### Study Information

1. Title (required)

Sample sizes in psychological research over time

1. Authors (required)

Marjan Bakker

Jake Marszalek

Yvonne Schittenhelm

1. Description (optional)

Many studies in the psychological literature have small sample sizes and are underpowered (Bakker, van Dijk, & Wicherts, 2012; Cohen, 1990; Maxwell, 2004; Stanley, Carter, & Doucouliagos, 2018), which affects generalizability, estimation of effect sizes, and accurate planning of future studies. The most effective way to increase statistical power is to increase sample size. Nevertheless, sample size is not often directly investigated. One of the studies examining sample size in psychological research over time is the study by Marszalek, Barber, Kohlhart, and Holmes (2011). In this study, data from Holmes (1979, 1983) was used. Holmes examined sample sizes in four APA journals in 1955 and 1977. These journals were Journal of Abnormal Psychology (JAbP), Journal of Applied Psychology (JApP), Journal of Experimental Psychology: Human Perception and Performance (JEP), and Developmental Psychology (DP). According to Holmes (1979), “these journals were selected because they represent a broad coverage of psychological research” (p. 284). Marszalek et al. examined and added the sample size data from papers published in the same journals in 1995 and 2006. These years were chosen to investigate whether sample sizes increased as a result of the recommendations by Wilkinson and the Task Force on Statistical Inference (Wilkinson, 1999). They found that the sample sizes in psychology remained small and that the recommendations for increasing sample size had not been integrated in core psychological research.

Around the time that the study by Marszalek et al. (2011) was published, the field of psychology went into a replicability crisis. Many studies failed to replicate, and the use of questionable research practices seemed to be widespread (Doyen, Klein, Pichon, & Cleeremans, 2012; John, Loewenstein, & Prelec, 2012; Open Science Collaboration, 2015). As a consequence, different initiatives were launched to improve psychological science. These initiatives include the introduction of preregistration and registered reports (Chambers, 2013; Nosek, Ebersole, DeHaven, & Mellor, 2018), the awarding of open science badges to studies that share data, materials, and/or are preregistered (Eich, 2014; Kidwell et al., 2016), and an increased focus on direct replication studies and studies with null results (Baxter & Burwell, 2017; Simons, Holcombe, & Spellman, 2014). The crisis has also increased attention to the problems of small sample sizes and underpowered studies and resulted in several calls to increase psychological studies' statistical power (Bakker, Hartgerink, Wicherts, & van der Maas, 2016; Button et al., 2013; Szucs & Ioannidis, 2017).

In the current study, we want to update the study by Marszalek et al. to investigate whether psychological studies' sample sizes have increased, specifically in response to the replicability crisis and the resulting focus on open science. We will add studies published in 2019 from the same journals. Because journals differ in how actively they adopted the open science practices, we will include two additional journals (Psychological Science (PS) and Journal of Experimental Social Psychology (JESP)) that have adopted open science practices more actively.

Our research questions are:

1. Have sample sizes increased over time?
2. Have sample sizes in psychological research increased as a reaction to the replicability crisis?
3. Does the amount of increase in sample sizes depend on Journal level policies (open science)?
4. Does the amount of increase in sample size depend on study level characteristics (open science)?
5. Additionally, we want to investigate how well automatic procedures can select the sample size compared to selecting sample sizes by hand.
6. Hypotheses (required)
7. Sample sizes in psychology have increased over time.
8. Sample sizes in psychology have increased as a reaction to the credibility crisis (i.e., the average sample size in 2019 is larger compared to the average sample size in 2006 and 1995).
9. Sample sizes of studies in journals that have adopted open sciences practices (PS and JESP) more actively have increased more than sample sizes in journals that have not actively adopted these open science practices (JAbP, JApP, JEP, and DP).
10. Sample sizes of studies that have adopted open science practices (e.g., shared data, shared materials, or preregistered studies) are higher compared to sample sizes of studies that have not adopted open science practices.

### Design Plan

1. Study type (required)

Observational study

1. Blinding (required)

No blinding is involved

1. Is there any additional blinding in this study?

No

1. Study design (required)

In this study we will examine the reported sample sizes of studies published in the journals JAbP, JApP, JEP, DP, PS, and JESP in the following years: 1995, 2006, and 2019.

1. Randomization (optional)

No randomization is involved

### 

### Sampling Plan

1. Existing data (required)

Registration prior to creation of data

1. Explanation of existing data (optional)

Please note that we will also include the data already collected and used in the study by Marszalek et al. (2011). However, we will go over that data again to make sure that we collect the sample sizes from the studies in the same way.

1. Data collection procedures (required)

We will include all published papers from the 1995, 2006, and 2019 volumes of the following journals: Journal of Abnormal Psychology, Journal of Applied Psychology, Journal of Experimental Psychology: Human Perception and Performance, and Developmental Psychology, Psychological Science, and Journal of Experimental Social Psychology. If a paper contains multiple studies (which samples are independent), we will include the studies separately. The data of the first four journals in the first two years are already available (Marszalek et al. 2011). However, we will go over these studies as well, to make the coding consistent between the years. We added the 2019 volumes to investigate whether the replicability crisis resulted in larger sample sizes. Because the included journals were not very active in adapting the open science practices (3 of them had a TOP score of 0, 1 (JApP) had a TOP score of 5), we also added these two new journals because these implemented open practices actively (they both have a TOP score of 12 and also award open science badges). See our supplement with information about the open science practices of each journal separately.

