Conducting rigorous research on large open-access developmental datasets

Amy Orben

Department of Experimental Psychology, University of Oxford

ABCD Workshop, Portland

@OrbenAmy

1. *Curbing analytical flexibility*
2. *Preregistration + Registered Reports*
3. *Specification Curve Analysis*
4. *Effect Sizes*

Derren Brown: The System

(Kate Button)

While there was a system to guarantee that she won,

it wasn’t the system she thought it was.

Race 1: 7776 people, randomly allocated a horse

Race 1: 7776 people, randomly allocated a horse

Race 2: 1296 race 1 winners, randomly allocated a horse

Race 1: 7776 people, randomly allocated a horse

Race 2: 1296 race 1 winners, randomly allocated a horse

Race 3: 216 race 2 winners, randomly allocated a horse

Race 1: 7776 people, randomly allocated a horse

Race 2: 1296 race 1 winners, randomly allocated a horse

Race 3: 216 race 2 winners, randomly allocated a horse

Race 4: 36 race 3 winners, randomly allocated a horse

Race 1: 7776 people, randomly allocated a horse

Race 2: 1296 race 1 winners, randomly allocated a horse

Race 3: 216 race 2 winners, randomly allocated a horse

Race 4: 36 race 3 winners, randomly allocated a horse

Race 5: 6 race 4 winners, randomly allocated a horse

Race 1: 7776 people, randomly allocated a horse

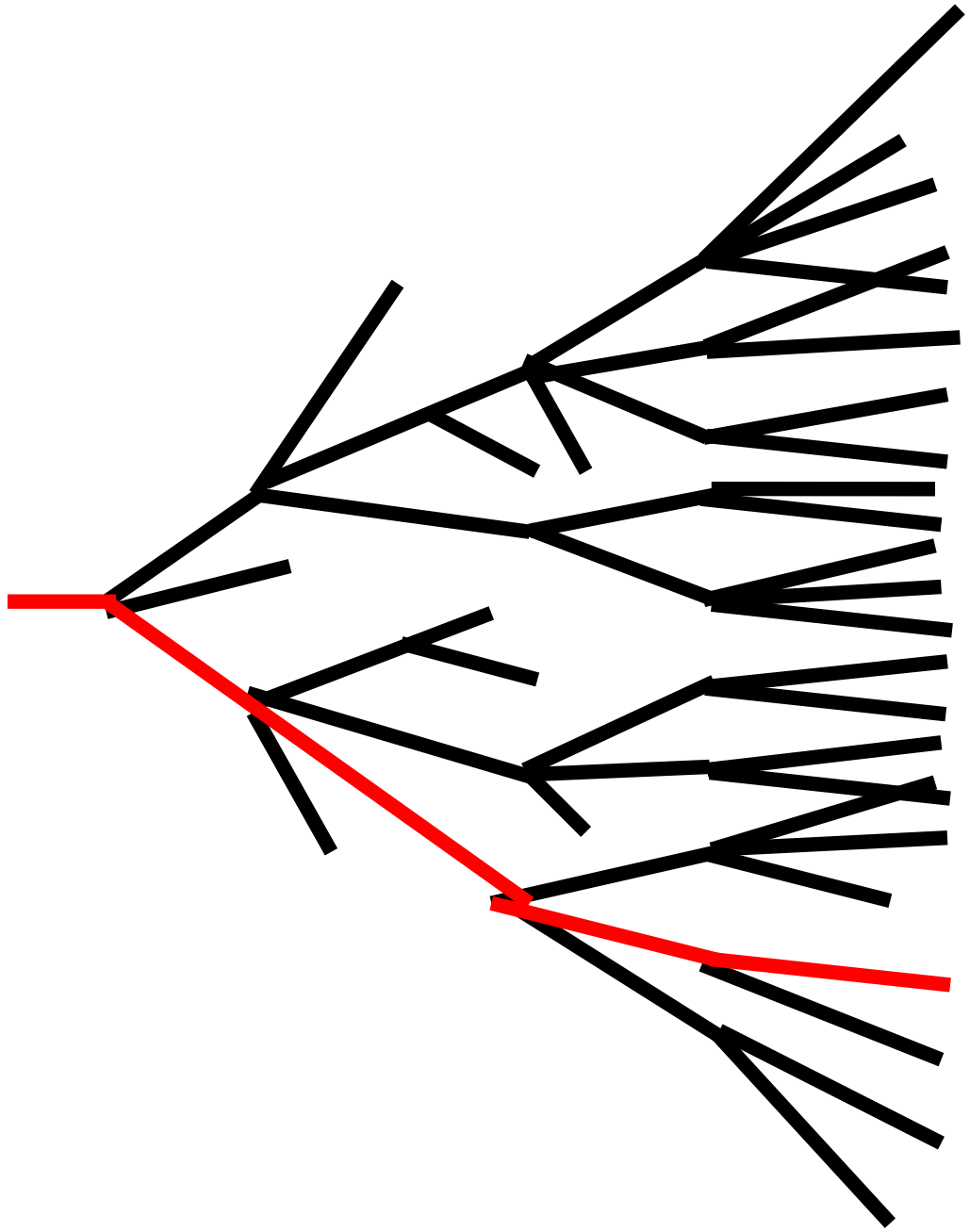
Race 2: 1296 race 1 winners, randomly allocated a horse

Race 3: 216 race 2 winners, randomly allocated a horse

Race 4: 36 race 3 winners, randomly allocated a horse

Race 5: 6 race 4 winners, randomly allocated a horse

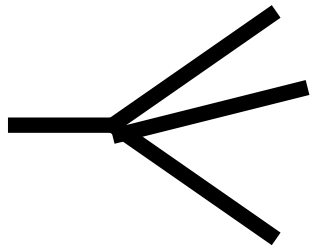
She was the 1 / 7776 who by chance had 5 consecutive wins

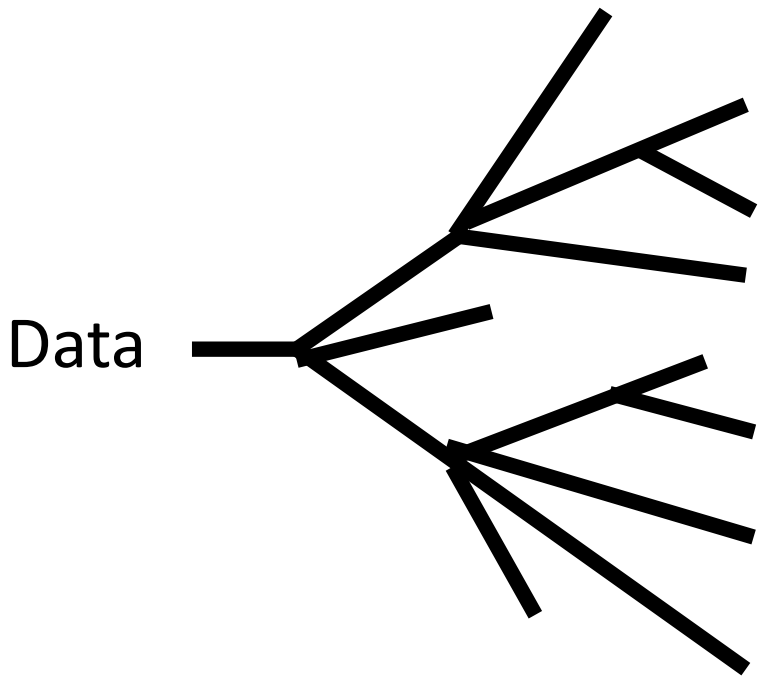


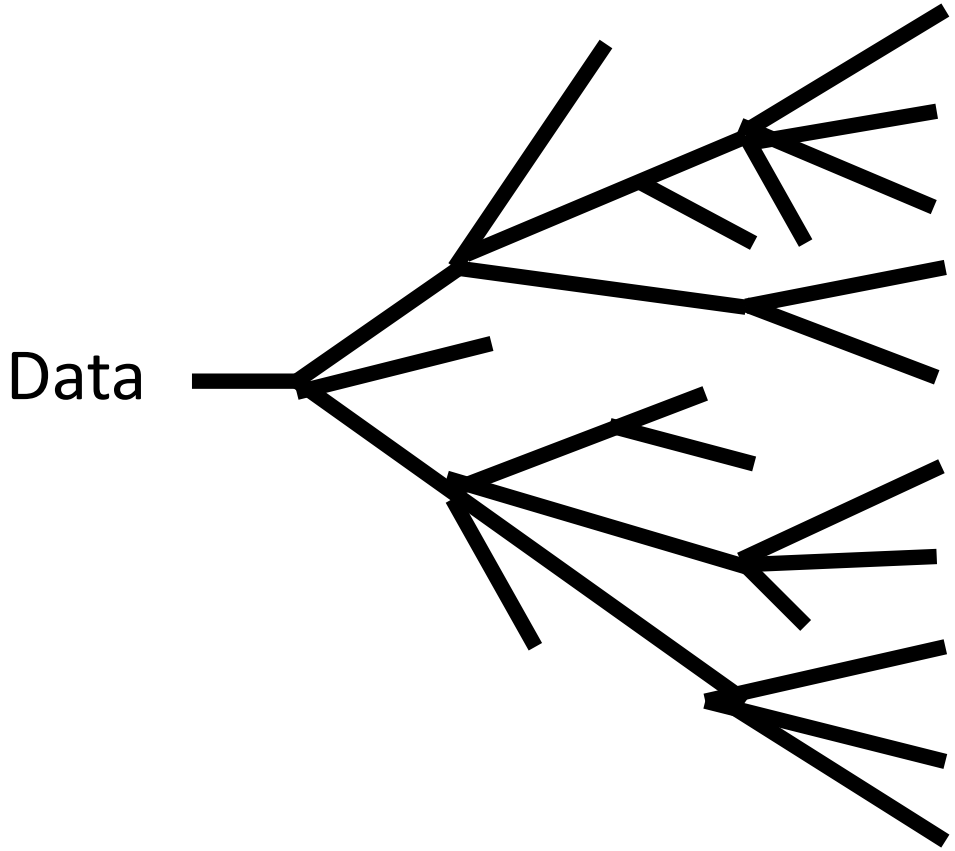
The “Winning Streak”

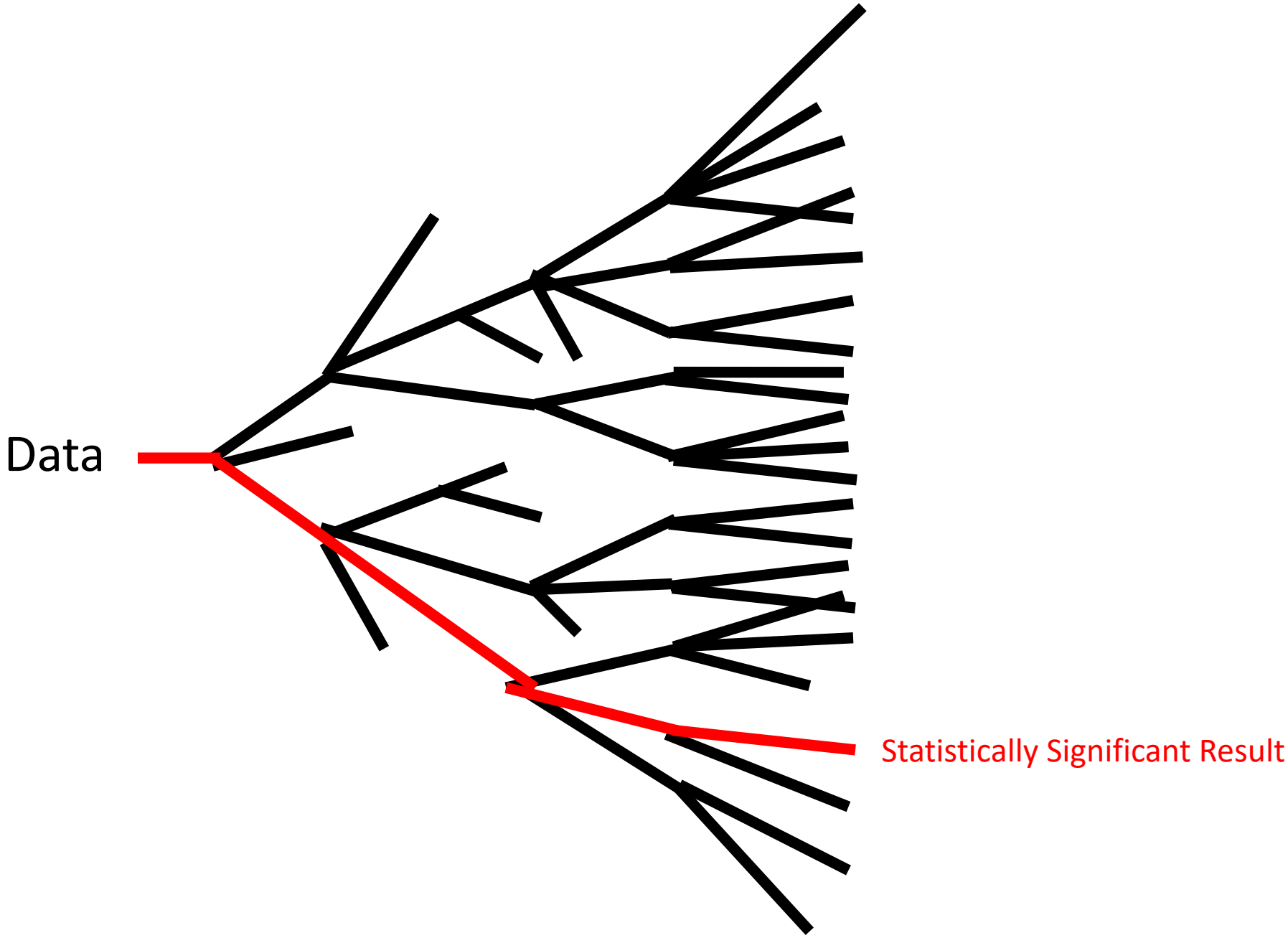
Data

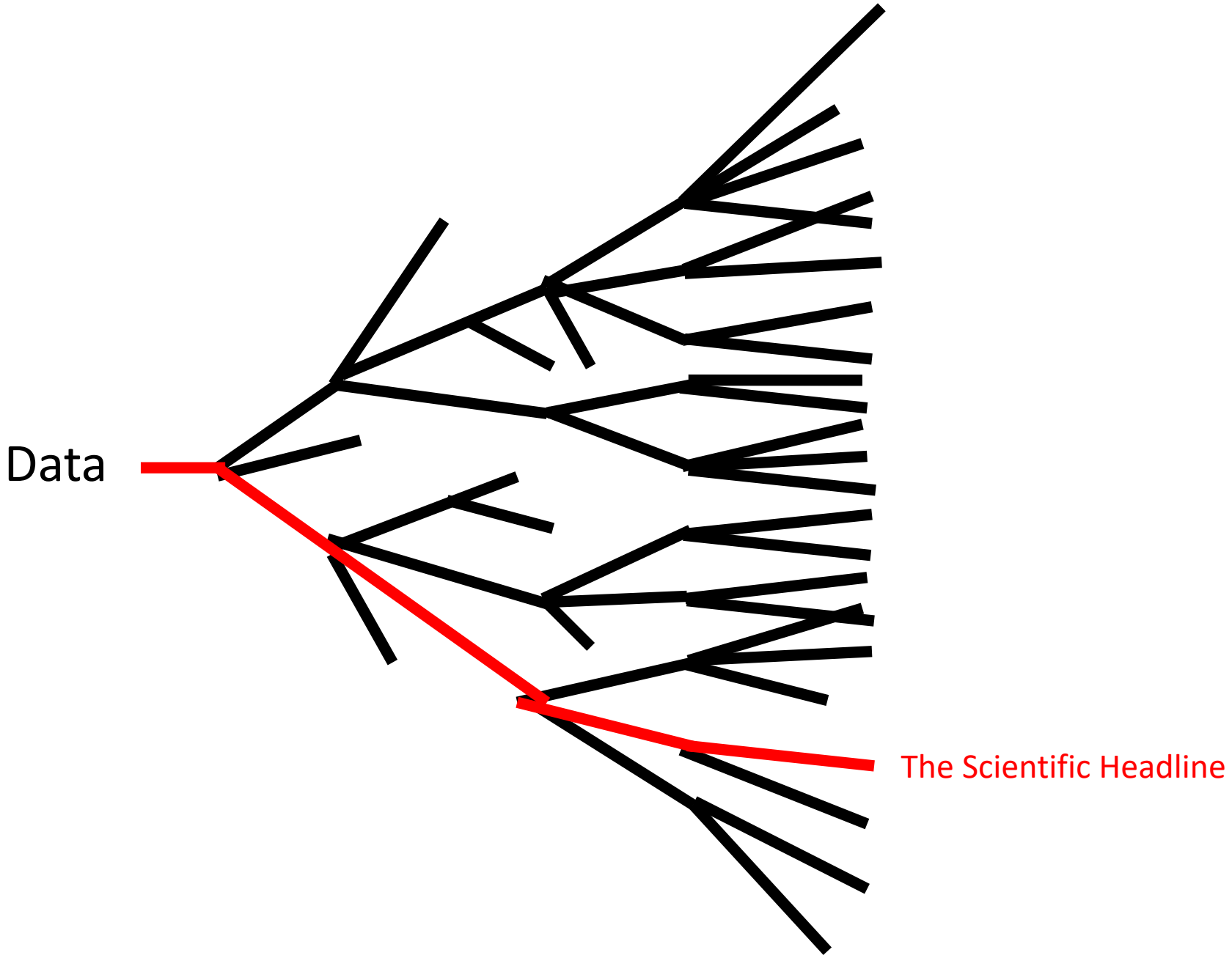
Gelman: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Data 


