Collaboration on Statistical Analyses and Statistical Reporting Errors in Published Psychology Articles: project 2

Pre-registration

# Study Information

## Title

Co-pilot 2.0: Collaboration on statistical analyses and statistical reporting errors in published Psychology Articles: project 2

## Authorship

Coosje L. S. Veldkamp, Chris H. J. Hartgerink, Marcel A. L.M. van Assen, Jelte M. Wicherts

## Background

Over the last decade, it has become clear that many published research articles contain errors in the reported statistical results (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; Caperos & Pardo, 2013; Garcia-Berthou & Alcaraz, 2004; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014; Wicherts, Bakker, & Molenaar, 2011). Veldkamp et al. (2014) investigated whether collaboration on statistical analyses, as proposed by Wicherts (2011), was associated with reduced error rates in psychology articles. Their hypothesis was that collaboration reduces errors in the reporting of statistical results: replication of statistical analysis and double checking of statistical results decrease the probability that researchers make rounding and retrieval errors and diminish the possibility that serious errors due to confirmation bias go unnoticed. While their study confirmed the high reporting error rates in psychology and established that collaboration on statistical analysis is uncommon among psychologists, they found no association between collaboration and reporting error rates. However, their data had several limitations that may have limited the study’s capability to detect this potential association. First, estimation of collaboration on statistical analysis was based on a questionnaire asking authors to recall who had been involved in the statistical analyses reported in their paper, which may have yielded inaccurate responses. Second, e-mails from the respondents to the authors in response to the survey showed that the questionnaire was perceived as sensitive, possibly leading to response bias and expectancy effects. Third, collaboration was defined as ‘more than one author being involved in the statistical analyses’, prohibiting analysis of the relationship between reporting errors and *number* of authors involved in the analysis. Fourth, the study suffered from lack of power due to a low overall response rate and the use of a relatively limited number of *p*-values and articles.

For these reasons, we propose an alternative design to examine the hypothesis that collaboration on statistical analysis can help reduce statistical reporting errors: by relating statistical reporting errors to the number of authors responsible for the data analysis as reported in the ‘author contributions’ section of all psychology research articles published in the multidisciplinary journal PLoS One to date (as of April 1, 2016: 13,480 articles). This does not only allow us to more objectively determine how many authors were involved in the statistical analyses, but also to examine in a much larger sample whether there is an association between the number of authors involved in the analyses and error rates. In addition, downloading other meta-data such as declared conflicts of interest or involvement of the first author in the statistical analyses will allow investigation of other potential factors associated with statistical reporting errors. While our software to detect statistical reporting can currently only be applied to articles employing the reporting style of the American Psychological Association (APA) style (American Psychological Association, 2010) we will also extract the author contribution information and other meta-data from all other research articles published in all PLoS journals to date (currently 168,665). This will enable us to compare major research areas in terms of mean number of authors responsible for the statistical analyses relative to the number of authors on the article, and in terms of how often the first author of the paper is listed among those responsible for the data analysis.

As in previous studies (Bakker & Wicherts, 2011; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015; Veldkamp, Nuijten, Dominguez-Alvarez, van Assen, & Wicherts, 2014), we differentiate between small errors (errors) and serious errors (gross errors). A small error constitutes a reporting error that does affect conclusions about the statistical significance of a test, such as rounding errors or minor retrieval errors. A serious error constitutes a reporting error that (may) have affected conclusions about statistical significance at a significance level of *p* < 0.05: when a *p*-value is reported as smaller than 0.05 when it was in fact larger than or equal to 0.05 or vice versa. We believe small errors to result from incorrect use of rounding rules, hurry, or sloppiness, whereas we believe serious errors to result from confirmation bias.

## Research Questions

Primary research question:

1. What is the association between collaboration on statistical analysis and the probability of making an error in the reporting of statistical results?

Secondary research questions:

1. What is the association between the first author being involved in the statistical analyses and the probability of making a serious error in the reporting of statistical results?
2. What is the association between having a conflict of interest and the probability of making a serious error in the reporting of statistical results?

Exploratory research questions:

1. What is the association between the number of collaborators on statistical analysis and the probability of making an error in the reporting of statistical results
2. What is the association between other article characteristics and the probability of making an error in the reporting of statistical results?

