Pre-Registration
[Power, PPV and Publication Bias of Cyber Security User Studies][*]

Thomas Groß
Newcastle University
United Kingdom

General Purpose of Pre-Registrations

Pre-registrations are research statements of intention established before a sample is evaluated and statistical inferences are undertaken. A pre-registration asserts the aim of a study, including its research questions and statistical hypotheses, methods, incl. operationalization of independent variables (IVs) and dependent variables (DVs), sample and analysis specification.

The primary reason for a pre-registration lies in the fact that a statistical inference (Null Hypothesis Significance Testing) is only valid if the statistical hypotheses are fixed before the inference is undertaken. This is grounded in a *p*-value being a conditional likelihood contingent on the fixed null hypothesis assumed to be true. Furthermore, pre-registrations serve as a ward against questionable research practices, such as outcome-switching, hypothesizing after the results are known (HARKing), or *p*-hacking…it is meant to counteract the many temptations of researcher degrees of freedom.

Pre-registrations are typically committed confidentially under embargo, with an immutable timestamp. Once the corresponding study is published, the embargo is lifted.

This is an experiment registration form for the Open Science Framework (OSF)[1]. It is modelled according to the format of AsPredicted[2].

Context of this Pre-Registration

Meta-Data of Pre-Registration.

- Open Science Framework Repository: `https://osf.io/bcyte/`
- Registered Registration File: `https://osf.io/68d3f/`—Prereg_Power_PPV_PubBias.pdf
- Timestamp: 2020-07-02 06:08 PM
- Archived Immutable Pre-Registration: `https://osf.io/k35yh`
- Timestamp: 2020-07-02 6:37 PM

Peer-Reviewed Publication. The definitive version of the study is published as:
Thomas Groß. Statistical Reliability of 10 Years of Cyber Security User Studies. In Proceedings of the 10th International Workshop on Socio-Technical Aspects in Security (STAST'2020), LNCS 11739, Springer Verlag, 2020.

ArXiv Report. Thomas Groß. Statistical Reliability of 10 Years of Cyber Security User Studies (Extended Version). arXiv:2010.02117, 2020. `https://arxiv.org/abs/2010.02117`

1 Structured Abstract

Background. Cyber security security user studies have been scrutinized in recent years on their reporting completeness for statistical inferences as

[*] Open Science Framework: `https://osf.io/bcyte/`

[1] `https://osf.io`

[2] `https://aspredicted.org`

well as their statistical reporting fidelity. However, other benchmarks of sound research, such as statistical power, estimates of Positive Predictive Value (PPV) and publication bias have been largely absent in the meta-research on the field.

Aim. We aim to estimate the power, PPV, and publication bias distribution over an SLR-derived sample of cyber security user studies.

Method. Based on an earlier published SLR of 146 cyber security user studies, we will extract correctly reported test triplets (test statistic, degrees of freedom, and *p*-value), the overall study sample sizes and group sizes of statistical tests, in addition to test families and multiple-comparison corrections. Based on that data we will compute effect sizes for parametric comparisons between conditions in the form of t-tests, χ^2-tests, or one-way F-tests. We will convert all such effect sizes into Standardized Mean Differences (SMD, Hedges g) for comparisons across studies. Based on these post-hoc effect size estimates, we will compute we will estimate confidence intervals as well funnel plots for the estimation of publication biases. Furthermore, we evaluate detection sensitivity, statistical power and PPV in face of parametrized *a priori* effect size thresholds.

Anticipated Results. While we expect based on earlier results that the sample will only partially yield usable effect size estimates (and thereby estimates for further benchmarks), we anticipate that the results will offer a plethora of data characterizing the field.

Anticipated Conclusions. We anticipate that the benchmarks provided will offer an empirical evidence base to inform the community how we are doing and substantiate recommendation on how to advance the field.

2 State of Data Collection

Have any data been collected for this study yet?

(a) □ **NO** data have been collected.
(b) □ Some data have been collected, but not analyzed.
(c) ☑Some data have been collected and analyzed.