Data

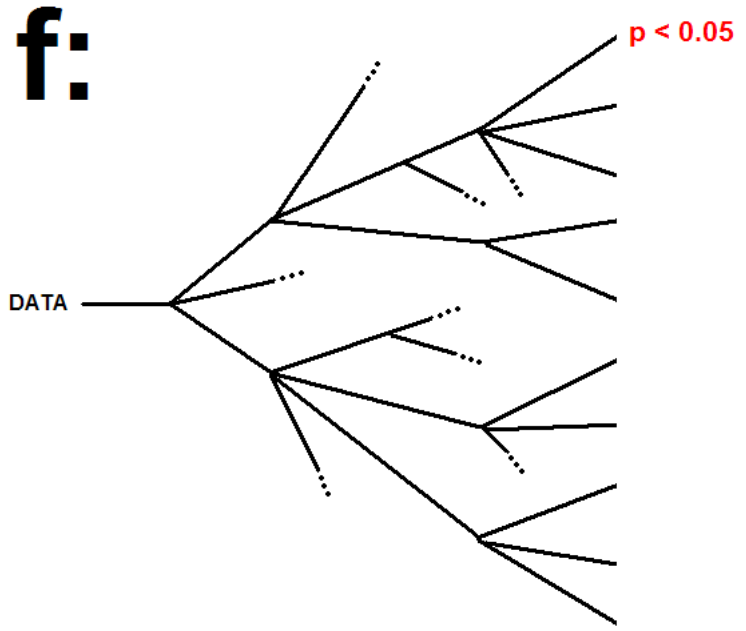

Data


Data
Statistically Significant Result


Data
The Scientific Headline

Garden of Forking Paths

“The researcher degrees of freedom do not feel like degrees of freedom because, conditional on the data, each choice appears to be deterministic. But if we average over all possible data that could have occurred, we need to look at the entire garden of forking paths and recognize how each path can lead to statistical significance in its own way."



Gelman: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

General Article

Psychological Science
22(11) 1359–1366

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

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

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

Abstract

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

Keywords

methodology, motivated reasoning, publication, disclosure

Does listening to the song ”When I’m Sixty-Four” cause people to become older?

20 University of Pennsylvania undergraduates

“When I’m Sixty-Four” or “Kalimba”

Indicate birthday and father’s age (control for baseline age across participants)

Does listening to the song ”When I’m Sixty-Four” cause people to become older?

20 University of Pennsylvania undergraduates

“When I’m Sixty-Four” or “Kalimba”

Indicate birthday and father’s age (control for baseline age across participants)

People were 1½ years younger after “When I’m Sixty-Four” $F(1,17) = 4.92, p = 0.040$

Table 3. Study 2: Original Report (in Bolded Text) and the Requirement-Compliant Report (With Addition of Gray Text)

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

An ANCOVA revealed the predicted effect: According to their birth dates, people were nearly a year-and-a-half younger after listening to "When I'm Sixty-Four" (adjusted *M* = 20.1 years) rather than to "Kalimba" (adjusted *M* = 21.5 years), *F*(1, 17) = 4.92, *p* = .040. Without controlling for father's age, the age difference was smaller and did not reach significance (*M*s = 20.3 and 21.2, respectively), *F*(1, 18) = 1.01, *p* = .33.

Simmons, Nelson, Simonsohn (2011)

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Simmons, Nelson, Simonsohn (2011)

Empirical Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(3) 337–356

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245917747646
www.psychologicalscience.org/AMPPS

SAGE

Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results

R. Silberzahn[1], E. L. Uhlmann[2], D. P. Martin[3], P. Anselmi[4], F. Aust[5], E. Awtrey[6], Š. Bahník[7], F. Bai[8], C. Bannard[9], E. Bonnier[10], R. Carlsson[11], F. Cheung[12], G. Christensen[13], R. Clay[14], M. A. Craig[15], A. Dalla Rosa[4], L. Dam[16], M. H. Evans[17], I. Flores Cervantes[18], N. Fong[19], M. Gamez-Djokic[20], A. Glenz[21], S. Gordon-McKeon[22], T. J. Heaton[23], K. Hederos[24], M. Heene[25], A. J. Hofelich Mohr[26], F. Högden[5], K. Hui[27], M. Johannesson[10], J. Kalodimos[28], E. Kaszubowski[29], D. M. Kennedy[30], R. Lei[15], T. A. Lindsay[26], S. Liverani[31], C. R. Madan[32], D. Molden[33], E. Molleman[16], R. D. Morey[34], L. B. Mulder[16], B. R. Nijstad[16], N. G. Pope[35], B. Pope[36], J. M. Prenoveau[37], F. Rink[16], E. Robusto[4], H. Roderique[38], A. Sandberg[24], E. Schlüter[39], F. D. Schönbrodt[25], M. F. Sherman[37], S. A. Sommer[40], K. Sotak[41], S. Spain[42], C. Spörlein[43], T. Stafford[44], L. Stefanutti[4], S. Tauber[16], J. Ullrich[21], M. Vianello[4], E.-J. Wagenmakers[45], M. Witkowiak[46], S. Yoon[19], and B. A. Nosek[3,47]

Project Stage	Work Package	Month																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	Building the Data Set																				
2	Recruitment and Initial Survey of Data Analysts																				
3	First Round of Data Analysis																				
4	Round-Robin Peer Evaluations																				
5	Second Round of Data Analysis																				
6a	Open Discussion and Debate, Further Analyses																				
6b	Write-Up of Manuscript																				
7	Internal Experts' Peer Review of Approaches																				
	Revision of Manuscript																				

Fig. 1. Overview of the project's stages.

Team	Analytic Approach	Odds Ratio
12	Zero-Inflated Poisson Regression	0.89
17	Bayesian Logistic Regression	0.96
15	Hierarchical Log-Linear Modeling	1.02
10	Multilevel Regression and Logistic Regression	1.03
18	Hierarchical Bayes Model	1.10
31	Logistic Regression	1.12
1	OLS Regression With Robust Standard Errors, Logistic Regression	1.18
4	Spearman Correlation	1.21
14	WLS Regression With Clustered Standard Errors	1.21
11	Multiple Linear Regression	1.25
30	Clustered Robust Binomial Logistic Regression	1.28
6	Linear Probability Model	1.28
26	Hierarchical Generalized Linear Modeling With Poisson Sampling	1.30
3	Multilevel Logistic Regression Using Bayesian Inference	1.31
23	Mixed-Model Logistic Regression	1.31
16	Hierarchical Poisson Regression	1.32
2	Linear Probability Model, Logistic Regression	1.34
5	Generalized Linear Mixed Models	1.38
24	Multilevel Logistic Regression	1.38
28	Mixed-Effects Logistic Regression	1.38
32	Generalized Linear Models for Binary Data	1.39
8	Negative Binomial Regression With a Log Link	1.39
20	Cross-Classified Multilevel Negative Binomial Model	1.40
13	Poisson Multilevel Modeling	1.41
25	Multilevel Logistic Binomial Regression	1.42
9	Generalized Linear Mixed-Effects Models With a Logit Link	1.48
7	Dirichlet-Process Bayesian Clustering	1.71
21	Tobit Regression	2.88
27	Poisson Regression	2.93

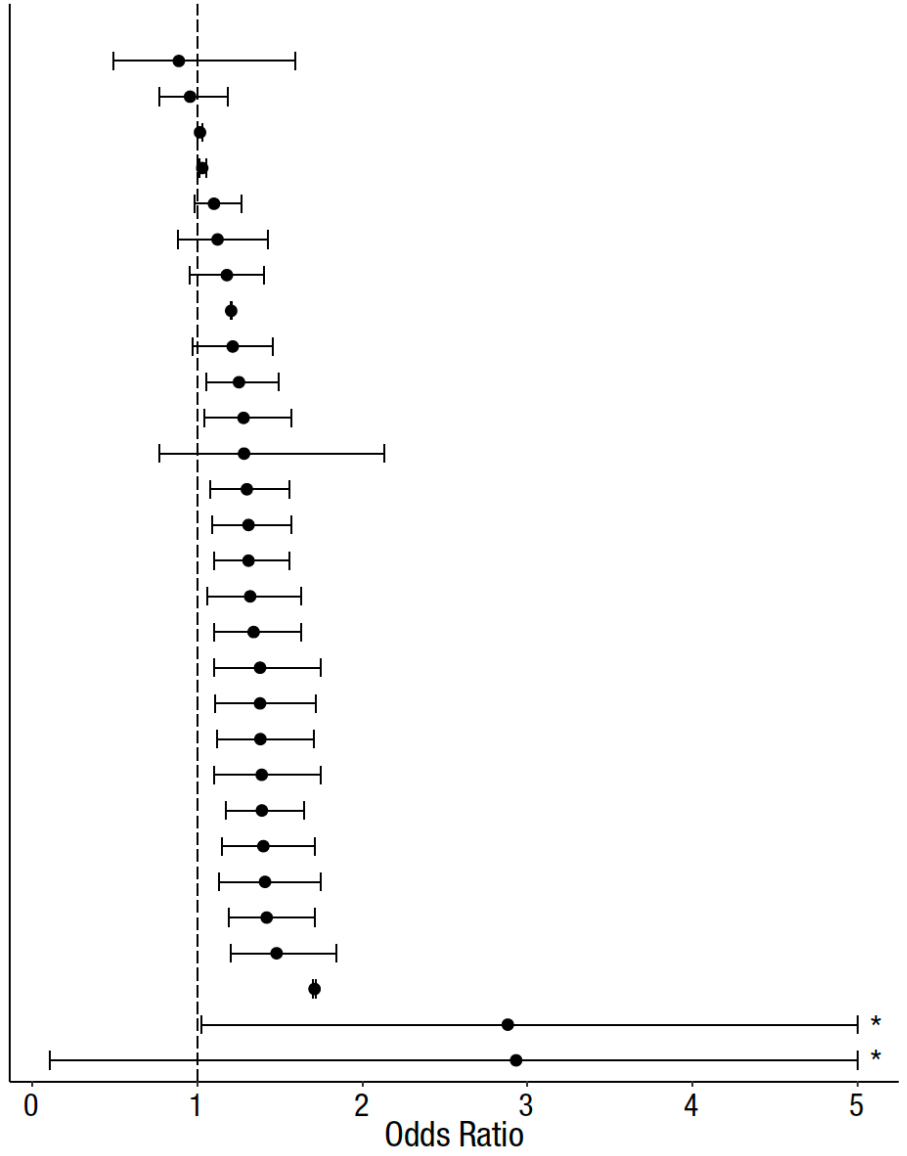


Fig. 2. Point estimates (in order of magnitude) and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees. Reported results, along with the analytic approach taken, are shown for each of the 29 analytic teams. The teams are ordered so that the smallest reported effect size is at the top and the largest is at the bottom. The asterisks indicate upper bounds that have been truncated to increase the interpretability of the plot; the actual upper bounds of the confidence intervals were 11.47 for Team 21 and 78.66 for Team 27. OLS = ordinary least squares; WLS = weighted least squares.

Why might these problems be amplified by large-scale openly accessible data?

An Example

Psychologically, however, they are more vulnerable than Millennials were: Rates of teen depression and suicide have skyrocketed since 2011. It's not an exaggeration to describe iGen as being on the brink of the worst mental-health crisis in decades. Much of this deterioration can be traced to their phones.

Smartphones
The Observer

Interview

Are smartphones really making our children sad?

By Interview by Ian Tucker

US psychologist Jean Twenge, who has claimed that social media is having a malign affect on the young, answers critics who accuse her of crying wolf

News

Jeremy Hunt: Social media poses as great a threat to children as obesity

share 5

Science must begin with myths, and with the criticism of myths.

Karl Popper

News › Education › Education News

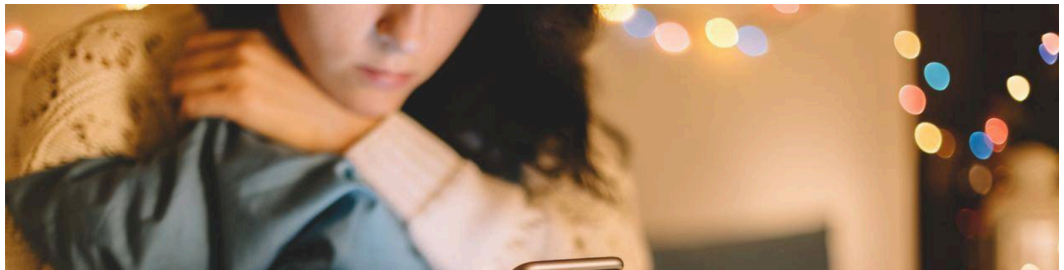
Giving your child a smartphone is like giving them a gram of cocaine, says top addiction expert

Harley Street clinic director Mandy Saligari says many of her patients are 13-year-old girls who see sexting as 'normal'

Rachael Pells Education Correspondent | @rachaelpells | Wednesday 7 June 2017 16:29 BST | 3 comments

149K shares

Like Click to follow The Independent Online



Psychologically, however, they are more vulnerable than Millennials were: Rates of teen depression and suicide have skyrocketed since 2011. It's not an exaggeration to describe iGen as being on the brink of the worst mental-health crisis in decades. Much of this deterioration can be traced to their phones.

Have Smartphones Destroyed a Generation?

More comfortable online than out partying, post-Millennials are safer, physically, than adolescents have ever been. But they're on the brink of a mental-health crisis.

