Experimental research: problems and opportunities in the big-data era

Cremers, H.

DOI
10.1075/z.210.02cre
Publication date
2017
Document Version
Author accepted manuscript
Published in
Crossroads Semantics

Experimental research: problems and opportunities in the big-data era

Henk Cremers

Abstract: Experimental research in psychology, psycholinguistics or medicine provides quantitative and therefore seemingly conclusive and trustworthy evidence. However, it has been convincingly shown that most research findings are actually false. This has hardly influenced the dominant scientific evaluation system which reflects a continued trust in the unbiasedness of data by a strong reliance on simple quantifications of scientific quality and productivity, such as number of publications and number of citations. This state of affairs is remarkable in the light of a long history of strong criticism of commonly used inference methods and scientific evaluation systems, which is now backed by large-scale research projects directly questioning the reproducibility of scientific findings. This way, the large amounts of data – "big-data" – have helped to uncover some of these problematic issues, but also provided a more open attitude towards data and code sharing. In addition, novel analytic frameworks may help to better integrate empirical data with computational models.

1. Introduction

The former British Prime Minister Benjamin Disraeli is sometimes credited for the phrase that "there are lies, dammed lies, and … statistics". Regardless of the origin of the saying, its relevance for modern science can hardly be overstated. Aside from questions whether data is theory-laden or not (Meehl 1993), data and statistics can be used and misused in a virtually infinite number of ways. They serve as the foundation for policy and policy change, and are instrumental in scientific inferences. For example: The Washington Post reports that researchers found an 8% increase in head injuries in cities that have adopted a bike-sharing program.[1] This is a remarkable finding, which may be used to argue against this bike-sharing program. Actually however, the head injuries dropped 14% in these cities, but the article compared this relative decline of head injuries to injuries overall (which was declining at an even faster rate) to infer the "increase" in injuries.[2] While this may seem a trivial example of misrepresentation of data, which may be confined to the media, similar examples are just as common in the empirical (but "softer") sciences. There have been cases of blatant scientific fraud (Wicherts & van Assen 2012), and reports of widespread error in statistical

[1] https://www.washingtonpost.com/news/to-your-health/wp/2014/06/12/cities-with-bike-share-programs-see-rise-in-cyclist-head-injuries/
[2] http://andrewgelman.com/2014/06/17/lie-statistics-example-23110/

analyses (Nuijten, Hartgerink, van Assen, Epskamp & Wicherts 2015). Yet neither one of these issues will be that problematic in the long run: fraudulent researchers are, it is safe to assume, a tiny minority, and random errors will cancel out. However, some damage is done in the grey area between fraud and random mistakes, and concerns -among others - weak statistical inferences and scientific evaluation and a strong believe in unambiguity of research data.

A painful example of such a case is the (social) priming effect shown by John Bargh (Bargh, Chen & Burrows 1996). In this famous study, Bargh and colleagues showed that participants primed with an elderly stereotype subsequently walked through the hallway more slowly than people in a control group. Such "strongly significant" effects were widely accepted as scientific evidence (Kahneman 2013). However, after decades of apparent replications and generalization of the findings, it was shown that these effects were not replicable and that they were actually due to a basic experimenter bias (Doyen, Klein, Pichon & Cleeremans 2012). This has urged Nobel prize laureate Daniel Kahneman, a long-time "believer" (Yong 2012) in this type of priming effects, to personally address the researchers involved, urging them to "clean up their mess", and he proposed an even longer chain of replication studies to establish the robustness of results.[3]

[3] http://www.nature.com/news/nobel-laureate-challenges-psychologists-to-clean-up-their-act-1.11535

Here I will discuss the appeal of such initial spectacular, but essentially non-replicable findings, and the culture of scientific evaluation and incentives in which they thrive. I will use psychology research as an example throughout, but it in fact more broadly concerns the social sciences, and certain "soft" medical and biological research (Fanelli 2010). To what extend it also relates to psycholinguistics is an open question. Perhaps the area is generally stronger theoretically embedded than most of the aforementioned research fields, although this doesn't necessary ameliorate problematic statistical issues. A key contributing factor to the unreliability of research findings is simply the use of small samples (Tversky & Kahneman 1971; Cohen 1994). Consequently, it may follow that in the current "big-data era" such problems are soon to be history, but it does not seem to be that easy (Lazer, Kennedy, King & Vespignani 2014). First I will start by discussing a few very well-known problems in the interpretations of research data, and then I will discuss the scientific evaluation system.