If (b) or (c), please explain briefly:

The sample we analyze has been collected and published in an earlier SLR by Coopamootoo and Groß [1] on 10 years of cyber security user studies (2006–2016); they have also computed effect sizes on a sub-sample of t-tests. Groß [5, 4] has already computed statistical inferences on that sample by recomputing statistical inferences with the R tool statcheck [3]. However, given that said sample serves as benchmark of the community's state-of-play and given that we focus on estimation theory and not null hypothesis significance testing, we believe focusing on the same sample is well called for.

3 Aims

Hypothesis: What's the main question being asked or hypothesis being tested?

We seek to quantify for a meaningful sub-sample of the SLR sample

1. standardized effect size of comparisons made,
2. group sample sizes,
3. statistical power of comparisons made in the form of (i) post-hoc power, (ii) simulated *a priori* power vis-à-vis of parametrized effect sizes in the field, (iii) sensitivity (min. effect size detectable at parametrized power),
4. positive predictive value (incl. accounting for bias and prior estimates),
5. publication bias.

We retain it as a secondary goal to consider the development of these variables over time, venues and sampling approaches (e.g., MTurk vs. lab), however, believe that—due to the need to exclude publications and statistical tests with with insufficient reporting for research synthesis—we will retain too small a sample for a robust estimation of a regression.

4 Methods

Give a brief overview of the methods used.

We are using a methodology founded in meta-analysis, however, do not intend to establish a meta-analysis itself. The reason for that is that the effects investigated in the sample are on vastly different constructs, direct or conceptual replications few and far between. Hence, the effects are not comparable on meta-analytic grounds. However, we intend to gain an estimation of the rough distribution of the variables in question, already informative in itself.

5 Data Preparation

Describe what measures will be taken to check assumptions and label outliers.

We will extract data from the papers with the R tool statcheck. We will manually code whether the assumptions of those extracted statistical tests were fulfilled, for instance, whether the correct tests were used in a dependent-sample situation. We will further check whether there seemed to have been errors in the computations.

6 Main Analyses

Describe what analyses (e.g., t-test, repeated-measures ANOVA) you will use to test your main hypotheses.

While we do not focus on null hypothesis significance testing, we compute a number of estimation methods.

Effect Sizes. We intend to compute effect sizes with the R packages (i) metafor, (ii) esc, and (iii) compute.es. These packages will be used for different estimation routes, namely (i) estimation from summary statistics (such as mean, SD, and group sizes), (ii) estimation from test statistics (such as χ^2 value and overall sample size), (iii) conversion from given effect sizes equivalents (such as from product-moment correlation r to Hedges g).

Therein, we largely focus on parametric and independent-sample effect sizes with one-by-one comparisons. That is largely, because (i) non-parametric statistics, in many cases, cannot be meaningfully standardized to parametric equivalents, (ii) often dependent-sample comparisons present in papers often either violate independent-observation assumptions (by using an independent-samples test for a dependent-samples setup) or fail to provide the data necessary to compute meaningful dependent-sample effect sizes (especially, the correlation between within-subject groups). (iii) multiple-comparison tests (such as multi-way ANOVAs) do not lend themselves easily to compute meaningful effect sizes unifiable to one-to-one comparisons unless variance explained on co-variates is available (usually unreported).

Sample and Group Sizes. We intend to extract total sample sizes of studies and group sizes of individual tests by manual coding from the papers.

Standard Errors and Confidence Intervals. We intend to estimate the standard errors from the estimated sampling variance, which is largely done with R packages metafor or esc out of the box. To compute confidence intervals, we use MBESS for estimates based on the standard errors observed.

Test Families and MCC. We intend to code multiple-comparison corrections and test families based on coding from the papers. Comparisons separate by different studies will constitute different test families. Within studies, we will consider as one test family tests of the same type executed on the same sample even if it is on multiple variables.

We intend to compute adjusted p-values and confidence intervals (for simplicity only with a Bonferroni correction) for test families we identify.

We take into account the MCC corrections reported in papers by manual coding from mentions of, e.g., “multiple comparison,” “correction,” “multiplicity,” “Bonferroni,” “Holm,” “Benjamini,” “Hochberg.” We intend to consider adjusted significance levels as well as adjusted p-values.