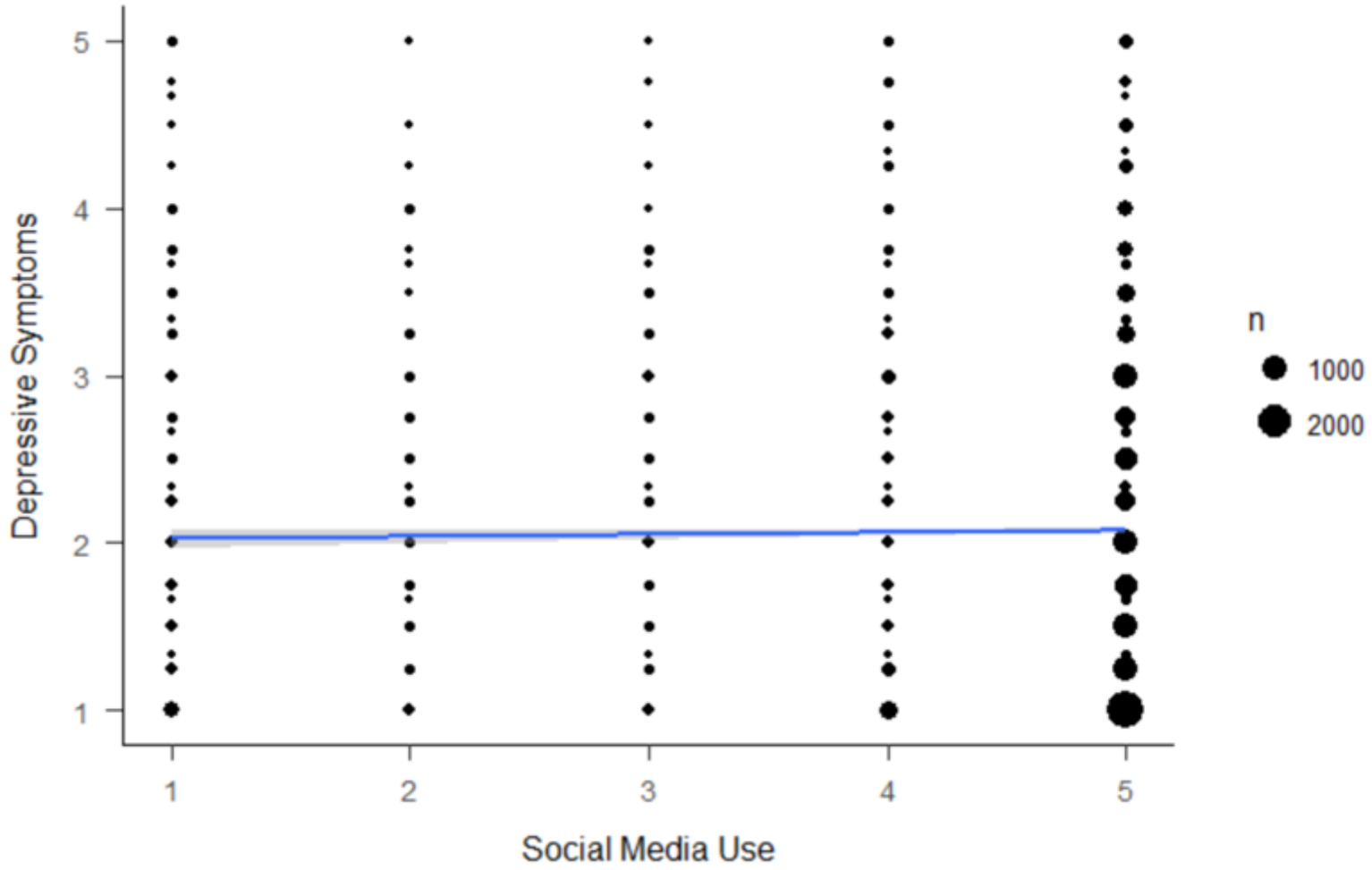
Empirical Article

Clinical Psychological Science
2018, Vol. 6(1) 3–17

DOI: 10.1177/2167702617723376
www.psychologicalscience.org/CPS

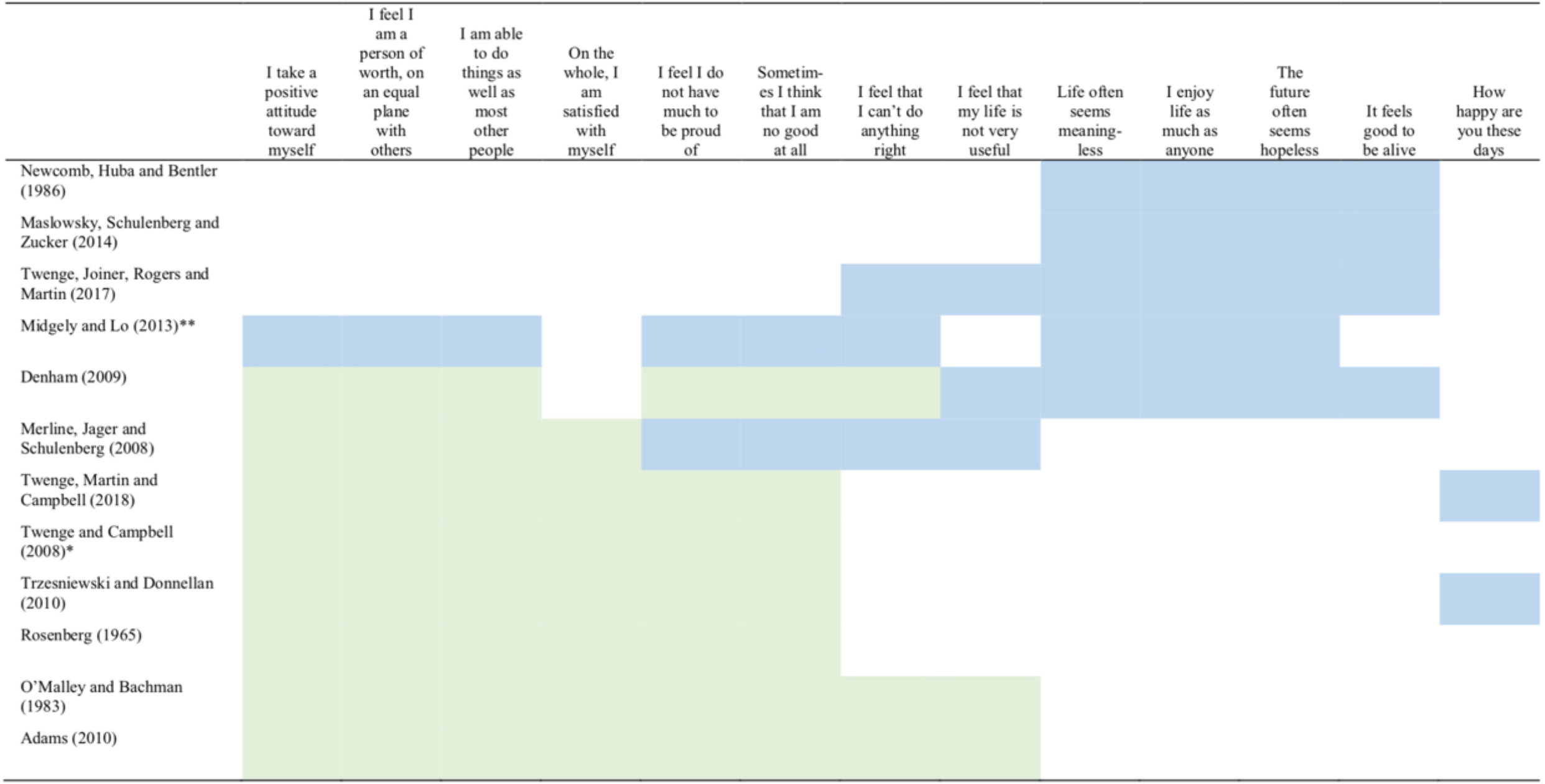
Increases in Depressive Symptoms, Suicide-Related Outcomes, and Suicide Rates Among U.S. Adolescents After 2010 and Links to Increased New Media Screen Time

Jean M. Twenge[1], Thomas E. Joiner[2], Megan L. Rogers[2], and Gabrielle N. Martin[1]
[1]San Diego State University and [2]Florida State University



Data from Twenge et al. (2017), Orben (2017)

Big Data – Small Effects



Orben and Przybylski (Nature Human Behaviour, 2019)

The Garden of Forking Paths

Data that is ”Too Big To Fail”

- **Large numbers of participants** ensure that even extremely modest covariations (e.g. *r’*s < 0.05) between self-report items will result in alpha levels typically interpreted as compelling evidence for rejecting the null hypothesis by psychological scientists (i.e. *p’*s < 0.05)
- **Large batteries** of **ill-defined questions** lead to an explosion of possible analytical pathways (researcher degrees of freedom)

Orben and Przybylski (Nature Human Behaviour, 2019)

What can we do?

Solutions to Analytical Flexibility

- Transparency:
  - Amount of variables
  - Termination rules
  - All experimental conditions
  - Observations that are eliminated
  - Covariates

Simmons, Nelson, Simonsohn (2011)

The 21-Word Solution

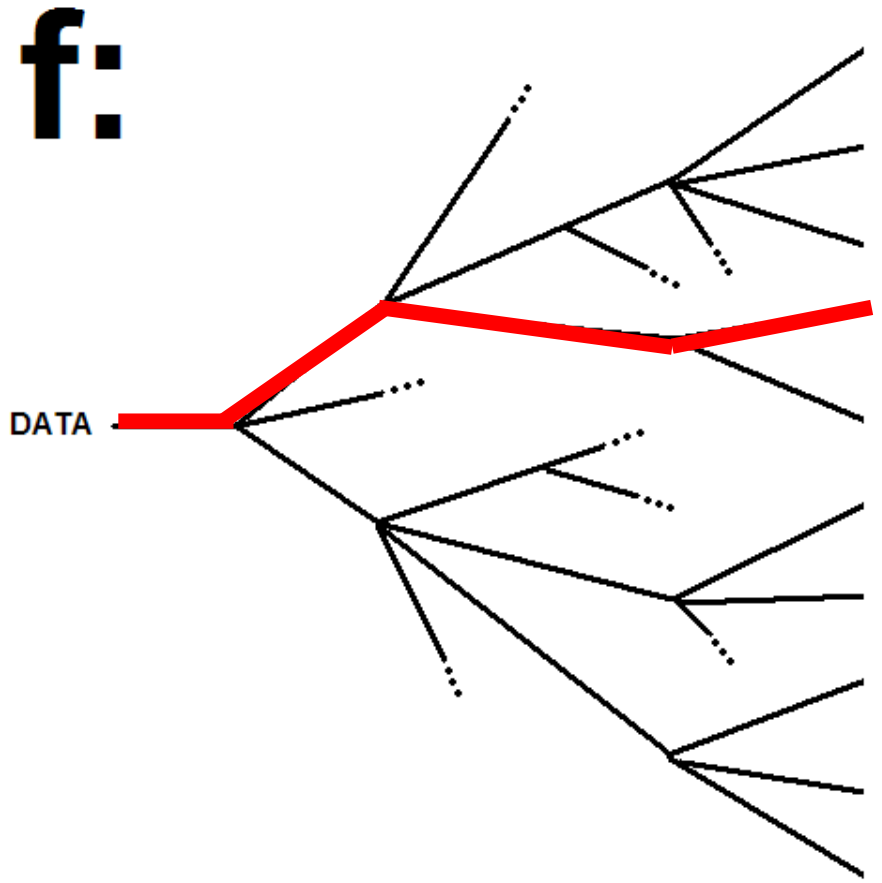
We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Solution #1

Decide on one analytical pathway beforehand using pre-registration or registered report methodologies

(Chambers, 2013; Munafò et al., 2017; van 't Veer, 2016; Lakens, 2014)

Pro: Simple way to decrease researcher degrees of freedom



http://blogs.discovermagazine.com/neuroskeptic/2013/10/16/the-f-problem/

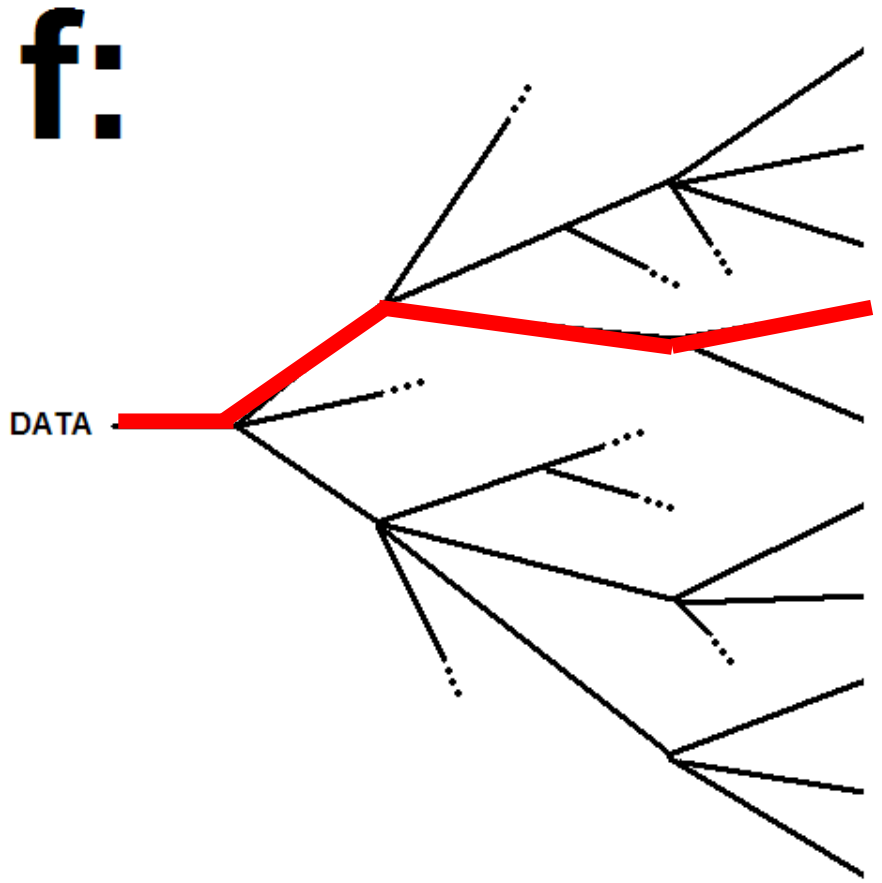
Solution #1

Decide on one analytical pathway beforehand using pre-registration or registered report methodologies

(Chambers, 2013; Munafò et al., 2017; van ’t Veer, 2016; Lakens, 2014)

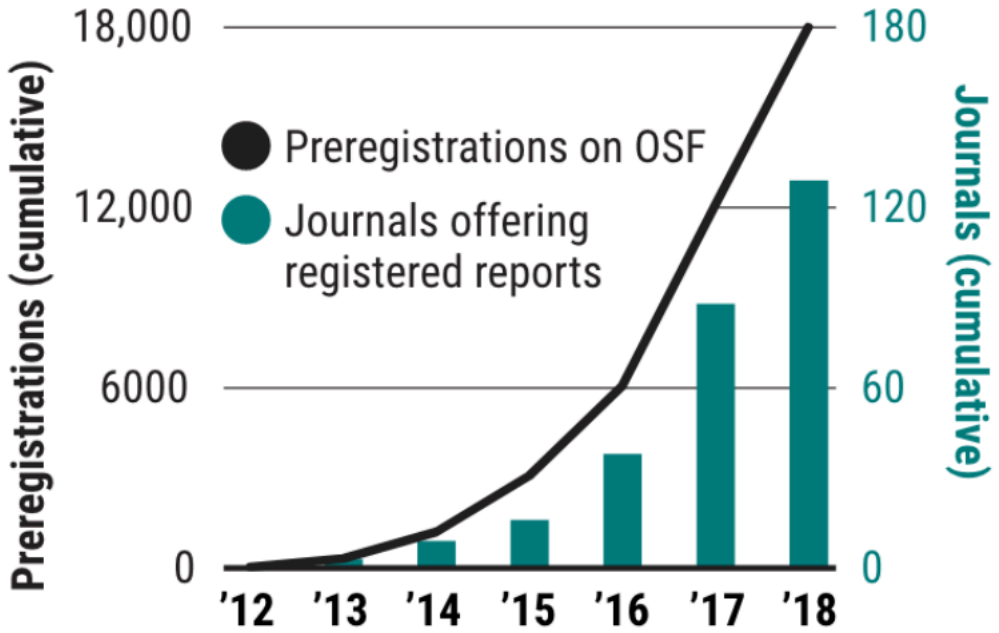
Pro: Simple way to decrease researcher degrees of freedom

Con: Researcher needs to prove that they have not previously seen or engaged with the data



Preregistration

Study preregistrations on the Open Science Framework (OSF) are doubling every year; more than 120 journals have introduced registered reports.