2. The questionable empirical toolbox

2.1 Bias

While most experimental researchers make a living gathering data, the biostatistician John Ioannidis has made a career out of questioning their

empirical status. Particularly remarkable examples of his work are the two papers *"Why most published research findings are false"* (Ioannidis 2005) and *"Why most discovered true associations are inflated"* (Ioannidis 2008), which re-initiated long-standing debates about the reliability of empirical science (e.g. Meehl 1978). By a modelling approach Ioannidis showed that under various estimates of biased research practices, it can be shown that more than 50% of research findings reported to be statistically significant are in fact false.[4] Interestingly, a large collaborative effort a decade later found empirically what Ioannidis predicted based on his model: most published findings in psychology do not replicate (Open Science Collaboration 2015).

The main concern addressed by Ioannidis is that the results of an analysis cannot be interpreted in a vacuum but need to be considered in their context, in particular the amount of *bias* (e.g. selective reporting) that is part of the results. This points at an essential issue in the interpretation of research findings: the validity and reliability never simply follow from the data itself, but always depend on the assumptions of the statistical framework. For example, it is well-known that when running multiple statistical tests, the chance of finding a "false positive" increases. Usually a threshold of $p<0.05$ is applied to consider a result statistically significant (i.e. a less than 5% chance that some observation is drawn from a null, or no effect, distribution). However, if one tests for example 20 hypotheses at this threshold, the overall

[4] that is, more likely to be drawn from a so called "null" distribution, see section 2.2

probability increases to 1-(1-0.05)^20 = 0.63, or about 63%. When all tests that are performed are actually reported, this so-called "multiple comparisons problem" allows for a fairly straightforward solution, a topic in any introductory statistics course. However, results can easily get biased when not all tests that are performed are reported. For example, several statistical tests are explored, the inclusion or exclusion of data points (for example subjects) is evaluated, et cetera, long before any correction for multiple comparisons is being applied. Considering this flexibility and several loosely defined statistical tests, it actually becomes easy to create a situation in which at least one finding is "significant" (e.g. Carp, 2012; Maxwell, 2004; Simmons, Nelson & Simonsohn 2011). In other words, given the undisclosed *flexibility* or researchers' "degrees of freedom" (Simmons et al. 2011) the phrase "statistical significance" becomes meaningless.

2.2 Null hypothesis testing

The above-mentioned issues are problems intrinsically related to the standard inference method that experimental psychology has adopted over the last 60 or so years. This standard framework became known as the *null hypothesis statistical testing* (NHST) (Nickerson 2000), and is a remarkable historical hybrid of two statistical philosophies: Fisher's null hypothesis testing and Neyman-Pearson's ideas on having "decision thresholds" (Gigerenzer 2004). Most of the problems with this NHST approach are fairly easily exposed from

a *Bayesian* perspective (Lindley 2000) yet have stayed remarkably persistent (Falk & Greenbaum 1995).