## Hypotheses

Primary hypotheses:

Collaboration reduces both small errors and serious errors in the reporting of statistical results: Replication of statistical analysis and double checking of statistical results decrease the probability that researchers make rounding and retrieval errors (small errors) and diminish the possibility that serious errors due to confirmation bias go unnoticed. Our specific hypotheses therefore are:

1. The probability that a paper contains at least one *p*-value that is incongruent with its test statistic and degrees of freedom is lower in papers in which more than one person was involved in the statistical analyses than in papers in which only one person was involved in the statistical analyses.
2. The probability that a paper contains at least one *p*-value that is incongruent with its test statistic and degrees of freedom to the extent that it affects statistical significance is lower in papers in which more than one person was involved in the statistical analyses than in papers in which only one person was involved in the statistical analyses.
3. The probability that a *p*-value is incongruent with its test statistic and degrees of freedom is lower if more than one person was involved in the statistical analyses than when only one person was involved in the statistical analyses.
4. The probability that a *p*-value is incongruent with its test statistic and degrees of freedom to the extent that it affects statistical significance is lower if more than one person was involved in the statistical analyses than when only one person was involved in the statistical analyses.

Secondary hypotheses:

It is widely believed that the chance that an article will be published is higher when hypotheses are confirmed than when hypotheses are not confirmed, which makes statistical analysis prone to confirmation bias, leading to *serious* reporting errors (gross errors). As the first author typically has a higher interest in getting the article published than co-authors have, we hypothesize that:

1. The probability that a paper contains at least one *p*-value that is incongruent with its test statistic and degrees of freedom to the extent that it affects statistical significance is higher in papers in which the first author was involved in the statistical analyses than in papers in which the first author was not involved in the statistical analyses.
2. The probability that a *p*-value is incongruent with its test statistic and degrees of freedom to the extent that it affects statistical significance is higher in papers in which the first author was involved in the statistical analyses than in papers in which the first author was not involved in the statistical analyses.

Authors’ conflicts of interest may also make statistical analysis more prone to confirmation bias, leading to *serious* reporting errors (gross errors). We therefore hypothesize that:

1. The probability that a paper contains at least one *p*-value that is incongruent with its test statistic and degrees of freedom to the extent that it affects statistical significance is higher in papers in which (at least one of) the authors declared a conflict of interest than in papers in which the authors declared no conflict of interest.
2. The probability that a *p*-value is incongruent with its test statistic and degrees of freedom to the extent that it affects statistical significance is higher in papers in which the authors declared a conflict of interest than in papers in which (at least one of) the authors declared no conflict of interest.

Exploratory hypotheses:

Because diffusion of responsibility (Wallach, Kogan, & Bem, 1964) and the potential for ‘social loafing’ (Karau & Williams, 1993; Latane, Williams, & Harkins, 1979), having too many people involved in the statistical analyses may increase the probability of making errors in the reporting of statistical results just as much as having only one person involved in the statistical analyses. Therefore, we explore a possible nonlinear association of the number of authors involved in the statistical analyses with the probability of making an error in the reporting of statistical results.

In addition, we explore whether there are associations between the probability of making a serious error in the reporting of statistical results and 1) number of authors on the paper, 2) research area (as defined by PLoS journal), and 3) publication year, and possibly some other variables as well.

# Sampling plan

## Existing data

This registration occurs prior to creation of data: as of the date of submission of this research plan for preregistration, the data have not yet been collected, created, or realized.

## Explanation of existing data

Not applicable

## Data collection procedures

Using an R script written by Chris Hartgerink (making use of the package ‘rplos’), we will download meta-data (see variables section) of all psychology-tagged articles published in all of the PLoS journals to date. Subsequently, all statistical results reported in these psychology articles will be retrieved from the downloaded results sections by the R package ‘statcheck’ (Epskamp & Nuijten, 2015; Nuijten et al., 2015) and collected into a second data file. ‘Statcheck’ will automatically indicate for each retrieved statistical result whether the *p*-value is consistent with its test statistic and the associated degrees of freedom, and if it is inconsistent, whether the inconsistency may have affected decisions about statistical significance (a ‘gross’ error).

In order to allow for future further analysis of author contributions in the entire population of PLoS articles, we will download the meta-data of the entire population of research articles published in all PLoS journals to date, which currently (April 1, 2016) amounts to 168.665 articles.

As we will write our report in an R markup file, we will be able to update results as the body of articles in PLoS One grows. The duration of the data collection procedure will depend on computer processing times, but we expect to have collected all data within two weeks of starting the collection procedure.

## Sample size

Currently, 13,480 psychology-tagged research articles have been published in PLoS One, while only 835 psychology-tagged articles have been published in the other PLoS journals. Because of this imbalance, we decided to only include the psychology-tagged articles published in PLoS One in most of our analyses. Only exploratory hypotheses 7 and 8 will be tested in the sample that does include the 835 psychology-tagged articles published in the other PLoS articles.