J. YOU/*SCIENCE*

Registered Reports

CORTEX 49 (2013) 609–610

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Editorial

Registered Reports: A new publishing initiative at *Cortex*

Christopher D. Chambers

Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff University, United Kingdom

Four central aspects of the Registered Reports model:

- Researchers decide hypotheses, experimental procedures, and main analyses *before* data collection
- Part of the peer review process takes place before experiments are conducted
- Passing this stage of review virtually guarantees publication
- Original studies and high-value replications are welcome

Stage 1 at *Cortex*

Title

The relationship between functional brain organization and self-reported digital screen engagement in late childhood

Authors

Kathryn L. Mills[1], Amy Orben[2], Andrew K. Przybylski[2,3]

[1] Department of Psychology, University of Oregon, 1227 University of Oregon Eugene OR, 97403, United States

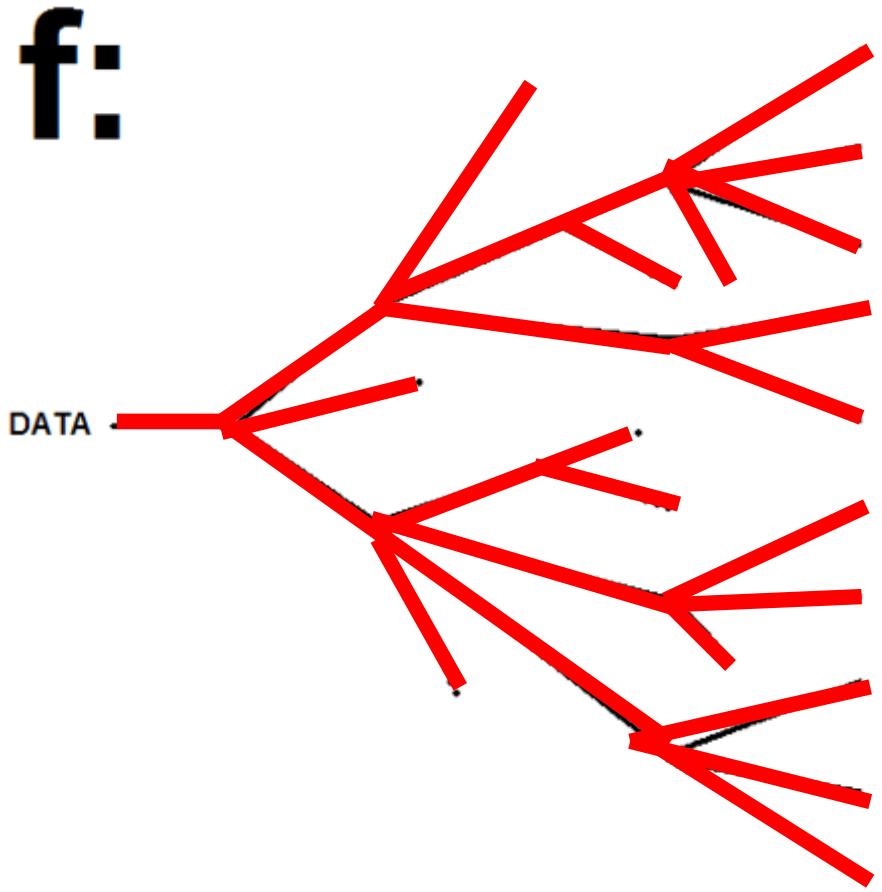
[2] Department of Experimental Psychology, University of Oxford, Anna Watts Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford OX2 6HG, United Kingdom

[3] Oxford Internet Institute, University of Oxford, 1 St Giles Oxford, OX1 3JS, United Kingdom

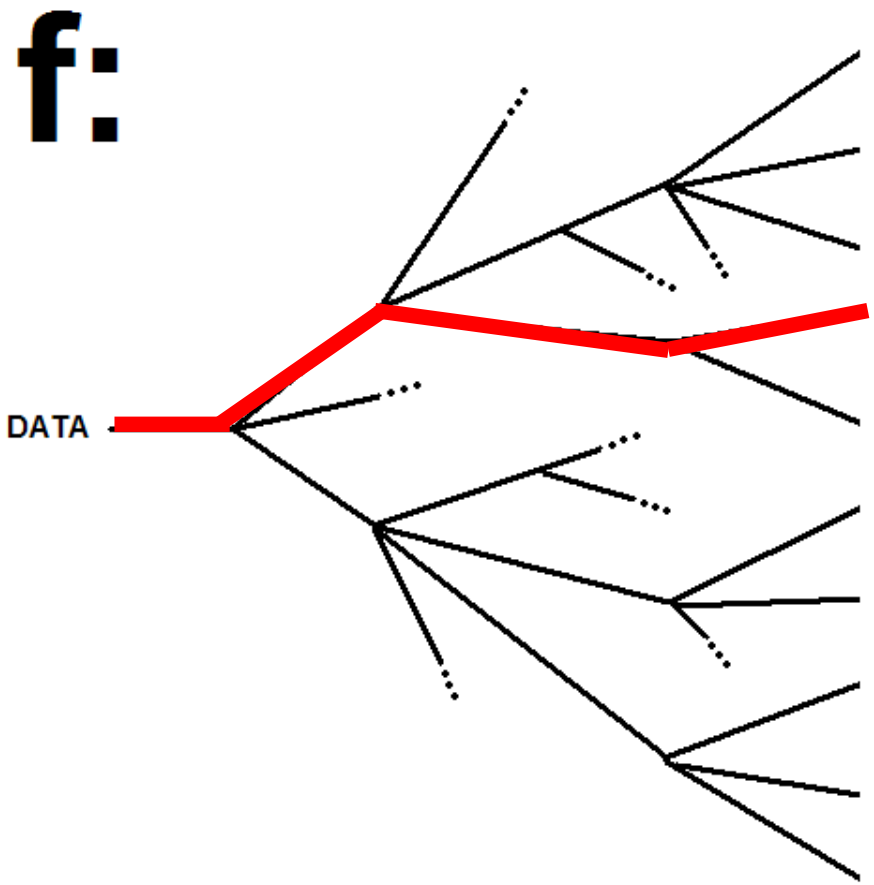
Solution #2

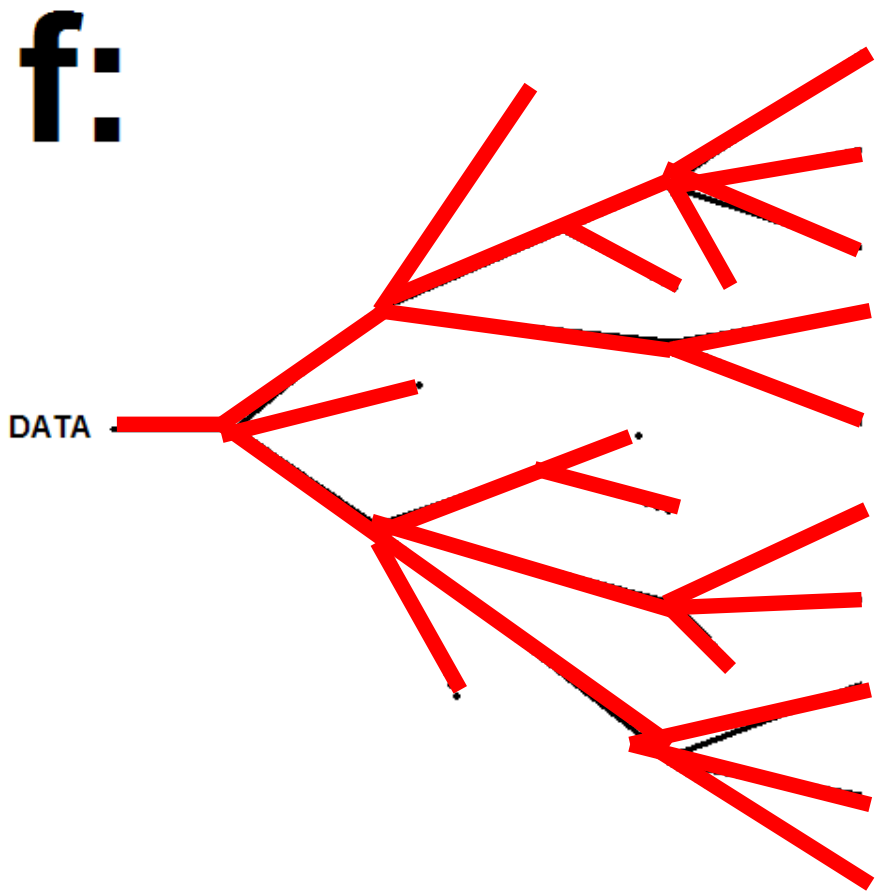
Examine all possible analytical pathways using Specification Curve Analysis
(SCA; Simonsohn, Simmons, & Nelson, 2015)

Pro: Works around researcher degrees of freedom even when data has been previously accessed



Simmonsohn, Simmons, Nelson (2015)





Simmonsohn, Simmons, Nelson (2015)

1
Identify Specifications

Decide on all possible analytical pathways

2
Implementing Specifications

Run all possible analyses and graph outcomes

3
Statistical Inferences

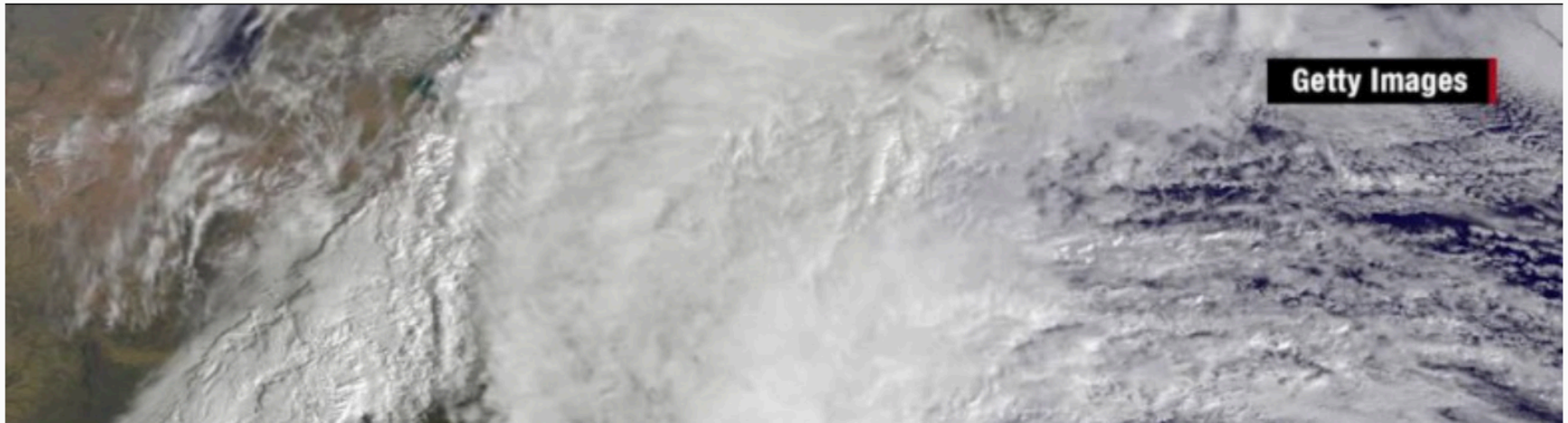
Run bootstraps to test whether original dataset has more significant specifications than a dataset where null hypothesis is true

Extreme Weather

Female hurricanes are deadlier than male hurricanes, study says

By Holly Yan, CNN

Updated 2127 GMT (0527 HKT) September 1, 2016



Are female hurricanes really deadlier than male hurricanes?

Jung et al. (1) claim to show that "feminine-named hurricanes cause significantly more deaths than do masculine-named hurricanes" (p. 1). This conclusion is mainly obtained by analyzing data on fatalities caused by hurricanes in the United States (1950–2012). By reanalyzing the same data, we show that the conclusion is based on biased presentation and invalid statistics.

The reasoning in ref. 1 is fundamentally based on the regression models reported in their table S2, in particular, model 4. However, due to the interaction terms combined with extreme values and weak significance, the analysis is based on a very fragile model; e.g., the model predicts almost 20,000 deaths for hurricane Sandy, which actually caused 159 fatalities. Their figure 1 and the discussion on p. 5, first paragraph, are not

Now, we explain our claim that the results are presented in a biased way. By holding the minimum pressure at its mean in prediction of counts of deaths, the authors only report the influence of MFI and normalized damage (figure 1 in ref. 1). This ignores the influence of the second interaction term MFI minimum pressure, which shows an opposite influence (see the estimated parameters on p. 5, first paragraph). By considering the counts of deaths under constant normalized damage, the results are contrary: male-named hurricanes with a low minimum pressure (strong hurricanes) are associated with more deaths than female ones (Fig. 1).

In the light of an alternating male-female naming system started in 1979, the authors claim that similar results can be obtained

differences between male- or female-named hurricanes for deaths, minimum pressure, category, and damages.

To conclude, the analyses given in ref. 1 are examples of the fact that prediction models using interaction terms have to be handled and interpreted carefully; in particular, using insignificant variables is not expedient and may lead to statistical artifacts.

To summarize, the data do not contain evidence that feminine-named hurricanes cause more deaths than masculine-named hurricanes.

Björn Christensen[a] ***and Sören Christensen***[b,1]

[a]*Institute for Statistics and Operations*

Population matters when modeling hurricane fatalities

Jung et al. (1) find gender bias leads to misperception of hurricane risk in both experimental and historical evidence. We affirm the rich literature on gender bias. However, we argue that Jung et al.'s empirical analysis suffers from endogeneity. Once addressed, we find the previous results to be of questionable robustness. Although gender bias may exist in limited-information experiments, historical evidence does not indicate gender bias in hurricane fatalities.