In NHST, the observed data are evaluated under the assumption (the null hypothesis and its associated null distribution) that there is no effect. The aforementioned infamous "p-values" scattered throughout many research papers refer exactly to this probability: the likelihood that a certain effect is observed under the assumption that there is actuality no (a null) effect. If this probability is small, then this so-called null hypothesis is rejected. For example, if one runs an experiment where priming of elderly-related words leads to a reduction in walking speed, one may conclude it is unlikely that such effect would be obtained if in actuality there is no effect of these prime words on walking speed. However, this is not what is often concluded. There is a tendency to interpret this as: it is unlikely that there is no effect of prime words given the data. In other words, it is not about the unlikeliness of the data under the assumption of no effect, but it is seen as a measure of the unlikeliness of the null hypothesis (and perhaps taken even one step further: the likelihood of an alternative hypothesis, i.e. priming, exists). This difference is not a play of words. The former correct interpretation refers to the conditional probability (that this the likelihood of some event given some other event) of observing some data (D) given a null hypothesis of no effect (H0): p(D|H0). The latter refers to another condition probability (the posterior), that of the existence of the null hypothesis given the data p(H0|D). That these two conditional probabilities are not identical follows from Bayes'

theorem and can be made intuitive by the following example. The probability that there are clouds, given the observation that it rains, approaches 1. However, the probability that it rains, given the observation that there are clouds is high, but certainly much lower. Simply put, you may see clouds without rain, but not rain without clouds. Using these probabilities interchangeably has consequences beyond empirical research or the weather. In the context of law enforcement for example, it is referred to as the *prosecutor fallacy* (Thompson & Schumann 1987): the difference between the likelihood of some evidence, given that someone is innocent, versus the probability that someone is innocent given the evidence. Mixing-up these probabilities has resulted in unwarranted years in jail (Hill 2004). While the difference between the probabilities potentially has far-stretching consequences for the interpretation, it is remarkable that many researches are not aware of the correct interpretation of p-values (Rosenthal & Gaito 1963), especially given the frequency with which you come across "$p<0.05$" in an empirical research article.

2.3 Theory testing

The weak empirical status of null-hypothesis testing in psychology stands in sharp contrast to other scientific disciplines like physics (Meehl 1967). The "softer" empirical sciences, unlike the physical sciences, generally do not make point estimates, that is, exact predictions on the strength of an

association: the *effect size* (Meehl 1967). Hypotheses are often stated in such a way that they can neither be corroborated nor refuted (Meehl 1978), and theories lack a cumulative character; like old generals, they never die, they just slowly fade away (Meehl 1978). Particularly in these softer sciences, the gap between some substantive theory (T) and hypothesis about an observation (O) is large, and these need to be glued together with an often problematic auxiliary (A) hypothesis and experimental particulars (C); (T.A.C) → O. Consequently, by falsifying some statistical hypothesis, one does not directly falsify the theory, but rather the conjunction: ¬O implies ¬ (T.A.C) or ¬T ∨¬A ∨ ¬C, (not just ¬T), which is arguably uninformative (Meehl 1978). This is a classic topic in the philosophy of science, and certainly not a problem specific to the social sciences, but because the auxiliary assumptions are much more problematic, and usually flexible, the burden of proof is low (Meehl 1978). Indeed, a large-scale study underscored this assertion and showed that the social sciences have a higher number of "positive" findings compared to other sciences, indicative of a weaker form of hypothesis testing (Fanelli 2010). Statistically the difference between these "soft" and "hard" sciences could be described as the difference between data-fitting versus data-predictions (Shmueli 2010). Along that line, with the application of basic techniques (split data into test and training sets) borrowed from machine-learning (Breiman 2001), a data-prediction approach can relatively easily be achieved. Linguistics and machine learning share a long and successful

history, for instance with he application neural networks in modelling natural language processing and language acquisition (e.g. Collobert & Weston 2008), yet even modern clever machines can get lost in translation (see Google Translate, figure 1c). However with respect to experimental research, a data-driven, test/training methodology allows for predictions on the effect size, even when theory and auxiliary assumptions do not directly forecast one.