Previous research on the prevalence of statistical reporting errors in published psychology articles shows that psychology journals and articles vary widely in the number of statistical results that they report (Bakker & Wicherts, 2011; Nuijten et al., 2015; Veldkamp et al., 2014). The average number of statistical results that the R package ‘statcheck’ retrieves from articles not only depends on the number of results that are reported in the articles in a given sample, but also on whether the results are reported exactly according to the APA reporting guidelines (American Psychological Association, 2010). From psychology articles published in all PLoS journals between 2003 and 2013, ‘statcheck’ retrieved statistical results from 24% of the articles, and retrieved an average of 9 statistical results from these articles (Nuijten et al., 2015). The percentage of articles from which statistical results were retrieved was lower in PLoS journals than in APA journals. The reason for this difference may be that while most psychological researchers do report their results according to the APA reporting guidelines, this reporting style is not required by PLoS. If we apply these latter numbers to our sample we estimate that statcheck will retrieve (0.24\*13,480\*9 =) 29,117 results from (0.24\*13,480 =) 3,235 papers to be used in our main analyses. For the analyses of exploratory analyses 7 and 8 we expect to be able to use an additional (0.24\*835\*9) = 1804 results from (0.24\*835 =) 200 papers.

## Sample size rationale

Our sample size is determined by the current availability of data: we can only use PLoS journals because they are currently the only journals that systematically report author contributions that are readily retrievable. Within the PLoS journals, we are limited to psychology articles because the package ‘statcheck’ is currently only able to retrieve statistical results reported according to APA guidelines.

## Stopping rule

Not applicable

# Variables

## Manipulated variables

Not applicable

## Measured variables

Data to collect

*All research articles PLoS journals*

* Doi
* Publication date
* Authors (full names, in order of authorship)
* Author contributions
* Journal

*For all psychology articles, also:*

* Full text paper

Variables to extract from papers

* Year of publication
* month of publication
* First author name
* List of abbreviated author names (initials first name and surname) [created based on author names. Note: initials will be checked by automatic comparison to all initials appearing in all author contributions. Only those that match will be used in the analyses (of secondary hypotheses 1 and 2).
* Count number of authors on paper [number] *(all research papers)*
* Competing interest [0/1]
* List of abbreviated author names of those who ‘Conceived and designed the experiments’ [in author contributions 1] *(all research papers)*
* List of abbreviated author names of those who ‘Performed the experiments’ [in author contributions 2] *(all research papers)*
* List of abbreviated author names of those who ‘Analyzed the data’ [in author contributions 3] *(all research papers)*
* List of abbreviated author names of those who ‘Wrote the paper’ [in author contributions 4] *(all research papers)*
* Count number of authors per contribution [number] (all research paper)
* Full-text *(psychology articles only)*
  + Extract from results section:
    - All test-statistics in APA format
    - All *p*-values in APA format
    - Whether these *p*-values are consistent or not
    - Whether the inconsistent *p*-value is a gross error or not
  + Also include in statcheck file
    - Number of contributors on data-analysis [count]
    - First author in data-analysis [0/1]
    - Competing interest [0/1]
    - Year
  + Number of *p*-values in APA format per paper [number]
  + Number of inconsistent *p*-values in APA format per paper [number]
  + Number of gross errors per paper [number]

Resulting variables collected in three data files:

1. Aggregate data ALL articles in PLOS journals
2. Aggregate data all PSYCH articles in PLOS journals
3. Statcheck data all PSYCH articles in PLOS journals

## Indices

Not applicable

# Design plan

## Study type

Observational study

## Blinding

Not applicable

## Study design

This is a cross-sectional study.

*Randomization*

Not applicable

# Analysis plan

Our sample is systematic rather than random, which precludes statistical inference. We will fit our models to the entire population of psychology articles published in PLoS One to date in order to predict statistical reporting errors in this population. However, in addition to our model parameter estimates, we will report test statistics and *p*-values for those interested.

## Statistical models

We will fit the following statistical models to test our hypotheses:

Primary hypotheses:

1. Generalized linear model (logistic)

* ***glm(y ~ 1 + co-piloted, data=df, family=binomial(link='logit'))***
* *Dependent variable (y): whether a paper contains at least one p-value that is congruent (1/0)*
* *Predictor (co-piloted): whether the analysis were conducted by more than one person (1/0)*

1. Generalized linear model (logistic)

* ***glm(y ~ 1 + co-piloted, data=df, family=binomial(link='logit'))***
* *Dependent variable (y): whether a paper contains at least one p-value that is congruent to the extent that it affects statistical significance (1/0)*
* *Predictor (co-piloted): whether the analysis were conducted by more than one person (1/0)*

1. Generalized linear mixed model (logistic)

* ***glmer(y ~ 1 + co-piloted + (1|article\_id) ,data=df, family=binomial(link='logit'))***
* *Dependent variable (y): whether a p-value is incongruent (1/0)*
* *Fixed factor (co-piloted): whether the analysis were conducted by more than one person (1/0)*
* *Random factor (article\_id): article from which the p-value was extracted*