Damages do not determine deaths, but rather both are simultaneously determined by multiple factors, including hurricane characteristics and the (omitted) underlying population and vulnerability (2), which lead to endogeneity, or correlation between damages and the error term that can bias estimated

point of landfall. (No annual county-level data exist back to 1950. We calculated the 2000 county-to-country population density ratio and, assuming it is constant across time, we scaled the ratio by annual US population density. We recommend future work refine this assumption.) Third, we normalized deaths by (*i*) dividing deaths by real damages (nominal damages taken from the International Catastrophe Insurance Managers; deflator information taken from the Bureau of Economic Analysis; population taken from the US Census) and (*ii*) dividing deaths by total US population in the year of landfall (we normalized deaths by the five-county population and found similar results). After a log-transformation,

The experiments in Jung et al.'s study (1) are interesting but the motivational facts are of questionable robustness. We establish this finding by controlling for population and correcting for endogeneity. Further research on the subject of hurricane naming is therefore warranted and encouraged.

***Laura A. Bakkensen*[a,1] *and William Larson*[b]**

[a]*School of Government and Public Policy, University of Arizona, Tucson, AZ 85721; and* [b]*Office of the Chief Statistician, Bureau of Economic Analysis, US Department of Commerce, Washington, DC 20230*

LETTER

Statistics show no evidence of gender bias in the public's hurricane preparedness

Jung et al. (1) make the bold claim that a storm's assigned gender (traditionally masculine vs. feminine name) predicts its destructive potential, such that a hypothetical Hurricane Eloise would have three times the expected death toll compared with a hypothetical Hurricane Charley, by 41 deaths to 15 deaths. They say that feminine names and pronouns are perceived by the public to be less threatening, resulting in lax preparations and fewer evacuations. Their conclusion was widely reported in the popular press.

During 30+ years on the Gulf Coast, my hurricane evacuations show no gender bias (Andrew, Lili, Rita, and Gustav). I am skeptical that sexism explains the noted effect.

Ninety-four hurricanes made landfall in

deaths raises the average death toll for all female storms by 85%, to 23.5 ($n = 63$ deaths; maximum, 256 deaths). Hurricanes Diane (1955), Camille (1969), Agnes (1972), and Sandy (2012) collectively accounted for 732, or 38.5%, of 1,900 total deaths in the study. Virtually all of the statistical difference between the deadliness of female vs. male storms is explained by the inclusion of these four events.

Surprisingly, of those 732 deaths, at least 327 (45%) occurred well inland, mostly due to flooding and landslides in the mountain valleys of the Appalachians as the largely spent storms dropped torrential rains. Diane's deluge hit a region already rain soaked from Hurricane Connie a mere 5 d before. An es-

Setting aside the issue of outliers, what are the odds that the six deadliest hurricanes of the study period would all bear feminine names?

If all of the names had been assigned randomly or alternately, the correct answer would be 1 in 2^6, or about 1.6%.

However, four of the six storms occurred during a period (1953–1978) when only female names were assigned. Only Katrina and Sandy had a chance to be "male." The chance that two randomly chosen hurricanes would both have female names is 1 in 2^2, or 25%: that's not bias, it's a coincidence, and not a strong one at that.

***Steve Maley*[1]**

LETTER

Female hurricanes are not deadlier than male hurricanes

Jung et al. (1) assert that hurricanes that made landfall in the United States killed more people when they had female names rather than male names. The article has stirred much controversy. Criticisms range from the inclusion of hurricanes from the era before they were given male names (2), over the selective interpretation and the overstatement of their results from the archival study in favor of their hypothesis (3), to the external validity of their six behavioral experiments for at-risk populations in at-risk situations (4).

The criticism of this letter is a different one: the results of their archival study are a function of the selective inclusion of regressions. Hurricanes with female names had smaller death tolls when the hurricanes were strong (lower pressure), but higher death tolls when the hurricanes were weak (higher pressure). The latter result is driven by the pre-1978 sample (model 5). In the post-1978 sample, the interaction effect becomes insignificant and the damage toll has a positive and significant relationship with the death toll (models 6 and 7).

Like the death toll, the damage toll is a simultaneous outcome of the storm and hence not a good explanatory variable. It merely reflects other underlying characteristics that could range from the size of the hurricane or its area of effect over the assets at risk and the safety infrastructures and more reflective of other characteristics for weaker storms and after 1978. Even though a lower importance of safety infrastructures during weaker storms and an overall improvement in or a convergence of safety infrastructures after 1978 might explain these results, the ambiguity of the death toll variable disallows any strong or definitive interpretation.

Daniel Malter[1]

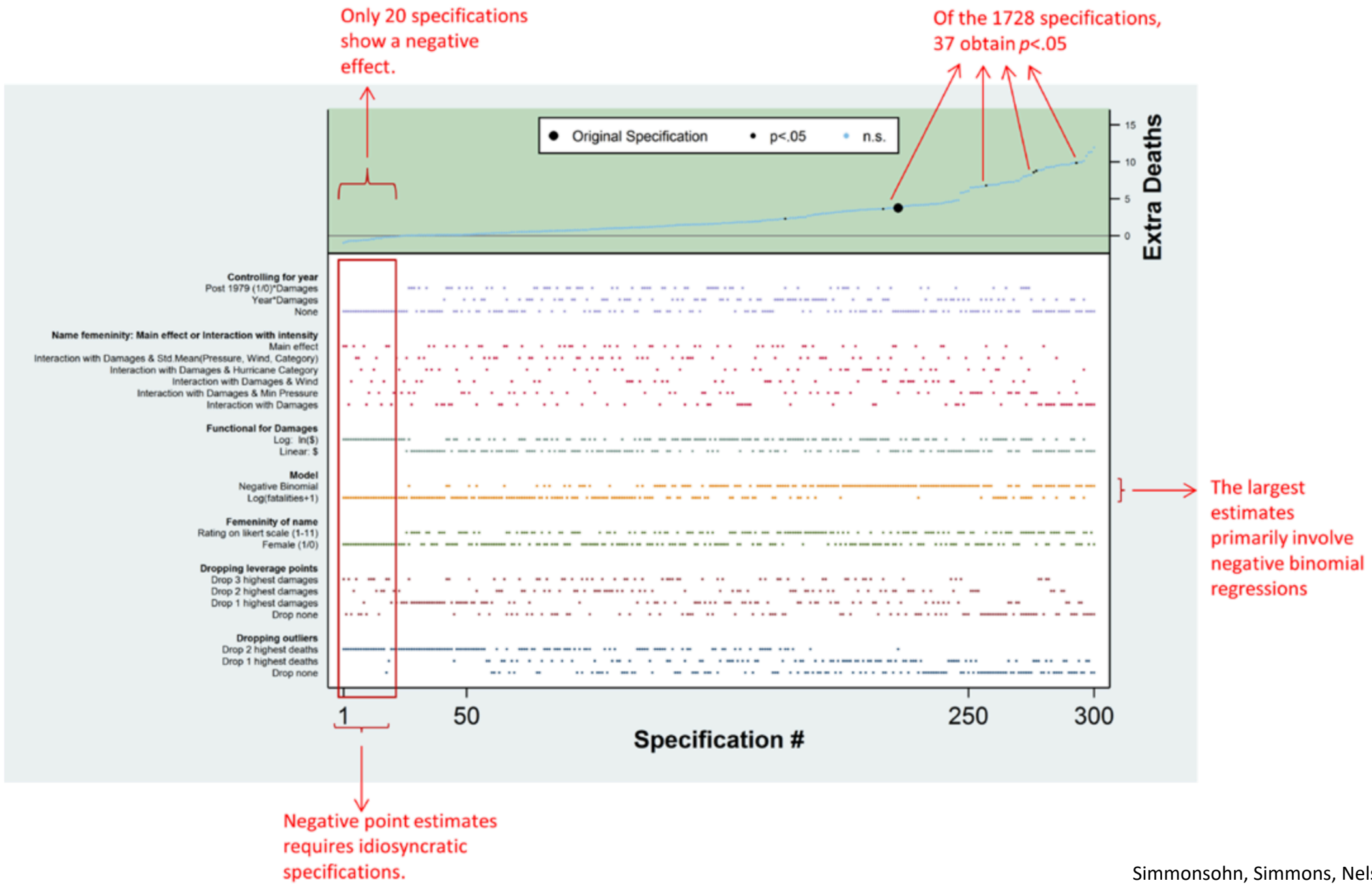
Strategy Unit, Harvard Business School, Harvard University, Boston, MA 02163

Specification Curve Analysis

Table 1. Original and alternative reasonable specifications used to test whether hurricanes with more feminine names were associated with more deaths.

Decision	Original Specifications	Alternative Specifications
1.Which storms to analyze	Excluded two outliers with the most deaths	Dropping fewer outliers (zero or one); dropping storms with extreme values on a predictor variable (e.g., hurricanes causing extreme damages)
2.Operationalizing hurricane names' femininity	Ratings of femininity by coders (1-11 scale)	Categorizing hurricanes names as male or female
3.Which covariates to include	Property damages in dollars interacted with femininity; minimum hurricane pressure interacted with femininity	Log of dollar damages; year; year interacted with damages
4.Type of regression model	Negative binomial regression	OLS with log(deaths+1) as the dependent variable
5.Functional form for femininity	Assessed whether the interaction of femininity with damages was greater than zero	Main effect of femininity; interacting femininity with other hurricane characteristics (e.g., wind or category) instead of damages

Simmonsohn, Simmons, Nelson (2015)



Simmonsohn, Simmons, Nelson (2015)

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Increasing Transparency Through a Multiverse Analysis

Perspectives on Psychological Science
2016, Vol. 11(5) 702–712

DOI: 10.1177/1745691616658637
pps.sagepub.com
SAGE

Sara Steegen[1], Francis Tuerlinckx[1], Andrew Gelman[2], and Wolf Vanpaemel[1]
[1]KU Leuven, University of Leuven and [2]Columbia University

Abstract
Empirical research inevitably includes constructing a data set by processing raw data into a form ready for statistical analysis. Data processing often involves choices among several reasonable options for excluding, transforming, and coding data. We suggest that instead of performing only one analysis, researchers could perform a multiverse analysis, which involves performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios. Using an example focusing on the effect of fertility on religiosity and political attitudes, we show that analyzing a single data set can be misleading and propose a multiverse analysis as an alternative practice. A multiverse analysis offers an idea of how much the conclusions change because of arbitrary choices in data construction and gives pointers as to which choices are most consequential in the fragility of the result.

PLOS | ONE

RESEARCH ARTICLE

I Just Ran a Thousand Analyses: Benefits of Multiple Testing in Understanding Equivocal Evidence on Gene-Environment Interactions

Vera E. Heininga[1]*, Albertine J. Oldehinkel[1], René Veenstra[2], Esther Nederhof[1]

1 University of Groningen, University Medical Center Groningen, Department of Psychiatry, Interdisciplinary Center Psychopathology and Emotion regulation, Groningen, the Netherlands, **2** University of Groningen, Department of Sociology, Groningen, the Netherlands

* v.e.heininga@umcg.nl

Citation: Heininga VE, Oldehinkel AJ, Veenstra R, Nederhof E (2015) I Just Ran a Thousand Analyses:

Abstract

Background

In psychiatric genetics research, the volume of ambivalent findings on gene-environment in-

frontiers in
NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 October 2012
doi: 10.3389/fnins.2012.00149

On the plurality of (methodological) worlds: estimating the analytic flexibility of fMRI experiments

Joshua Carp*

Department of Psychology, University of Michigan, Ann Arbor, MI, USA

Edited by:
Satrajit S. Ghosh, Massachusetts Institute of Technology, USA

Reviewed by:
Jonathan E. Peelle, Washington University, USA
Eugene Duff, University of Oxford, UK

****Correspondence:***
Joshua Carp, Department of Psychology, University of Michigan, 530 Church Street, Ann Arbor, MI 48109, USA.
e-mail: jmcarp@umich.edu

How likely are published findings in the functional neuroimaging literature to be false? According to a recent mathematical model, the potential for false positives increases with the flexibility of analysis methods. Functional MRI (fMRI) experiments can be analyzed using a large number of commonly used tools, with little consensus on how, when, or whether to apply each one. This situation may lead to substantial variability in analysis outcomes. Thus, the present study sought to estimate the flexibility of neuroimaging analysis by submitting a single event-related fMRI experiment to a large number of unique analysis procedures. Ten analysis steps for which multiple strategies appear in the literature were identified, and two to four strategies were enumerated for each step. Considering all possible combinations of these strategies yielded 6,912 unique analysis pipelines. Activation maps from each pipeline were corrected for multiple comparisons using five thresholding approaches, yielding 34,560 significance maps. While some outcomes were relatively consistent across pipelines, others showed substantial methods-related variability in activation

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY;

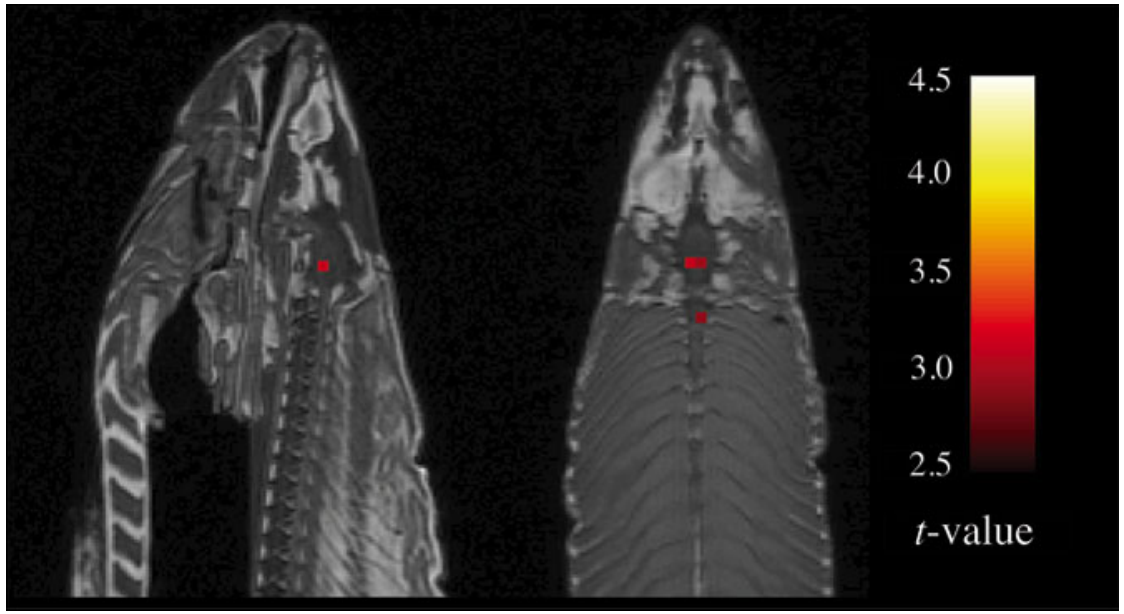
[3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.



Processing step	Reason	Options [suboptions]	Number of plausible options
Motion correction	Correct for head motion during scanning	• 'Interpolation' [linear or sinc] • 'Reference volume' [single or mean]	4
Slice timing correction	Correct for differences in acquisition timing of different slices	'No', 'before motion correction' or 'after motion correction'	3
Field map correction	Correct for distortion owing to magnetic susceptibility	'Yes' or 'no'	2
Spatial smoothing	Increase SNR for larger activations and ensure assumptions of GRF theory	'FWHM' [4 mm, 6 mm or 8 mm]	3
Spatial normalization	Warps an individual brain to match a group template	'Method' [linear or nonlinear]	2
High-pass filter	Remove low-frequency nuisance signals from data	'Frequency cut-off' [100 s or 120 s]	2
Head motion regressors	Remove remaining signals owing to head motion via statistical model	'Yes' or 'no' [if yes: 6/12/24 parameters or single time point 'scrubbing' regressors]	5
Haemodynamic response	Account for delayed nature of haemodynamic response to neuronal activity	• 'Basis function' ['single-gamma' or 'double-gamma'] • 'Derivatives' ['none', 'shift' or 'dispersion']	6
Temporal autocorrelation model	Model for the temporal autocorrelation inherent in fMRI signals	'Yes' or 'no'	2
Multiple-comparison correction	Correct for large number of comparisons across the brain	'Voxel-based GRF', 'cluster-based GRF', 'FDR' or 'non-parametric'	4
Total possible workflows			69,120

FDR, false discovery rate; FWHM, full width at half maximum; GRF, Gaussian random field; SNR, signal-to-noise ratio.