A coder will go over all included papers. First the coder identifies whether the paper used data (thereby excluding editorials and or simulation studies, also meta-analyses were excluded). Secondly, when the paper included multiple studies that all used a separate sample, we included them as separate studies (separate rows). Of each study the variables defined under 17 were collected. A random set of 10% of the studies will be checked by a second coder to estimate the inter rater reliability. Discrepancies will be solved by discussion. The second coder will also evaluate the studies of which the first coder isn’t sure.

1. Sample size (required)

We will include all studies from the volumes of the included journals. Based on the study by Marszalek we expect the following numbers of total sample sizes for JAbP, JApP, JEP, and DP, for the years 1995 and 2006. The number of total sample sizes for the year 2019 and for the Journals PS and JESP are based on the number of papers in that journal. Note that some papers might not include studies with participants (e.g., editorials, reviews), and other papers might include multiple studies (with independent sample sizes). The latter will be treated as separate studies, each having a total sample size. Thus final sample sizes will differ.

Table Number of studies per year

Journal 1995 2006 2019

JAbP 74 107 80

JApP 69 112 82

JEP 274 334 109

DP 110 145 216

PS 59 179 152

JESP 24 79 131

1. Sample size rationale (optional)

We included all studies from these journals from these years and expect a total of approximately 2336 studies (see table above).

We used the estimated number of included studies presented in the table above in a power analysis for hypothesis 2 and 3. Based on the numbers above, and assuming an average total sample size of 200 in the years 1995 and 2006, a theta of .5, and alpha = .05, we would have > .8 power to detect an effect of 1.2 (factor increase) or higher for hypothesis 2 and > .8 power to detect an effect of 1.45 (factor increase) or higher for hypothesis 3. See the plots for the power to detect different effects (factor increase). Assuming a total of 183 included studies (PS and JESP in 2019) and half of them having an open science badge, we have 80% power to detect a Cohen’s d of 0.36 with alpha = .05 (two-sided). Please see the plots for the statistical power given different values of cohen’s d and different allocation percentages. See the code for the power calculations of H2 and H3. To calculate this power we used G\*power:

**t tests -** Means: Wilcoxon-Mann-Whitney test (two groups)

**Options:** A.R.E. method

**Analysis:** Sensitivity: Compute required effect size

**Input:** Tail(s) = Two

Parent distribution = min ARE

α err prob = 0.05

Power (1-β err prob) = 0.8

Sample size group 1 = 141

Sample size group 2 = 142

**Output:** Noncentrality parameter δ = 2.8127371

Critical t = 1.9697942

Df = 242.512

1. Stopping rule (optional)

We will include all studies from the specified years.

### Variables

1. Manipulated variables (optional)

No manipulated variables are included

1. Measured variables (required)

*Year*: 1995, 2006, and 2019

*Journal*: Journal of Abnormal Psychology (JAbP), Journal of Applied Psychology (JApP), Journal of Experimental Psychology: Human Perception and Performance (JEP), and Developmental Psychology (DP), Psychological Science (PS) and Journal of Experimental Social Psychology (JESP).

*Open science journal*: Inactive (= 0; JAbP, JApP, JEP, DP) and active (= 1; PS, JESP)

*Sample size (N):* This is our primary outcome variable. For each study (a paper can contain multiple studies, the samples of these studies need to be independent), we collect the total sample size after exclusions. If we find inconsistencies between reported sample sizes (e.g., between sample sizes reported in the text and sample sizes reported in a table), we choose the sample size that matches the degrees of freedom of the analysis. Furthermore, if the level of analysis of the main research question is not on the individual level (e.g., groups or parent-child dyads), we will use the number of observations (groups/dyads) at the level of analysis.

*Individual sample size (N\_i):* If the study compares groups (e.g., different experimental conditions, existing groups like gender), the sample sizes per group will be collected. If the exact individual sample sizes are not given, we assume that participants are divided equally over the groups.

*Open science paper; preregistered (OS\_prereg):* No (= 0), yes (= 1; paper has preregistration badge)

*Open science paper; open data (OS\_data):* No (= 0), yes (= 1; paper has open data badge)

*Open science paper; open materials (OS\_mat):* No (= 0), yes (= 1; paper has open materials badge)

If different names used in the analysis code, they are added between brackets.

1. Indices (optional)

We use the following indices:

*YearC*: 1995 = 0; 2006 = 2; 2019 = 3

*Replicability crisis (RepCris)*: before (= 0; year 1995 and 2006) and after (= 1; year 2019)

*Open science journal (OpenScJour)*: Not active (= 0; JAbP, JApP, JEP, DP) and active (= 1; PS, JESP)

*Open science paper: general (OS)*: We combine the preregistration, open data, open materials items into one single open science general item. This item is scored 1 if at least one of the three items scores 1 and 0 otherwise.