3. Scientific publications and evaluation

Results from experimental (or observational) research thus seem to convey much more than is warranted. The inferences are dependent on basic statistical assumptions (e.g. normality, multiple comparison) and a host of (questionable) research practices (e.g. flexibility in data analyses). What is even more remarkable is how this set of common practices has "survived" the decades of convincing criticism and viable alternative inference procedures that for instance (one way or another) simply address the uncertainty around the strength (effect size) of statistical results (Loftus 1993).[5] In this context, it is also interesting to consider the *confirmation bias* among scientists (Fugelsang, Stein, Green & Dunbar 2004), a tendency to interpret information that confirms one's hypotheses: How is it possible that despite the abundance

[5] For instance with the *bayes factor* or by providing *confidence intervals*.

of evidence against the repertoire of a set of common research practices, they remain so widely used and strongly believed? One reason may be that the idea that data give unambiguous "answers" is deeply embedded in the system of scientific publication and evaluation. The idea, for example, that scientific quality can easily be quantified (in terms of number of papers, number of citations etc), and the fact that novelty is more highly regarded than replications, and a review process has been adopted that blatantly ignores statistical laws (Tversky & Kahneman 1971) have led to a system where spectacular but often non-replicable results thrive and evolve. Clearly, a complex process like the evaluation of research findings or scientific quality does not lend itself to a simple optimal solution. However, beyond this process being just not optimal, it may even be essentially non self-correcting (Ioannidis 2012).

3.1 The economy of the publication and evaluation systems

The scientific publication system has been described according to a set of economics principles were articles are considered as commodities, that may help explain the persistence of certain research practices (Young, Ioannidis & Al-Ubaydli 2008). According to this work, scientific reports suffer from what is called the *winner's curse.* In auctions, and particularly in a situation where no one exactly knows how much an object is worth, the "winner of the bid" on average tends to overpay for whatever was for sale (Young et al. 2008). In

science one example of this phenomenon is that only "significant" findings will get published (the famous "file drawer problem"; Rosenthal 1979) and among the published findings, especially those in selective, high-impact journals, tend to contain results that are overstated (Ioannidis 2008). This occurs because from a pool of findings, the strongest, and hence most spectacular results also are the ones that on average tend to overestimate the underlying population effect most strongly. Often subsequent replication studies tend to show different results (but get less attention, see below) – outside the realm of economics referred to as the proteus phenomenon (Pfeiffer, Bertram & Ioannidis 2011). Remarkably, largely publicly funded research is freely available to commercial intermediates (journals) who sell the articles back to the author(s)! This system can be explained by the inherent *uncertainty* in science; it is impossible to predict beforehand the future value, extensions, and practical applications (Young et al. 2008). Journals provide authority independent of the content of a manuscript, and hence give the commodity "value". A relatively small amount of a high impact and therefore powerful journals (an oligopoly) with limited publication slots thus determine the visibility of research. Effectively these journals create *artificial scarcity,* low acceptance rates, numbers of papers and print page limits. These limitations stem from the paper publications age, but are entirely artificial in the digital age, and fuel the rise of open-access publications. Notable is the switch that the editorial team of the journal *Lingua* made by starting a similar new journal *Glossa*, in response to their publisher Elsevier's reluctance to

support open-access publication.[6]

Citations are of course essential and mandatory to the scientific process, simply to acknowledge other work. But the reliance on counting papers and the importance of citations as evidence of scientific quality has several drawbacks. The role of high-impact journals is peculiar even if the number of citations of a paper actually says anything about the quality or importance of a paper. It has been shown that citations of individual papers do not really correlate with the impact factor of a journal (Munafo 2013). In other words, the actual "impact" of a research paper (measured in number of citations) is hardly predictable by the journal rank – which is based on the citations of its manuscripts! Citations for an individual paper could still be useful as a measure of scientific quality regardless. However, a survey under highly-cited biomedical researchers showed a limited relation between what researchers themselves regarded as their most important work, and the number of times this work was cited (Ioannidis, Boyack, Small, Sorensen & Klavans 2014). Citations can also start to live a life of their own; scientific citation networks can create a blur with unfounded authority; for example through a citation bias against papers that weakened a belief, the amplification of results without new data, or even invention of results (Greenberg 2009). A particular form of citation, self-citation, further helps to create this unfounded authority. An example here is research on the so-called "Type-D" personality

[6] http://kaivonfintel.org/2015/11/02/lingua-glossa/

construct supposedly psychometrically distinct form other personality constructs and highly predictive of some medical conditions, a claim that could not be replicated (Coyne & de Voogd 2012).