Poldrack et al. (2017)

MCS

1
Identify Specifications

Decide on all possible analytical pathways

Well-being
Any possible combination of 24 questions about well-being, self-esteem and feelings (cohort members) or of 25 questions of strengths and difficulties questionnaire (caregivers)

Technology Use
Mean of any possible combination of 5 questions concerning TV use, electronic games, social media use, owning a computer and using internet at home

Covariates
Included or not
(mother's ethnicity, education, employment, psychological distress, equivalised household income, whether biological father is present, number of siblings in household, conflict in mother-child relationship, frequency of mother-child interaction, long-term illness, negative attitudes towards school, mother's word activity score)

Total 3,221,225,472 specifications

2
Implementing Specifications

Run all possible analyses and graph outcomes

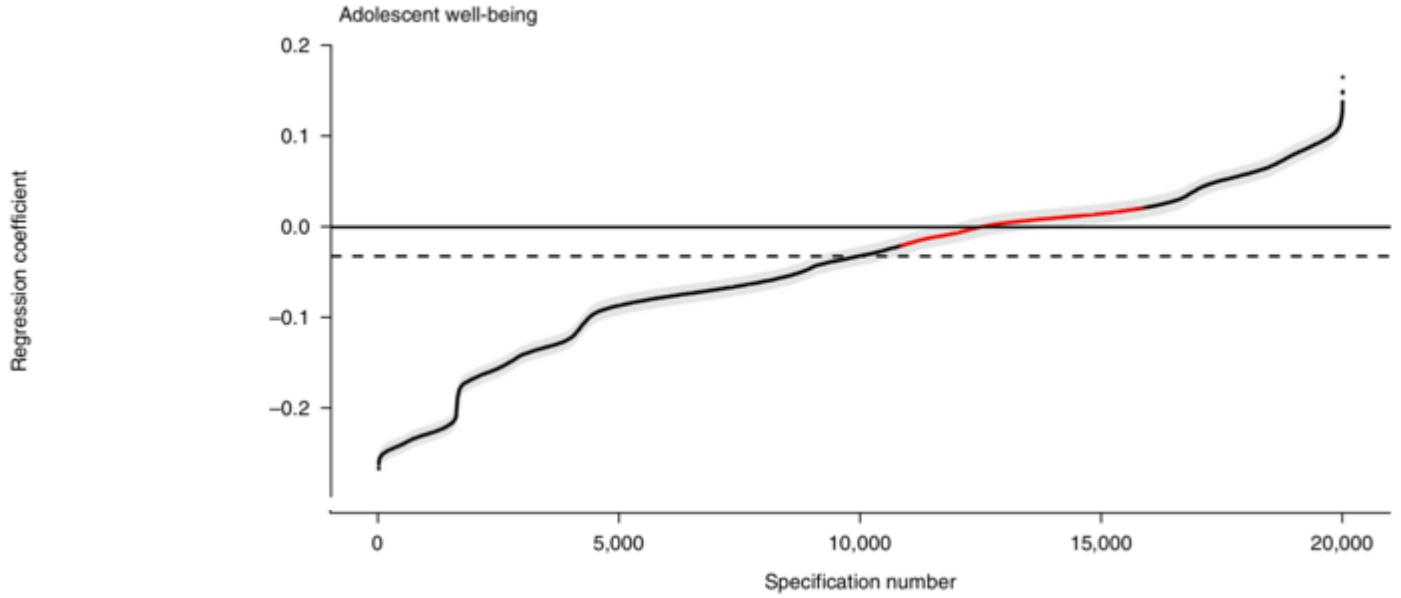


Fig. 3 | Results of SCA for MCS. Specification curve analysis showing the range of possible results for a simple cross-sectional regression of digital technology use on adolescent well-being. Each point on the *x* axis represents a different combination of analytical decisions, which are displayed in the 'dashboard' at the bottom of the graph. The resulting standardized regression coefficient is shown at the top of the graph; the error bars visualize the standard error. Red represents non-significant outcomes while black represents significant outcomes. To ease interpretation, the dotted line indicates the median standardized regression coefficient found in the SCA: $\beta = -0.032$ (partial $\eta^2 = 0.004$, median $n = 7{,}968$, median standard error = 0.010).

Orben and Przybylski (Nature Human Behaviour, 2019)

2
Implementing Specifications

Run all possible analyses and graph outcomes

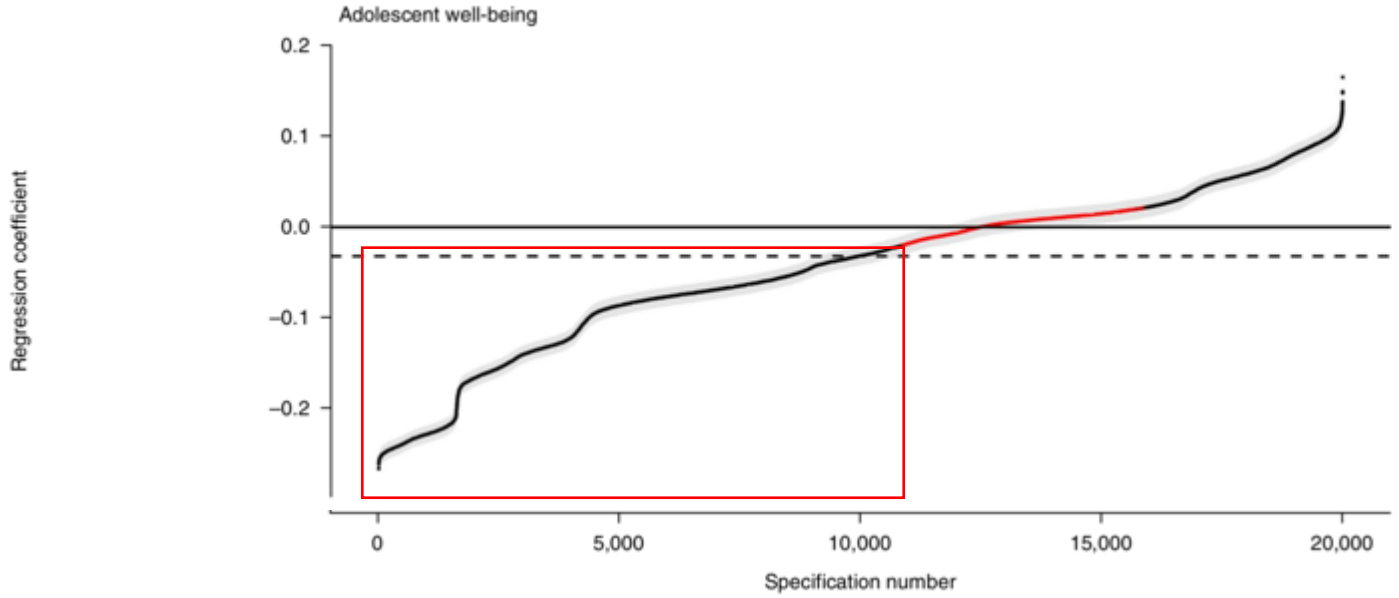


Fig. 3 | Results of SCA for MCS. Specification curve analysis showing the range of possible results for a simple cross-sectional regression of digital technology use on adolescent well-being. Each point on the *x* axis represents a different combination of analytical decisions, which are displayed in the 'dashboard' at the bottom of the graph. The resulting standardized regression coefficient is shown at the top of the graph; the error bars visualize the standard error. Red represents non-significant outcomes while black represents significant outcomes. To ease interpretation, the dotted line indicates the median standardized regression coefficient found in the SCA: $\beta = -0.032$ (partial $\eta^2 = 0.004$, median $n = 7{,}968$, median standard error = 0.010).

Orben and Przybylski (Nature Human Behaviour, 2019)

2
Implementing Specifications

Run all possible analyses and graph outcomes

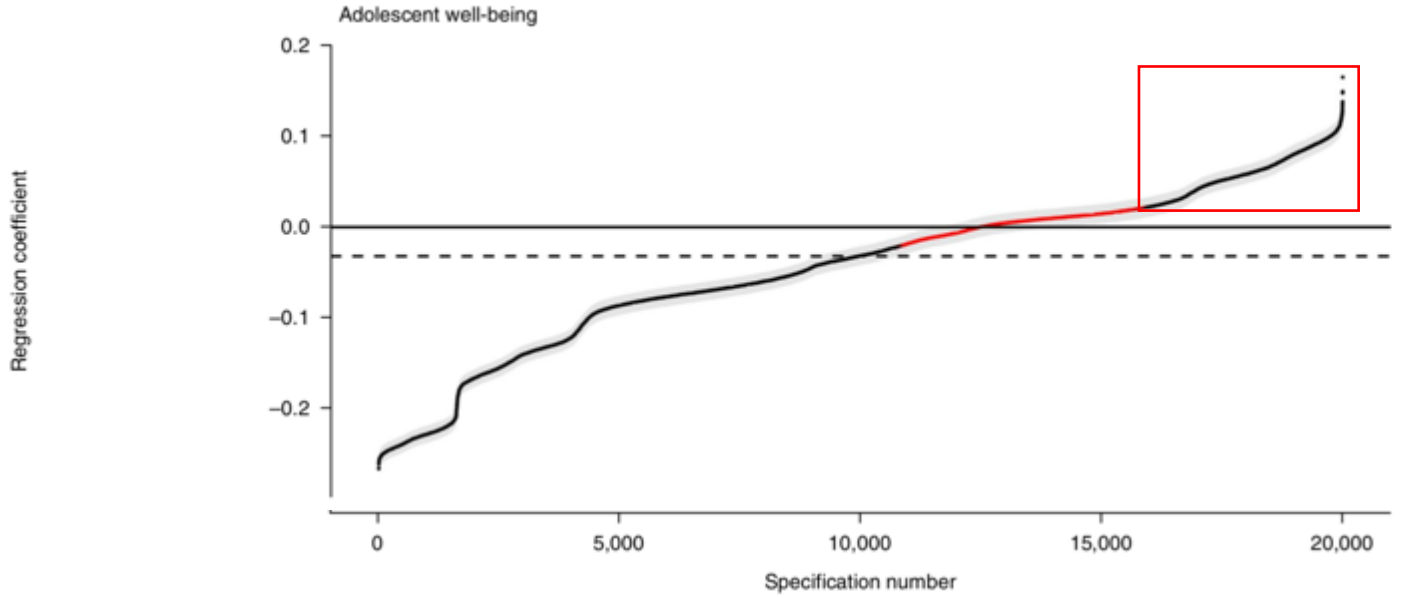


Fig. 3 | Results of SCA for MCS. Specification curve analysis showing the range of possible results for a simple cross-sectional regression of digital technology use on adolescent well-being. Each point on the *x* axis represents a different combination of analytical decisions, which are displayed in the 'dashboard' at the bottom of the graph. The resulting standardized regression coefficient is shown at the top of the graph; the error bars visualize the standard error. Red represents non-significant outcomes while black represents significant outcomes. To ease interpretation, the dotted line indicates the median standardized regression coefficient found in the SCA: $\beta = -0.032$ (partial $\eta^2 = 0.004$, median $n = 7{,}968$, median standard error = 0.010).

Orben and Przybylski (Nature Human Behaviour, 2019)

2
Implementing Specifications

Run all possible analyses and graph outcomes

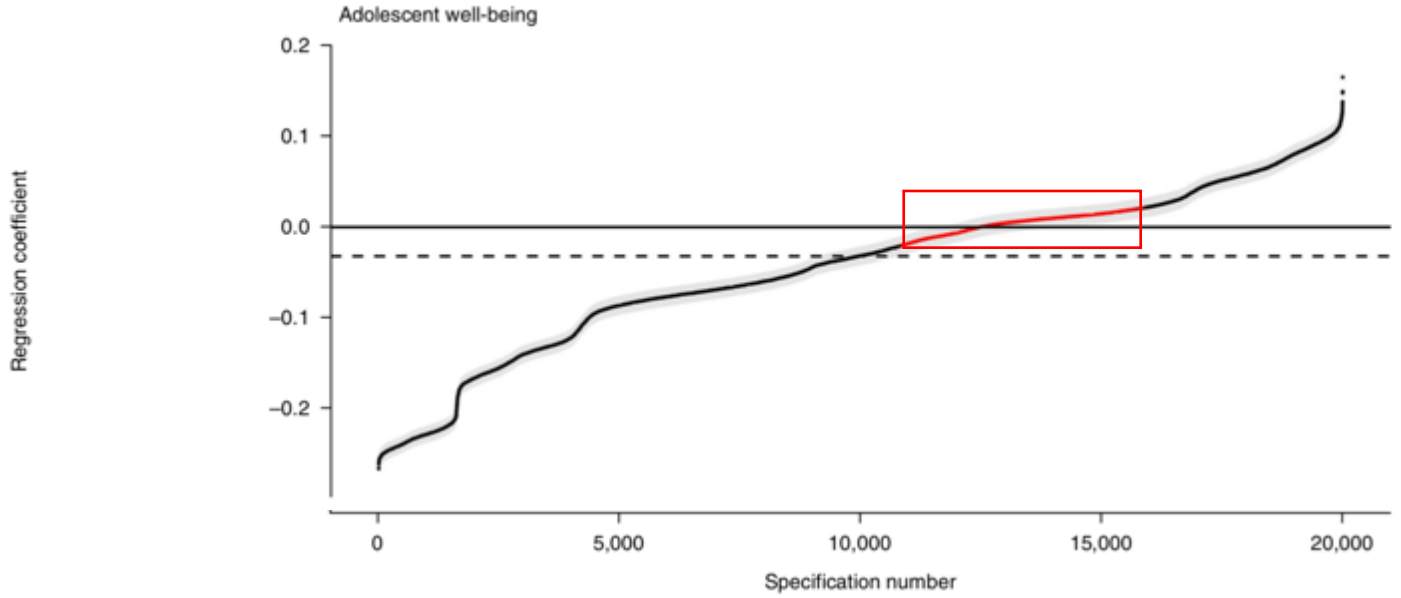


Fig. 3 | Results of SCA for MCS. Specification curve analysis showing the range of possible results for a simple cross-sectional regression of digital technology use on adolescent well-being. Each point on the *x* axis represents a different combination of analytical decisions, which are displayed in the 'dashboard' at the bottom of the graph. The resulting standardized regression coefficient is shown at the top of the graph; the error bars visualize the standard error. Red represents non-significant outcomes while black represents significant outcomes. To ease interpretation, the dotted line indicates the median standardized regression coefficient found in the SCA: $\beta = -0.032$ (partial $\eta^2 = 0.004$, median $n = 7{,}968$, median standard error = 0.010).