If different names used in the analysis code, they are added between brackets.

### Analysis Plan

1. Statistical models

Our primary outcome variable is sample size, which resemble count data. We therefore will use Hierarchical Gereralized Linear Modelling (HGLM) to model the sample size development over time and to take both the count data as the nesting (studies, year, journals) into account. Because the data in Marszalek et al showed overdispersion, we will use the negative binomial distribution.

To answer the first hypothesis we used the following model:

Level 1:

Level 2:

Level 3:

Combined:

Specifically, refers to the overall log average sample size and to the fixed effect of the time. The random effects are represented with for the journal intercept, and for the random slope of year (effect of year within journal), and for the variance within journal year. We expect a positive , which would indicate that sample sizes have increased over time (hypothesis 1).

With this analysis, we follow the analysis plan of Marszalek et al. However, we might not be able to fit this model (e.g., failure to converge). In these cases, we will simplify the model or switch to a model with fixed effects of journal and year.

To test the second and third hypothesis we will only include fixed effects and switch to a negative binomial regression model:

We expect a positive *b3*, which would indicate that Journals that adapted the Open Science Practices showed a stronger increase in sample sizes than the other journals (hypothesis 3). If *b3* is signficantly different from zero, we will continue with simple effects analyses to test seperately for the journals that did and did not adapted open science practices showed an increase in sample size. When *b3* is not significant, we will run the model again without the interaction and test over all journals whether sample sizes increased after the replication crisis. More specifically, to confirm hypothesis 2, we would expect a significant *b1*.

To test the fourth hypothesis, we will only include the sample sizes from the open science journals from the year 2019. Sample sizes of studies that used at least one open science practice are compared with the sample sizes of studies that didn’t use an open science practice with a Mann-Whitney U test.

We will also run the models both for the total sample sizes (main analysis) as for the individual sample size (additional explorative analyses).

See also our analysis code for the specific implementation

1. Transformations (optional)

We will use dummy coding for the categorical variables. In the variable section we indicated which of the categories is the reference category (=0).

1. Inference criteria (optional)

Because we test 4 hypotheses, our inference criteria is α = .05/4.

1. Data exclusion (optional)

We will include papers that analyze data from persons (or groups/dyads), but will exclude the following papers:

* meta-analyses
* editorials
* simulation studies
* reviews
* we might encounter other types of papers studies that we will exclude. We will mention these other types clearly as additional exclusion types in the final manuscript.

1. Missing data (optional)

All studies that report a sample size will be included, no missing data

1. Exploratory analysis (optional)

Exploratory, we will do the same analyses (described above) on the individual sample sizes.

To develop statcheck (Nuijten & Epskamp, 2018) further, we will select the texts that we used to base the sample sizes on. Furthermore, we will run statcheck on the included papers to collect the degrees of freedom to estimate the sample sizes. By doing so, we can compare our manually methods with automatic procedures.

### Other

1. Other (Optional)

Refererences

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers’ intuitions about power in psychological research. *Psychological Science, 27*(8), 1069-1077. doi:10.1177/0956797616647519

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554. doi:10.1177/1745691612459060

Baxter, M. G., & Burwell, R. D. (2017). Promoting transparency and reproducibility in Behavioral Neuroscience: Publishing replications, registered reports, and null results. *Behavioral Neuroscience, 131*(4), 275-276. doi:10.1037/bne0000207

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 1-12. doi:10.1038/nrn3475

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609-610. doi:10.1016/j.cortex.2012.12.016

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist, 45*, 1304-1312. doi:10.1037/0003-066X.45.12.1304

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). doi:10.1371/journal.pone.0029081

Eich, E. (2014). Business not as usual. *Psychological Science, 25*, 3-6. doi:10.1177/0956797613512465

Holmes, C. B. (1979). Sample size in psychological research. *Perceptual and Motor Skills, 49*(1), 283-288.

Holmes, C. B. (1983). Sample Size in Four Areas of Psychological Research. *Transactions of the Kansas Academy of Science (1903-), 86*(2/3), 76-80. Retrieved from <http://www.jstor.org/stable/3627914>

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. doi:10.1177/0956797611430953

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., . . . Hess-Holden, C. (2016). Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLoS Biol, 14*(5), e1002456. doi:10.1371/journal.pbio.1002456

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*(2), 331-348. doi:10.2466/03.11.pms.112.2.331-348

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147-163. doi:10.1037/1082-989X.9.2.147

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600-2606. doi:10.1073/pnas.1708274114

Nuijten, M. B., & Epskamp, S. (2018). statcheck: Extract statistics from articles and recompute p-values (1.3.1) [R package]. Retrieved from <https://cran.r-project.org/web/packages/statcheck/index.html>

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943. doi:10.1126/science.aac4716

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to registered replication reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9*(5), 552-555. doi:10.1177/1745691614543974

Stanley, T., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*. doi:10.1037/bul0000169

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 15*(3), e2000797. doi:10.1371/journal.pbio.2000797

Wilkinson, L. (1999). Statistical methods in psychology journals - Guidelines and explanations. *American Psychologist, 54*(8), 594-604. Retrieved from <Go to ISI>://000081919300009