3.2 Alternatives for the evaluation system

A host of solutions have been proposed that are promising in ameliorating the current scientific evaluation system (Kriegeskorte 2012). A key problem is the closed evaluation system based on anonymous reviews. Research has shown that the system is highly unreliable (Bornmann, Mutz & Daniel 2010). A critical change might therefore be an open, and more importantly, post-publication review system (Kriegeskorte 2012). That is, after the paper has appeared, researchers should be able to comment on the findings et cetera, which could refine a paper. This is markedly different from the current process were research findings are presented "as-is" after the anonymous review process. However, for such a novel system to work reviewers should also be rewarded for their contribution, by getting votes on the quality of the review (Kriegeskorte 2012). This would essentially be the type of highly-effective review system that Amazon uses for its products. Another area where a similar review system has proven invaluable is the stackoverflow.com library of questions and answers on computer code. Votes are given on the quality of answers ("reviews") and this way, users are rewarded for their contributions. Such applications at least suggest that the

current review system is not questioned for the skill to review material, but rather for its closed nature and the usually small sample (maybe 2 or 3 reviewers per paper or grant application). Moreover, in the slightly different setting of a so-called prediction-market, it turns out that a group of reviewers is actually good at predicting the replicability of research findings (Dreber et al. 2015).

4. More Data, More Problems?

We seemed to have unofficially entered the big-data era, although we don't really seem to know how much data (the 20 petabytes of data Google processes, per day?) one needs to be considered to be working on big data.[7] Experimental research data still seems far away from these numbers, for the simple reason that the acquisition of data remains time-consuming. Moreover, if the interpretation of research data is so difficult, one may wonder why more data would actually be helpful at all. The most obvious advantage of having more data is that any statistical estimate simply becomes more reliable (Wainer 2007). Secondly, larger data-sets allow for machine-learning approaches: predictions are being made and tested on "unseen" data, instead of just fitting a statistical model (Breiman 2001). As discussed, this can be considered a much stronger form of scientific inference. Data collection

[7] http://www.talyarkoni.org/blog/2014/05/19/big-data-n-a-kind-of-black-magic/

through internet and with smartphones (Miller 2012) has certainly helped to make the acquisition of experimental data much easier. Notable examples include a study on emotion contagion measured on Facebook, testing 689,003 subjects (Kramer, Guillory & Hancock 2014) and a project on subjective well-being using a smartphone-survey that included 18,420 participants (Rutledge, Skandali, Dayan & Dolan 2014). In addition, large collaborative efforts help to collect large amounts of data relatively fast. Two examples here are the Many Labs replication project (Klein, Ratliff, Vianello, Adams & Bahník 2014) and the open science consortium (Open Science Collaboration 2015). The first showed that high-powered (that is large sample sizes) studies could replicate 10 out of 13 "classic" and modern psychological phenomena. The second study came to a perhaps more pessimistic conclusion: the effect sizes of the replication studies were roughly half of what the original studies found, and among the replication studies less than 40% showed significant result, compared to 97% in the original publication pool (Open Science Collaboration 2015). Unsurprisingly, the interpretation of these findings led to debate (Gilbert, King, Pettigrew & Wilson 2016). Regardless of how positive or worrisome these findings may be regarded, they certainly are good examples of how an open and collaborative effort can lead to much less biased and more robust experimental data.