ES Unification. We will standardize effect sizes across studies as Standardized Mean Difference (Hedges g). Even if the constructs investigated are not comparable, this estimate will show us quantitatively how effects investigated are distributed vis-à-vis of their standard errors and confidence intervals. It thereby answers what sizes of effects does the community usually investigate with what confidence.

Publication Bias. We will illustrate publication bias with effect size vs. standard error funnel plots, computed with metafor.

Per-Study Aggregation. To gain summary statistics per study, we consider computing the average Standardized Mean Difference (Hedges *g*) together with the average standard error over the tests reported in the study.

Power. Of course, post-hoc power drawn from observed effect sizes in given studies is notoriously unreliable—and well known for that fact. The principal reason for that is that effect sizes drawn from small-sample studies tend to be unreliably estimated and tend to be over-estimated because of publication bias. Thereby, the post-hoc power on those over-estimated effect sizes tends to be over-estimated as well. By considering post-hoc power, researchers tend to fool themselves.

While we use observed effect sizes for the publication bias analysis, for power we predominately will look at parametrized effect size thresholds. That is, we will ask what power would that study have had, if it were conducted on a hypothetical population effect size of say, $\gamma = 0.5$. We compute this with the sample and group sizes presented in the papers, at a significance level of $\alpha = .05$.

For sensitivity analyses, we intend to use power thresholds of $1 - \beta \in \{.60, .70, .80, .90, .95, .99\}$ and significance level $\alpha = .05$.

PPV. The Positive Predictive Value (PPV) estimates how likely a positive report is true in reality. We use Ioannidis' estimation formulae [7] to compute the PPV including estimations of bias and prior probability.

In first instance, we will simulate vs. parametrized thresholds of prior and bias, basically offering variants of the analysis for fundamental assumptions on the field as a whole.

Estimating Prior and Bias. As a secondary analysis, we intend attempt a rough estimation on prior and bias, as well.

To estimate the prior, we intend to use his proposal of computing the odds of likely true and likely false comparisons.

To estimate the bias, we orientate on Ioannidis' estimations of biases [7] of specific study types, e.g., 0.2 for well-run RCTs. We intend estimate the bias with an empirical grounding as follows:

1. base bias for all studies of 0.2 (which Ioannidis claims for very well run RCTs),
2. difference between RCT and non-RCT as estimated from Jadad scoring [8, 6]. That is, a study with full Jadad score will be considered an RCT; proportional to diminishing Jadad score, the study will incur increased bias down to the non-RCT level.
3. difference between well-run and not-well run estimated from the reporting completeness. That is, studies which offer complete reporting in terms of the Coopamootoo-Groß [2] coding of nine completeness indicators will be considered well-run and gain the corresponding bias proposed by Ioannidis; studies which fail their reporting completeness are considered not to have provided evidence to be well-run and are, thereby, hypothesized to have incurred more bias.

We note that it is of course impossible to *know* the bias of the study itself from the published report.

Especially the final estimate from reporting completeness to well-run-ness of the study is tenuous. Therein, we hypothesize that investigators who were diligent in their reporting were equally diligent in conducting their study. While this seems plausible to us as a general rule, it us not universally true. Specifically, we have encountered studies which seemed to tick the boxes in terms of reporting, but were utterly invalid in their research setup, reasoning, and statistical inference.

At the same time, we hope that such cases are the exception. We hypothesize that, as a general rule, well-reported papers will correlate with well-run studies, at least enough to offer a glimpse at the state of the field, albeit with a grain of salt.

7 Secondary Analyses

Describe what secondary analyses you plan to conduct (e.g., order or gender effects).

We intend to compute secondary analysis on dependence on publication year and venue, use of MTurk vs. lab samples, sample size permitting.

8 Validation

Describe what diagnostics or validation methods you plan to employ to check the soundness of the analyses.

We intend to recompute effect sizes etc. by hand with a secondary software (e.g., G*Power) to validate our estimates.

9 Sample

Where and from whom will data be collected? How will you decide when to stop collecting data (e.g., target sample size based on power analysis or accuracy in parameter estimation, set amount of time)? If you plan to look at the data using sequential analysis, describe that here.