Orben and Przybylski (Nature Human Behaviour, 2019)

2
Implementing Specifications

Run all possible analyses and graph outcomes

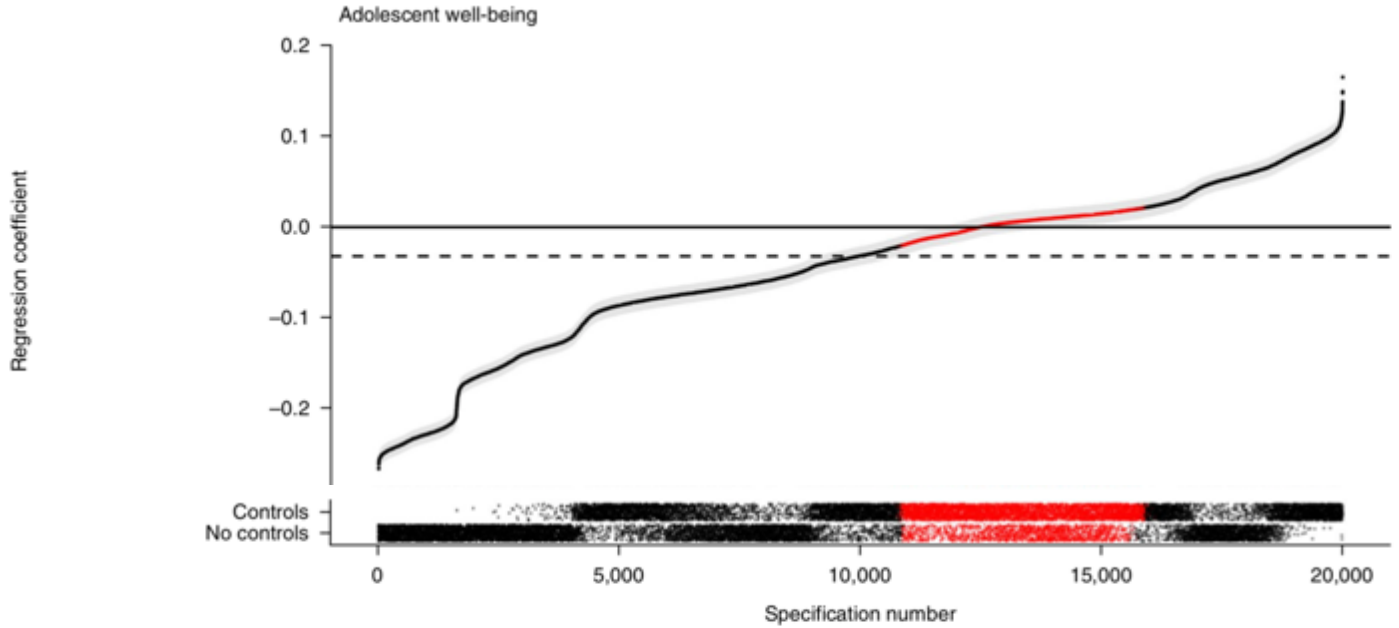


Fig. 3 | Results of SCA for MCS. Specification curve analysis showing the range of possible results for a simple cross-sectional regression of digital technology use on adolescent well-being. Each point on the *x* axis represents a different combination of analytical decisions, which are displayed in the 'dashboard' at the bottom of the graph. The resulting standardized regression coefficient is shown at the top of the graph; the error bars visualize the standard error. Red represents non-significant outcomes while black represents significant outcomes. To ease interpretation, the dotted line indicates the median standardized regression coefficient found in the SCA: $\beta = -0.032$ (partial $\eta^2 = 0.004$, median $n = 7{,}968$, median standard error = 0.010).

Orben and Przybylski (Nature Human Behaviour, 2019)

2
Implementing Specifications

Run all possible analyses and graph outcomes

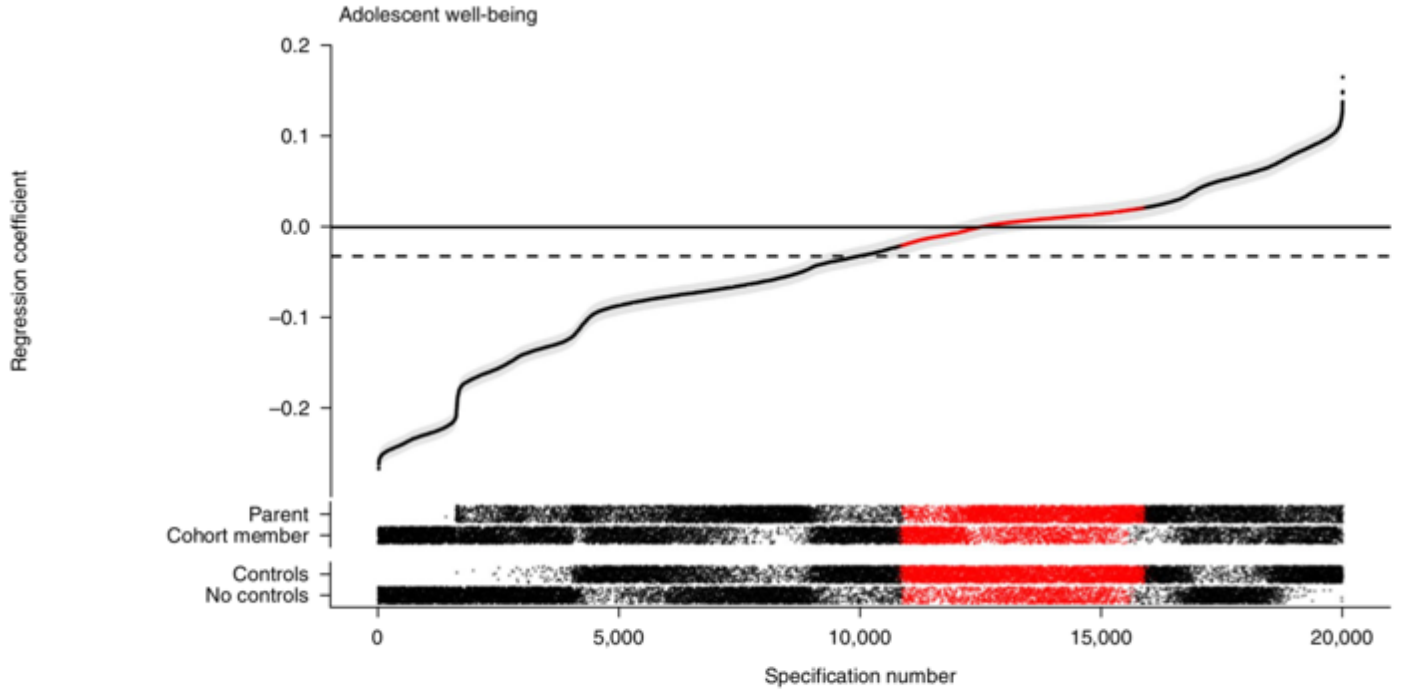


Fig. 3 | Results of SCA for MCS. Specification curve analysis showing the range of possible results for a simple cross-sectional regression of digital technology use on adolescent well-being. Each point on the *x* axis represents a different combination of analytical decisions, which are displayed in the 'dashboard' at the bottom of the graph. The resulting standardized regression coefficient is shown at the top of the graph; the error bars visualize the standard error. Red represents non-significant outcomes while black represents significant outcomes. To ease interpretation, the dotted line indicates the median standardized regression coefficient found in the SCA: $\beta = -0.032$ (partial $\eta^2 = 0.004$, median $n = 7{,}968$, median standard error = 0.010).

Orben and Przybylski (Nature Human Behaviour, 2019)

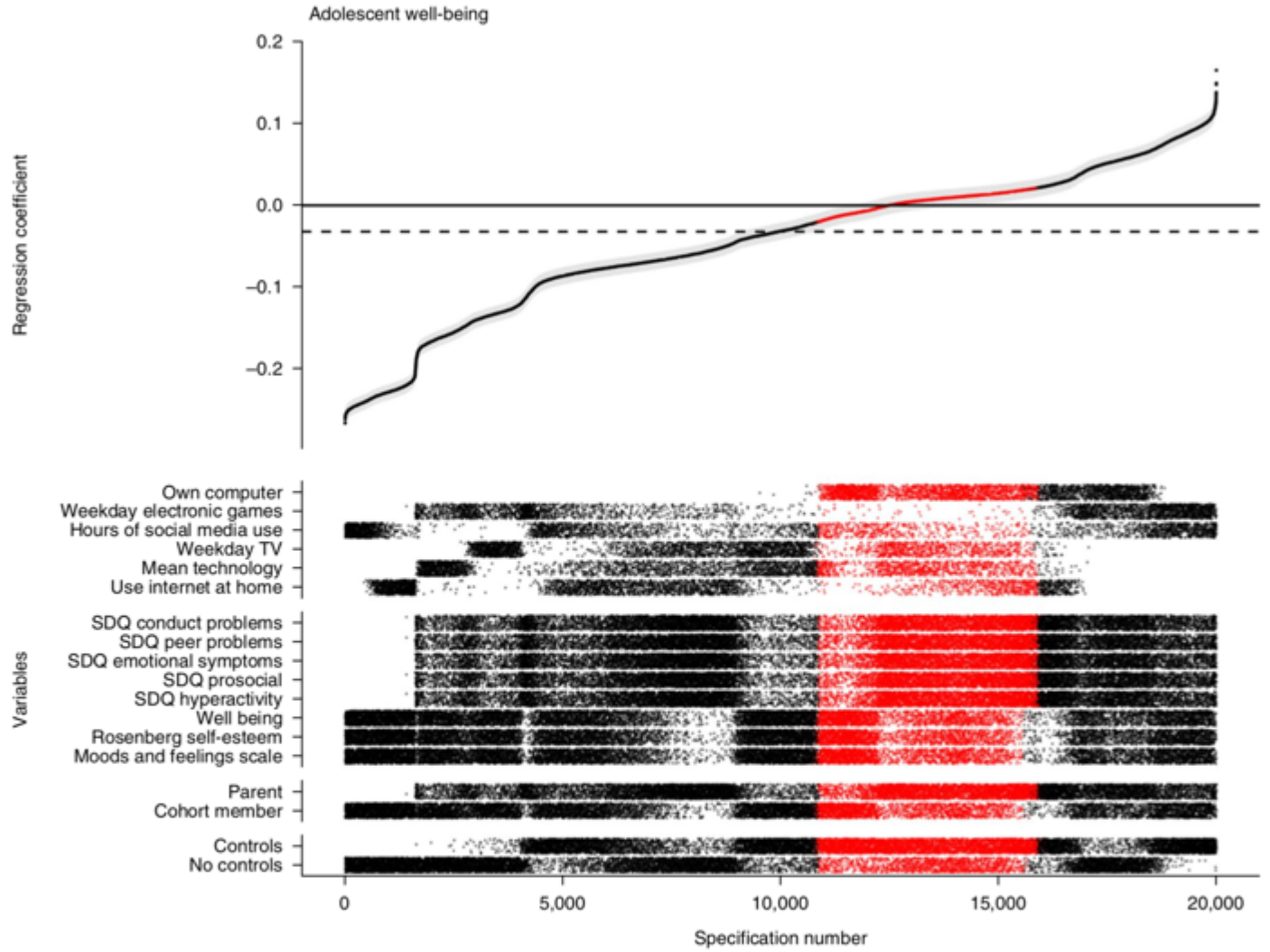


Fig. 3 | Results of SCA for MCS. Specification curve analysis showing the range of possible results for a simple cross-sectional regression of digital technology use on adolescent well-being. Each point on the *x* axis represents a different combination of analytical decisions, which are displayed in the 'dashboard' at the bottom of the graph. The resulting standardized regression coefficient is shown at the top of the graph; the error bars visualize the standard error. Red represents non-significant outcomes while black represents significant outcomes. To ease interpretation, the dotted line indicates the median standardized regression coefficient found in the SCA: $\beta = -0.032$ (partial $\eta^2 = 0.004$, median $n = 7{,}968$, median standard error = 0.010).

Orben and Przybylski (Nature Human Behaviour, 2019)

2 Implementing Specifications

Run all possible analyses and graph outcomes

Other Examples

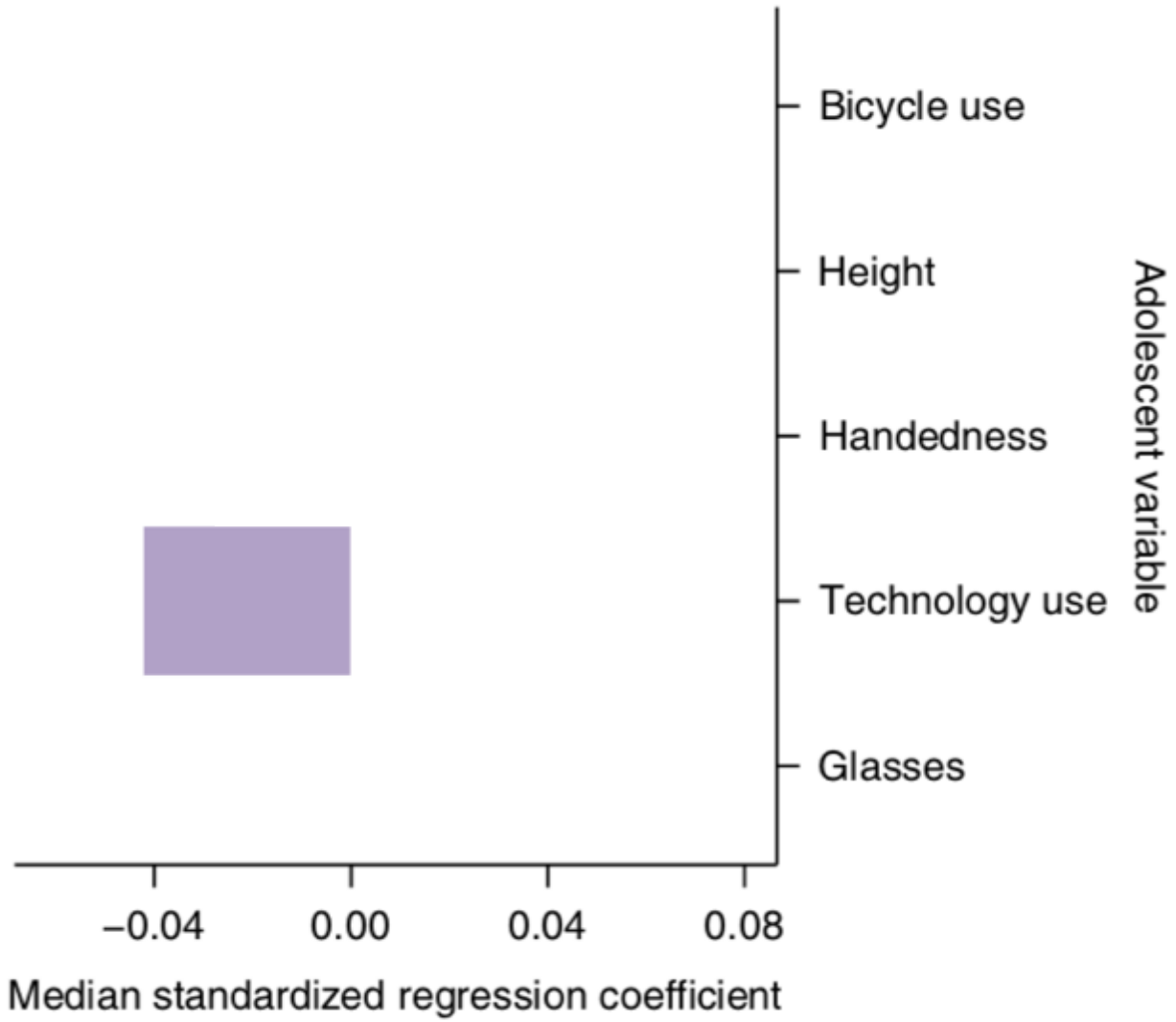
Preregistered with 3 datasets: Orben and Przybylski (Psychological Science, 2019)