1. Generalized linear mixed model (logistic)

* ***glmer(y ~ 1 + co-piloted + (1|article\_id) ,data=df, family=binomial(link='logit'))***
* *Dependent variable (y): whether a p-value is incongruent to the extent that it affects statistical significance (1/0)*
* *Fixed factor (co-piloted): whether the analysis were conducted by more than one person (1/0)*
* *Random factor (article\_id): article from which the p-value was extracted*

Secondary hypotheses:

1. Generalized linear model (logistic)

* ***glm(y ~ 1 + first\_author\_involved, data=df, family=binomial(link='logit'))***
* *Dependent variable (y): whether a paper contains at least one p-value that is incongruent to the extent that it affects statistical significance (1/0)*
* *Predictor (first\_author\_involved): whether the first author was involved in the analyses (1/0)*

1. Generalized linear mixed model (logistic)

* ***glmer(y ~ 1 + first\_author\_involved + (1|article\_id) ,data=df, family=binomial(link='logit'))***
* *Dependent variable (y): whether a p-value is incongruent to the extent that it affects statistical significance (1/0)*
* *Fixed factor (first\_author\_involved): whether the first author was involved in the analyses (1/0)*
* *Random factor (article\_id): article from which the p-value was extracted*

1. Generalized linear model (logistic)

* ***glm(y ~ 1 + conflict\_of\_interest, data=df, family=binomial(link='logit'))***
* *Dependent variable (y): whether a paper contains at least one p-value that is incongruent to the extent that it affects statistical significance (1/0)*
* *Predictor (conflict\_of\_interest): whether the authors declared a conflict of interest (1/0)*

1. Generalized linear mixed model (logistic)

* ***glmer(y ~ 1 + conflict\_of\_interest + (1|article\_id) ,data=df, family=binomial(link='logit'))***
* *Dependent variable (y): whether a p-value is incongruent to the extent that it affects statistical significance (1/0)*
* *Fixed factor (conflict\_of\_interest): whether the authors declared a conflict of interest (1/0)*
* *Random factor (article\_id): article from which the p-value was extracted*

## Transformations

No transformations are planned for our confirmatory analyses. For our exploratory analyses, we will dummy code the number of authors in involved in the analyses and examine which function (linear or non-linear) fits best.

## Follow-up analyses

Our exploratory analyses in which we examine a possible nonlinear association of the number of authors involved in the statistical analyses with the probability of making an error in the reporting of statistical results essentially constitutes a follow-up analysis for each of the four primary analyses. No other follow-up analyses are planned.

## Inference criteria

As specified at the beginning of the analysis plan, our sampling technique precludes inference. We will however report test statistics and *p*-values for those interested, maintaining an alpha of 0.05.

## Data exclusion

We will not exclude any data points unless they obviously constitute errors in data extraction. An example of such obvious errors would be *p*-values larger than 1, which has been found to occur when the package statcheck retrieves a p-value is reported as ‘p times 10 to the power of’ (Veldkamp et al., 2014). Exclusion of data point due to such error will be described in detail in the sample description in our report.

## Missing data

Missing values will be excluded based on pairwise deletion.

## Exploratory analysis

See analysis plan.

# Script

See project page for the data extraction script.

# References

American Psychological Association. (2010). *Publication Manual of the American Psychological Association. Sixth Edition*. Washington, DC: American Psychological Association.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43*, 666-678. doi: 10.3758/s13428-011-0089-5

Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research, 16*(4), 202-207. doi: 10.1002/mpr.225

Caperos, J. M., & Pardo, A. (2013). Consistency errors in p-values reported in Spanish psychology journals. *Psicothema, 25*(3), 408-414. doi: 10.7334/psicothema2012.207

Epskamp, S., & Nuijten, M. B. (2015). R package “statcheck”: Extract statistics from articles and recompute values. Retrieved from <https://github.com/MicheleNuijten/statcheck>

Garcia-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and P values in medical papers. *Bmc Medical Research Methodology, 4*, 13. doi: 10.1186/1471-2288-4-13

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of personality and social psychology, 65*(4), 681.

Latane, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of personality and social psychology, 37*(6), 822.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 1-22.

Veldkamp, C. L..S., Nuijten, M. B., Dominguez-Alvarez, L., van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PLoS One, 9*(12), e114876.

Wallach, M. A., Kogan, N., & Bem, D. J. (1964). Diffusion of responsibility and level of risk taking in groups. *The Journal of Abnormal and Social Psychology, 68*(3), 263.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature, 480*, 7. doi: 10.1038/480007a

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS One, 6*(11), e26828. doi: 10.1371/journal.pone.0026828