The sample is collected from an existing published SLR by Coopamootoo and Groß [1]. The sample size is thereby fixed *a priori*. Of the 146 paper selected in Coopamootoo's and Groß' SLR, we will focus on the sub-set of papers that have complete reporting for their tests (according to statcheck [3] analysis, will refine to the statistical inferences that are reported as complete. We will further focus on the subset containing one-to-one comparison (i) t-tests, (ii) F-tests, (iii) χ^2 tests, (iv) r.

Sample Size for Regressions. Of course, sample size is limited by how many tests are completely reported and how many fulfil their assumptions to vouch for a sound effect size estimation.

However, let us consider the *a priori* power requirements for a linear ordinary-least-square (OLS) multiple regression with four predictors (Year, Venue, Platform, and Study; the latter for a mixed-methods account of repeated samples of the same study).

To reach $1 - \beta = .80$ power at a significance level of $\alpha = .05$ and an effect size of $f^2 = .15$, we would need a sample size of $N = 85$. To reach $1 - \beta = .95$ power with the same setup, we would need $N = 129$ as sample size.

Groß [4] reported a final sample size of 114 papers, with 252 correctly parsed test triplets (34 containing an error, 10 containing an decision error). Hence, depending on the required exclusions for dependent sample statistics and failed assumptions, the required sample size is still achievable.

10 Exclusion Criteria

Who will be excluded (e.g., outliers, participant who fail manipulation check, demographic exclusions)? Will they be replaced by other participants?

We will not exclude outliers (as we are reporting standardized effect sizes). We will exclude observations which violate assumptions or provide incomplete reporting to extract meaningful effect sizes.

The most predominant case, therein, will the in face of dependent-sample scenarios, in which we will exclude tests in which investigators have used independent-sample statistics when dependent-sample statistics would have been in order (violating the independent-observation assumption) as well as correctly conducted dependent-sample tests for which the correlation between within-subject groups is not reported (because it is needed for meaningful effect-size estimation).

11 Exception Handling

Should exceptions from the planned study occur (e.g., unexpected effects observed), how will they be handled?

Exceptions will be explicitly declared and considered exploratory.

12 Sign-Off

Pre-registration written by: T.G.
Pre-registration reviewed by: T.G.

Change Management

2020-11-26: The pre-registration was amended with author disclosure and project acknowledgment.

Acknowledgment

The work was in parts supported by the UK Research Institute in the Science of Cyber Security (RISCS) under a National Cyber Security Centre (NCSC) grant on ”Pathways to Enhancing Evidence-Based Research Methods for Cyber Security.” The author was supported by ERC Starting Grant CASCAde (GA n$^{\text{o}}$716980).

References

[1] K. Coopamootoo and T. Groß. Systematic evaluation for evidence-based methods in cyber security. Technical Report TR-1528, Newcastle University, 2017.

[2] K. P. Coopamootoo and T. Groß. A codebook for experimental research: The nifty nine indicators v1.0. Technical Report TR-1514, Newcastle University, November 2017.

[3] S. Epskamp and M. Nuijten. statcheck: Extract statistics from articles and recompute *p*-values (R package version 1.0.0.). `https://cran.r-project.org/web/packages/statcheck/index.html`, 2014.

[4] T. Groß. Fidelity of statistical reporting in 10 years of cyber security user studies. In *Proceedings of the 9th International Workshop on Socio-Technical Aspects in Security (STAST’2019)*, volume 11739 of *LNCS*, pages 1–24. Springer Verlag, 2019.

[5] T. Groß. Fidelity of statistical reporting in 10 years of cyber security user studies [extended version]. arXiv Report arXiv:2004.06672, Newcastle University, 2020.

[6] S. H. Halpern and M. J. Douglas. Jadad scale for reporting randomized controlled trials. *Evidence-based Obstetric Anesthesia*, 2005.

[7] J. P. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.

[8] A. R. Jadad, R. A. Moore, D. Carroll, C. Jenkinson, D. J. M. Reynolds, D. J. Gavaghan, and H. J. McQuay. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled clinical trials*, 17(1):1–12, 1996.