A second development is the expansion of sophisticated, and sometimes automatic, meta-analyses. These types of developments have led to both the uncovering and quantification of some problematic issues in

experimental research. For example by plotting a large amount of reported p-values, the phenomenon of "p-hacking" can be uncovered: there are many more p-values just under the significance threshold than would be expected, indicative of selective reporting (Simonsohn, Nelson & Simmons 2014). Large automatic meta-analytic techniques are also necessary to keep track of the tremendous amount of research papers and results, an example of this is the automated curation of facts in the biomedical literature (Rodriguez-Esteban, Iossifov & Rzhetsky 2006). Another example in brain imaging is the Neurosynth framework: an automated way of aggregating thousands of research articles and performing automated meta-analyses on brain imaging research (Yarkoni, Poldrack, Nichols, Van Essen & Wager 2011). Figure 1b shows such an automated meta-analysis simply on the term "semantics": among around 11000 articles, 70 studies are identified on which inferences are made about the likelihood of brain activation in relation to this term. There is no further specification of the process, no theory or even any guarantee that this does not also reflect concepts like "semantic memory". However it could provide a useful starting point as an upper bound (since it is relatively sensitive, but highly unspecific) of a brain network underlying specific semantic processes, more than any single study could if it was based on a small sample (Button et al. 2013).

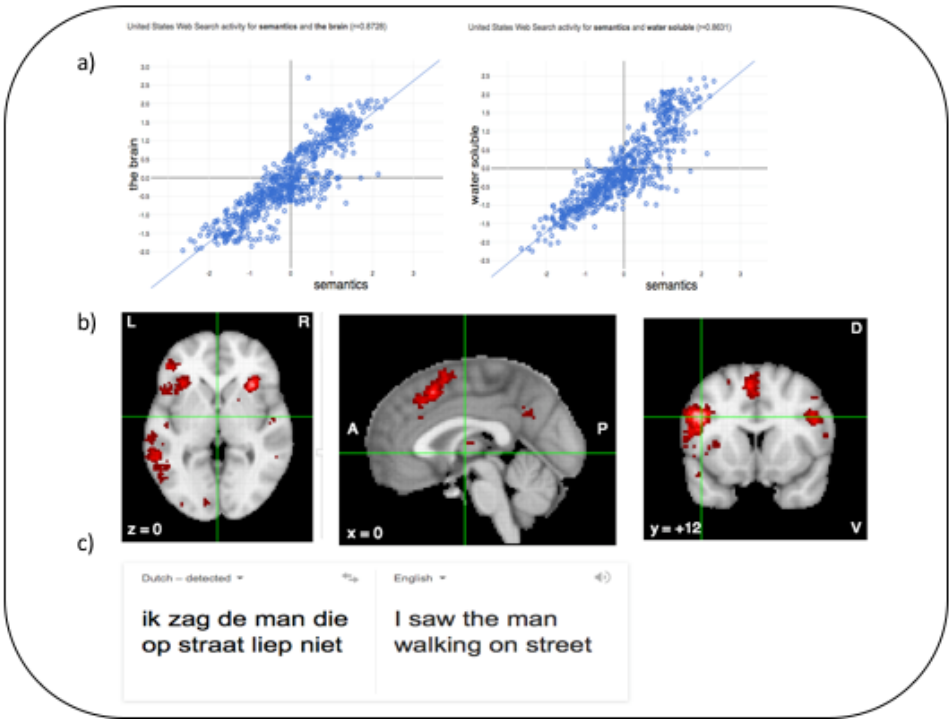

Figure 1. Hits and misses of big-data

a) Correlations between Google searches for “semantics” and “in the brain” (left) and “semantics” and “water soluble” (right)
b) Automated meta-analyses on the brain regions involved in semantics.
c) Google Translate’s understanding of Dutch

