact same size as the original). Figure S1 plots it for sample size ranging between n=10 and n=100 (using R code available here: https://osf.io/adweh/).

As n increases, the percentage of replications that reject d_1 converges to that we would obtain with the normal distribution because the student distribution converges to the normal as n increases.

Figure S1. Proportion of replications obtaining an effect size significantly smaller than the false-positive original, when d_{true}=0 and original obtains *p*=.049 or *p*=.025



Supplement 4. Deriving result that replications with 2.5*original sample size have 80% power to reject d33%.

The goal is to determine the required sample size for a replication to obtain 80% power to reject the null that the underlying effect size is $d_{33\%}$ (the effect size that gives the original study 33% power) against the one-sided alternative that the effect is smaller than that, if the true effect were d=0. In simpler terms: find *n* to make the replication be powered at 80% to conclude it informatively failed, if the true effect being studied does not exist.

As shown below, the required sample size depends on the noncentrality parameter of the underlying test-statistic distribution (e.g., t or F), and the noncentrality parameter is itself a function of sample size, complicating the problem.

I proceed by solving it for the normal distribution, where this simultaneity problem is absent. The solution for the normal distribution indicates replication samples need to be (somewhat greater than) 2.5 times the original to obtain 80% power. I then asses the resulting power of replications with 2.5 times the original sample size for the student distribution and find that it leads to power very close to 80%.

Normal case

Consider a simple case where two means from samples of size *n* each are subjected to a normal test (σ known). The statistical power of that test, for a given α level, for the null that the effect is d_{null}, when it actually is d_{true}, is a function only of μ, where:

(Eq.1) $\mu = \sqrt{\frac{n}{2}}(d_{null}-d_{true})$.

We want to test the null hypothesis H0: the effect is $d=d_{33\%}$ against the one sided alternative H1: the effect is $d<d_{33\%}$. We are interested in the power of the test when the true effect is $d_{true}=0$, leading to:

(Eq. 1') $\mu = \sqrt{\frac{n}{2}}\, d_{null.}$

Let $\mu_{33\%}$ bet the value of μ that leads this two-sided test have 33% power (for $\alpha=.05$). When assuming normality, $\mu_{33\%}$ is a constant, the same for all n. It will be useful to rearrange terms and express $d_{33\%}$ as a function of this constant and the sample size of the original study (n_o):

(Eq.2) $$d_{33\%}= \mu_{33\%}\sqrt{\frac{2}{n_o}}$$

Now, for the replication we wish to test the null that $d_{true}=d_{33\%}$, against the one-sided alternative that $d_{true}<d_{33\%}$ and to obtain power of 80% for rejecting this null if $d_{true}=0$. We shall refer to the constant associated with this level of power, for a one-sided test, with $\mu_{80\%}$, and to the replication's sample size with $n_{R.}$ From (Eq. 1') we have:

(Eq.3) $$\mu_{80\%}= \sqrt{\frac{n_R}{2}}d_{33\%}.$$

Manipulating (Eq.3) to leave n_R in the left-hand-side, and substituting the right-hand-side of (Eq.2) for $d_{33\%}$ we arrive at an expression of n_R as a function of n_O:

(Eq.4) $$n_R=n_O * (\frac{\mu_{80\%}}{\mu_{33\%}})^2$$

Again, $\mu_{80\%}$ and $\mu_{33\%}$ are constants, because of the normality assumption, and they are easy to compute (e.g., using standard power formulas or software like gPower; Faul, Erdfelder, Lang, & Buchner, 2007). Their values are $\mu_{33\%}= 1.528$ and $\mu_{80\%} = 2.486$ leading to:

(Eq.4') n_R=n_O*2.64.

In words, for a test based on the normal distribution, to obtain 80% power of rejecting $d_{33\%}$ against the one-sided alternative of $d<d_{33\%}$, if $d_{true}=0$, the replication needs to include 2.64 times as many participants as the original study did.

For tests where σ is not known, e.g., two sample t-test, ANOVA, a given increase in sample size leads to greater increases in power, so the ratio will be slightly lower than 2.64, hence the rule of thumb 2.5. Figure 4 in the paper documents the performance of the 2.5 rule of thumb on the student distribution and $\chi^2(1)$ test performed on binary data. The R Code behind that figure I available here https://osf.io/adweh/files/

Supplement 5. Actual power of replications using reported effect size in original study to set sample size in replication.

I discuss the approach using an original study with n=20 per cell and with underlying power of 50%. The R program used to generate Figure 3, and posted online here: https://osf.io/adweh/files/, allows users to set both of these parameters so that the results can be easily extended.

When n=20, 50% power is obtained if the original effect size is d_{true}=.64. Of course the estimated effect size will not be exactly $\hat{d}$=.64. In fact, because if $\hat{d}$=.64 we get exactly *p*=.05, *all* significant estimates will be larger than .64, and hence *all* replications will have less than the desired 80% power.

For example, if the original study obtained $\hat{d}$=.75 (again d_{true} = .64, n=20) then a replicator setting 80% power would set sample size by finding the *n* that gives 80% to d=.75, which happens to be n_{rep}=29. So the replication would have n=29 compared to the original n=20, but because d_{true}=.64<.75, the actual power of the replication is <80%, in fact, it is just 67%.The R program conducts this calculation for the distribution of possible effect size estimates if d_{true}=.64, gets the sample sizes that would give those estimates 80% power, and then computes the effective power for that distribution of possible effect size. Using that distribution I then compute the expected actual power (the mean of the distribution) and the share <33%, <50%, <80% and <90%.