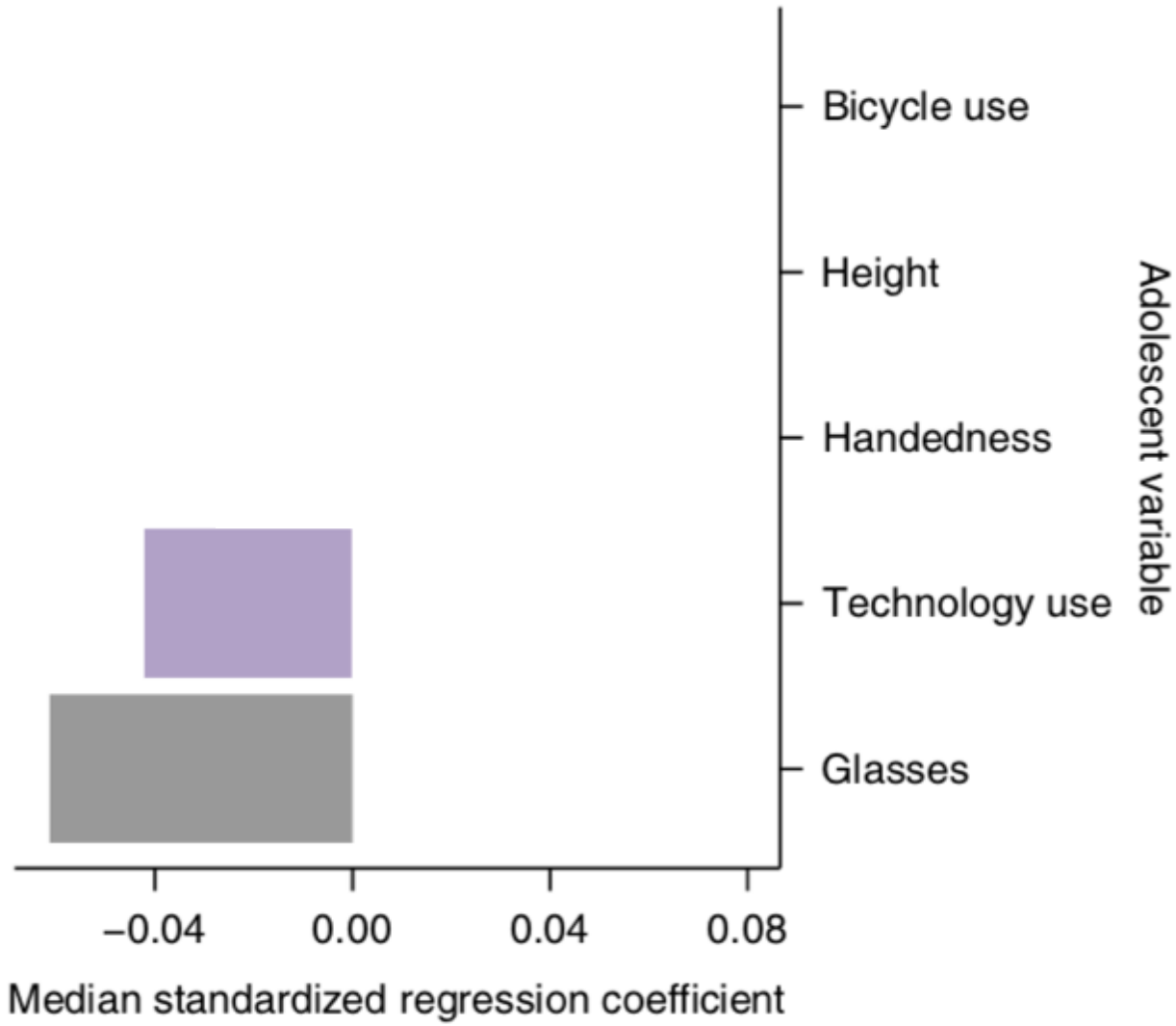
Longitudinal: Orben, Dienlin and Przybylski (PNAS, 2019)

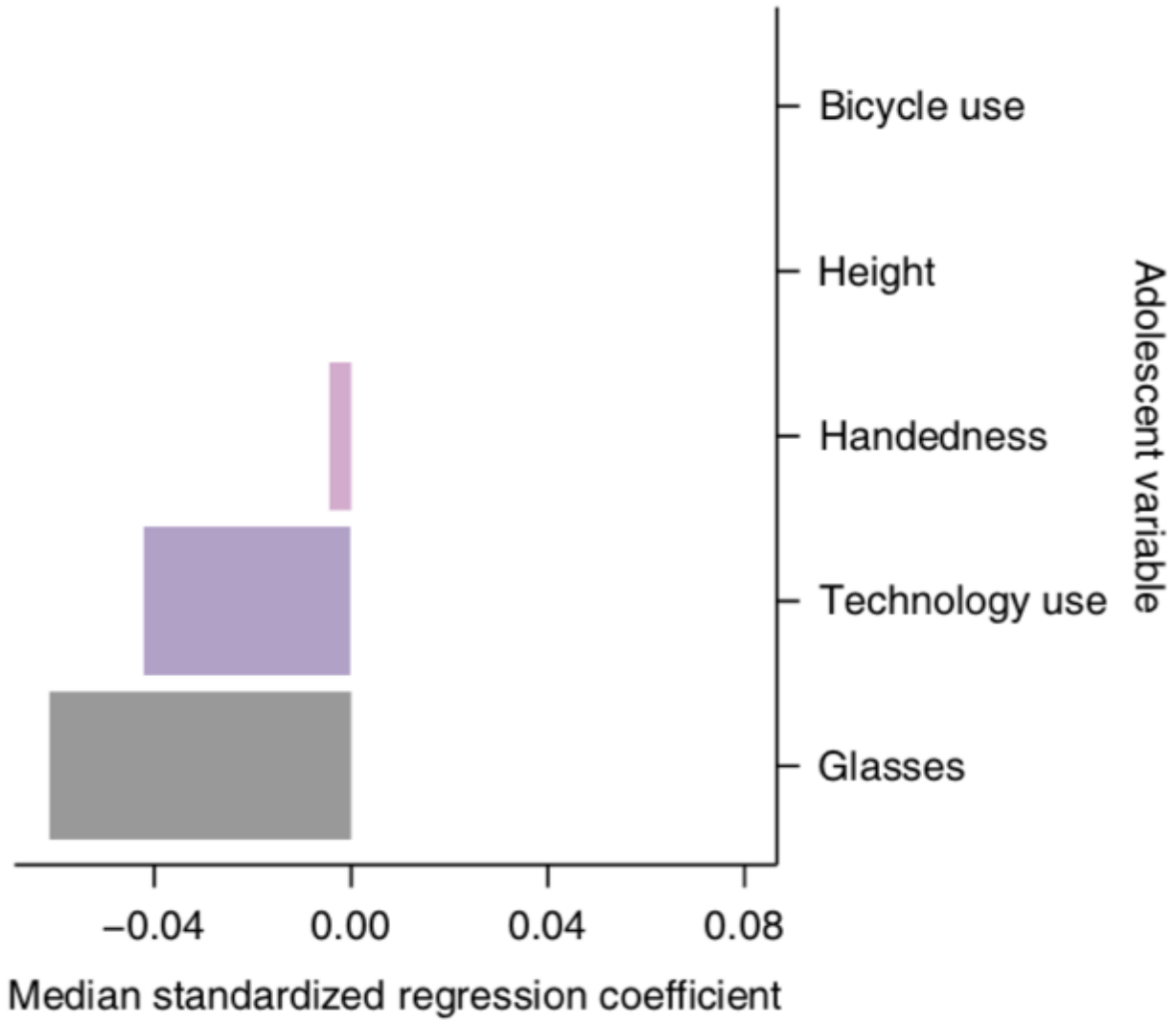
Solution #3

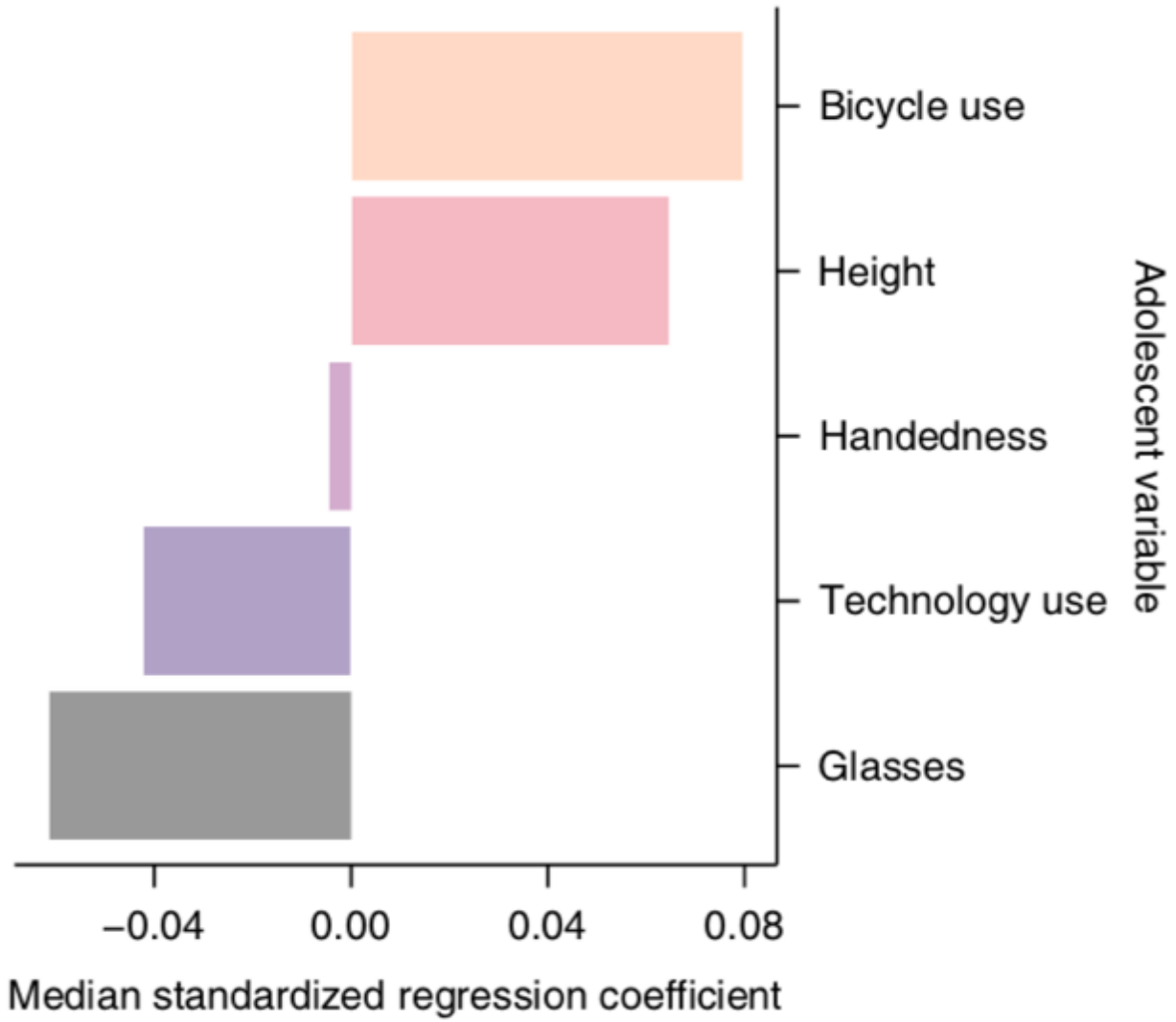
Include extra transparency about effect sizes

This can be putting effect sizes into perspective using other variables, Smallest Effect Sizes of Interest or real-life cut-offs


Bicycle use
Height
Handedness
Technology use
Glasses
Adolescent variable
−0.04
0.00
0.04
0.08
Median standardized regression coefficient


Bicycle use
Height
Handedness
Technology use
Glasses
Adolescent variable
−0.04
0.00
0.04
0.08
Median standardized regression coefficient


Bicycle use
Height
Handedness
Technology use
Glasses
Adolescent variable
−0.04
0.00
0.04
0.08
Median standardized regression coefficient


Bicycle use
Height
Handedness
Technology use
Glasses
Adolescent variable
−0.04
0.00
0.04
0.08
Median standardized regression coefficient

Equivalence Testing for Psychological Research: A Tutorial

Advances in Methods and
Practices in Psychological Science
2018, Vol. 1(2) 259–269

Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2515245918770963
www.psychologicalscience.org/AMPPS

Daniël Lakens, **Anne M. Scheel**, **and Peder M. Isager**
Human-Technology Interaction Group, Eindhoven University of Technology

Abstract

Psychologists must be able to test both for the presence of an effect and for the absence of an effect. In addition to testing against zero, researchers can use the two one-sided tests (TOST) procedure to test for *equivalence* and reject the presence of a smallest effect size of interest (SESOI). The TOST procedure can be used to determine if an observed effect is surprisingly small, given that a true effect at least as extreme as the SESOI exists. We explain a range of approaches to determine the SESOI in psychological science and provide detailed examples of how equivalence tests should be performed and reported. Equivalence tests are an important extension of the statistical tools psychologists currently use and enable researchers to falsify predictions about the presence, and declare the absence, of meaningful effects.

Or: https://psyarxiv.com/syp5a/

Psychological Medicine (2015), **45**, 3269–3279. © Cambridge University Press 2015
doi:10.1017/S0033291715001270

ORIGINAL ARTICLE

Minimal clinically important difference on the Beck Depression Inventory – II according to the patient's perspective

K. S. Button[1]*, D. Kounali[1], L. Thomas[1], N. J. Wiles[1], T. J. Peters[2], N. J. Welton[1], A. E. Ades[1] and G. Lewis[3]

[1] *School of Social and Community Medicine, University of Bristol, Bristol, UK*
[2] *School of Clinical Sciences, University of Bristol, Bristol, UK*
[3] *Division of Psychiatry, University College London, London, UK*

Background. The Beck Depression Inventory, 2nd edition (BDI-II) is widely used in research on depression. However, the minimal clinically important difference (MCID) is unknown. MCID can be estimated in several ways. Here we take a patient-centred approach, anchoring the change on the BDI-II to the patient's global report of improvement.

Method. We used data collected ($n = 1039$) from three randomized controlled trials for the management of depression. Improvement on a 'global rating of change' question was compared with changes in BDI-II scores using general linear modelling to explore baseline dependency, assessing whether MCID is best measured in absolute terms (i.e. difference) or as percent reduction in scores from baseline (i.e. ratio), and receiver operator characteristics (ROC) to estimate MCID according to the optimal threshold above which individuals report feeling 'better'.

Results. Improvement in BDI-II scores associated with reporting feeling 'better' depended on initial depression severity, and statistical modelling indicated that MCID is best measured on a ratio scale as a percentage reduction of score. We estimated a MCID of a 17.5% reduction in scores from baseline from ROC analyses. The corresponding estimate for individuals with longer duration depression who had not responded to antidepressants was higher at 32%.

Good analysis of large-scale data is inherently rooted in **transparency**

Some of the tools to help are:

1. *Preregistration + Registered Reports*
2. *Specification Curve Analysis*
3. *Considering Effect Sizes*

Thank you

Professor Andrew Przybylski

Professor Robin Dunbar

Professor Dorothy Bishop

Conducting rigorous research on large open-access developmental datasets

Amy Orben

Department of Experimental Psychology, University of Oxford

ABCD Workshop, Portland

@OrbenAmy