However, even incredible amounts of data and tremendous computational power to process it can be mistaken. A now famous example is the large error of Google Flu Trends in predicting influenza prevalence based on Google search terms. In 2012 their estimation was twice as high as the actual prevalence (Lazer et al. 2014). When I tried to “predict” interest in

semantics based on Google search terms, a term like "the brain" did very well (a correlation coefficient r=0.87; see figure 1a), which may seem reasonable given the wide network of brain regions related to semantics we just observed. However, the term "water soluble" does about as good a job, with a correlation coefficient of r=0.83, another number that would make most researchers jealous, yet the relationship seems random. In other words, one may still need some theory to understand data, since huge amounts of data easily produce strong but spurious results. As discussed, one of the core problems of inferences of experimental research data and theory testing is the gap between some theory and its hypothesis. This gap is a necessary evil since hypotheses need to be molded into highly simplified conditions in order to become testable with basic statistics. It has been argued that there is nothing so theoretical as a good method (Greenwald 2012), and I would like to end here by briefly mentioning a highly promising technique in that regard to combine computational models with brain imaging data: representational-similarity analyses (Kriegeskorte, Mur & Bandettini 2008). The idea behind this approach is that you can evaluate predictions from computational models against a pattern of brain imaging data and as such, test the neural representation of these models (Kriegeskorte et al. 2008). Tyler and colleagues, for example, associated different syntactic computations with the similarity structure of brain waves measured with MEG data, and found different computational representations in the inferior frontal and temporal gyri (Tyler, Cheung, Devereux & Clarke 2013). Contrasted with "traditional"

techniques, representational-similarity analysis overcomes the common correspondence problem between the units of a computational model and units of neural data (Kriegeskorte et al. 2008) and could therefore be highly influential in integrating (big) (brain) data with computational models.

References

Bargh, J. A., Chen, M. & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*(2), 230–244. http://doi.org/10.1037/0022-3514.71.2.230

Bornmann, L., Mutz, R. & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: a multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE*, *5*(12), e14331. http://doi.org/10.1371/journal.pone.0014331

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. http://doi.org/10.1038/nrn3475

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300. http://doi.org/doi: 10.1016/j.neuroimage.2012.07.004

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. http://doi.org/10.1037/0003-066X.49.12.997

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing: deep neural networks with multitask learning. *the 25th international conference* (pp. 160–167). New York, New York, USA: ACM. http://doi.org/10.1145/1390156.1390177

Coyne, J. C. & de Voogd, J. N. (2012). Are we witnessing the decline effect in the Type D personality literature? What can be learned? *Journal of Psychosomatic Research*, *73*(6), 401–407. http://doi.org/10.1016/j.jpsychores.2012.09.016

Doyen, S., Klein, O., Pichon, C.-L. & Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, but Whose Mind? *PLoS ONE*, *7*(1), e29081. http://doi.org/10.1371/journal.pone.0029081

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y. et al. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347. http://doi.org/10.1073/pnas.1516179112

Falk, R. & Greenbaum, C. W. (1995). Significance Tests Die Hard: The Amazing Persistence of a Probabilistic Misconception. *Theory & Psychology*, *5*(1), 75–98. http://doi.org/10.1177/0959354395051004

Fanelli, D. (2010). “Positive” Results Increase Down the Hierarchy of the Sciences. *PLoS ONE*, *5*(4), e10068. http://doi.org/10.1371/journal.pone.0010068.t003

Fugelsang, J. A., Stein, C. B., Green, A. E. & Dunbar, K. N. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, *58*(2), 86–95. http://doi.org/10.1037/h0085799

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. http://doi.org/doi: 10.1016/j.socec.2004.09.033

Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*(6277), 1037–1037. http://doi.org/10.1126/science.aad7243

Greenberg, S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *BMJ (Clinical Research Ed.)*, *339*, b2680.

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science*.

Hill, R. (2004). Multiple sudden infant deaths -- coincidence or beyond coincidence? *Paediatric and Perinatal Epidemiology*, *18*(5), 320–326. http://doi.org/10.1111/j.1365-3016.2004.00560.x

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124.

http://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology*, *19*(5), 640–648. http://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science : a Journal of the Association for Psychological Science*, *7*(6), 645–654. http://doi.org/10.1177/1745691612464056

Ioannidis, J. P. A., Boyack, K. W., Small, H., Sorensen, A. A. & Klavans, R. (2014). Bibliometrics: Is your most cited work your best? *Nature*, *514*(7524), 561–562. http://doi.org/10.1038/514561a

Kahneman, D. (2013). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Klein, R. A., Ratliff, K., Vianello, M., Adams, R. B., Jr & Bahník, S. (2014). Investigating variation in replicability: A“ many labs” replication project. Retrieved from Open Science Framework.