The results vary slightly as we increase the original sample size, until the student distribution converges to the normal.

#R PROGRAM

```
library(pwr)     #Need to install this package
#FUNCTION 1
#  Find sample size given desired power and effect size
  getn = function(d,power) {
    n=pwr.t.test(d=d,power=power)$n
    return(round(n,digits=0))
#FUNCTION 2
# Find power for given sample size and effect size
  getpower = function(d,n) {return(pwr.t.test(d=d,n=n)$power)}

#FUNCTION 3
#MAIN FUNCTION FOR ALL CALCULATIONS
realpower=function(nori,powerori,powerrep) {
  #SYNTAX:
  #nori: per cell sample size of original study (comparing two means)
  #powerori: true power of original study
  #powerrep: claimed power of repliation

  #compute df, ncp, tc, true effect size
    #degrees of freedom
      dfori=2*nori-2
    #true effect size leading to set power for original study
      dori=pwr.t.test(n=nori,power=powerori)$d
    #noncentrality parameter for student distribution n and d
      ncpori=sqrt(nori)*dori
    #critical t-value for p=.05 given original sample size
      tc=qt(.975,dfori)

  #generate original studies
  #the way I do this is to get the distribution of possible t-values obtained,
I start creating a vector with percentiles between 1-power and 99.9%
    rp=seq(from=1-powerori+.001, to=.999,by=.001)
  #I then find the noncentral t-values associated with each of those
percentiles, effectively the noncentral distribution every 1/1000 point
    t_publish=qt(p=rp,df=dfori,ncp=sqrt(nori/2)*dori)
  #convert that distribution of t-values into effect sizes
    d_publish=(2*t_publish)/sqrt(dfori)

  #Find nrep for powerrep
  #for each of the possible values obtained, I compute the sample size a
replicator would set seeking the claimed level of power
    nrep=mapply(getn,d=d_publish,power=powerrep)  #nrep: sample size of
replication

  #find actual power
  #the true level of power is different from the published, that's why actual
power is not claimed power, i here compute actual power
    real=mapply(getpower,d=dori,n=nrep)  #for every sample size run by the
replicator, i compute the effective power

    m=mean(real)  #every sample size in the vector is equally likely (by
construction, becuase of line 38 of code above) so the simple mean
                  #is the expected value of power
  cat("Mean power:       ",m,"\n")
  }
```

```
#Underlying power is 50%
  realpower(nori=20,powerori=.5,powerrep=.80)
  realpower(nori=20,powerori=.5,powerrep=.90)
  realpower(nori=20,powerori=.5,powerrep=.95)

   #Robust to different sample size (within 3%)
      realpower(nori=100,powerori=.5,powerrep=.80)
      realpower(nori=100,powerori=.5,powerrep=.90)
      realpower(nori=100,powerori=.5,powerrep=.95)


#Underlying power is 35
  realpower(nori=20,powerori=.35,powerrep=.80)
  realpower(nori=20,powerori=.35,powerrep=.90)
  realpower(nori=20,powerori=.35,powerrep=.95)

   #Robust to different sample size (within 3%)
    realpower(nori=100,powerori=.35,powerrep=.80)
    realpower(nori=100,powerori=.35,powerrep=.90)
    realpower(nori=100,powerori=.35,powerrep=.95)
```

Supplement 6. Details on the summary statistics of sample size in *Psychological Science* 2003-2010

For an unrelated project with a University of Pennsylvania undergraduate student, Daniel Li, we compiled a dataset that included all tests statistics reported in Psych Science between 2003 and 2010. These originally consisted of free text extracted from the published papers. It was later parsed into subfield that includes variables that indicate the test t, F, χ^2 or n, the degrees of freedom and the value of the test statistic.

Our dataset includes 11,292 test results, of which 4,275 are t-tests (i.e., those relying on the student distribution). The dataset does not include information on the origin of the test, they could consists of differences of means, point estimates in regressions, planned contrasts, linear trends, etc. To proxy for sample size we treat all these tests as comparing two conditions (some undoubtedly consist of within subject comparison, some of multiple cell designs) leading to the per-condition imputed sample size of n=(df+2)/2.

The median sample size so computed is 19.35 (the paper reports it as "about 20"). 94.27% of tests would involve samples n<150, the paper describes it as "about 95%"

These are undoubtedly noisy estimates, but because the reporting of sample size is not standardized across papers, it is difficult to automatize the collection of these numbers. The collection of d.f. was trivial to execute.

References

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.

Feddersen, J., Metcalfe, R., & Wooden, M. (2012). Subjective Well-Being: Weather Matters; Climate Doesn't.

Gámez, E., Díaz, J. M., & Marrero, H. (2011). The Uncertain Universality of the Macbeth Effect with a Spanish Sample. *The Spanish journal of psychology*(1), 156-162.

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*: Academic Press United Kingdom.

Schwarz, N., & Clore, G. (1983). Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States. *Journal of Personality and Social Psychology, 45*(3), 513-523.

Siev, J. (2012, December 20th, 2012).

Smithson, M. (2003). *Confidence Intervals*: SAGE Publications, Incorporated.

Zhong, C. B., & Liljenquist, K. (2006). Washing Away Your Sins: Threatened Morality and Physical Cleansing. *Science, 313*(5792), 1451.