Kramer, A. D. I., Guillory, J. E. & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. http://doi.org/10.1073/pnas.1320040111

Kriegeskorte, N. (2012). An emerging consensus for open evaluation: 18 visions for the future of scientific publishing, 1–5. http://doi.org/10.3389/fncom.2012.00094/abstract

Kriegeskorte, N., Mur, M. & Bandettini, P. (2008). Representational

similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*.

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205. http://doi.org/10.1126/science.1248506

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (the Statistician)*, *49*(3), 293–337.

Loftus, G. R. (1993). A picture is worth a thousandp values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, *25*(2), 250–256. http://doi.org/10.3758/BF03204506

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, *9*(2), 147–163. http://doi.org/10.1037/1082-989X.9.2.147

Meehl, P. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*.

Meehl, P. E. (1993). Philosophy of science: Help or hindrance? *Psychological Reports*, *72*(3), 707–733. http://doi.org/10.2466/pr0.1993.72.3.707

Miller, G. (2012). The Smartphone Psychology Manifesto. *Perspectives on Psychological Science : a Journal of the Association for Psychological Science*, *7*(3), 221–237. http://doi.org/10.1177/1745691612441215

Munafo, M. (2013). Deep Impact: Unintended consequences of Jorunal Rank, 1–33.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S. & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 1–22. http://doi.org/10.3758/s13428-015-0664-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. http://doi.org/10.1126/science.aac4716

Pfeiffer, T., Bertram, L. & Ioannidis, J. P. A. (2011). Quantifying selective reporting and the Proteus phenomenon for multiple datasets with similar bias. *PLoS ONE*, *6*(3), e18362. http://doi.org/10.1371/journal.pone.0018362

Rodriguez-Esteban, R., Iossifov, I. & Rzhetsky, A. (2006). Imitating manual curation of text-mined facts in biomedicine. *PLoS Computational Biology*, *2*(9), e118. http://doi.org/10.1371/journal.pcbi.0020118

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin; Psychological Bulletin*, *86*(3), 638–641.

http://doi.org/10.1037/0033-2909.86.3.638

Rosenthal, R. & Gaito, J. (1963). The Interpretation of Levels of Significance by Psychological Researchers. *The Journal of Psychology*, *55*(1), 33–38. http://doi.org/10.1080/00223980.1963.9916596

Rutledge, R. B., Skandali, N., Dayan, P. & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257. http://doi.org/10.1073/pnas.1407535111

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. http://doi.org/10.1214/10-STS330

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. http://doi.org/10.1177/0956797611417632

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. http://doi.org/10.1037/a0033242.supp

Thompson, W. C. & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor“s fallacy and the defense attorney”s fallacy. *Law and Human Behavior*, *11*(3), 167–187. http://doi.org/10.1007/BF01044641

Tversky, A. & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105.

Tyler, L. K., Cheung, T. P. L., Devereux, B. J. & Clarke, A. (2013). Syntactic computations in the language network: characterizing dynamic network properties using representational similarity analysis. *Frontiers in Psychology*, *4*, 271. http://doi.org/10.3389/fpsyg.2013.00271

Wainer, H. (2007). The most dangerous equation. *American Scientist*.

Wicherts, J. M. & van Assen, M. A. L. M. (2012). Research fraud: Speed up reviews of misconduct. *Nature*, *488*(7413), 591–591. http://doi.org/10.1038/488591b

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. http://doi.org/10.1038/nmeth.1635

Yong, E. (2012, May 17). Replication studies: Bad copy. *Nature*, pp. 298–300. http://doi.org/10.1038/485298a

Young, N. S., Ioannidis, J. P. A. & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, *5*(10), e201. http://doi.org/10.1371/journal.pmed.0050201