Pre-Registration
Statistical Reporting in 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].

[*]Open Science Framework: `https://osf.io/549qn/`
[1]`https://osf.io`
[2]`https://aspredicted.org`

Context of this Pre-Registration

Meta-Data of Pre-Registration.

- Open Science Framework Repository: `https://osf.io/549qn/`
- Registered Registration File: `https://osf.io/54xpt/—Prereg_StatCheck_Cyber_Security_User_Studies.pdf`
- Timestamp: 2018-09-20 05:45 PM
- Archived Immutable Pre-Registration: `https://osf.io/yqs7w`
- Timestamp: 2019-02-05 5:28 PM

Peer-Reviewed Publication. The definitive version of the study is published as:
Thomas 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), LNCS 11739, Springer Verlag, 2020, pp. 1–24.

ArXiv Report. Thomas Groß. Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies. arXiv:2004.06672, 2020. `https://arxiv.org/abs/2004.06672`

1 Structured Abstract

Background. User studies in cyber security in the widest sense often rely on statistical inference to

show that effects seen in (quasi-)experiments and surveys are significant, and that the null hypothesis can be rejected. In such Null Hypothesis Significance Testing (NHST), the community relies on sound reporting to ascertain the credibility of the results.

Aim. We investigate the prevalence of statistical misreporting as well as the relation to publication venue and year.

Method. Based on a systematic literature review of user studies in cyber security from selected venues in the 10 years 2006–2016, we will evaluate that prevalence of statistical misreporting using the R package statcheck. We will offer a systematic quantification of insufficient reporting, reporting inconsistencies and decision errors. We further conduct correlational ordinal/multinomial logistic regressions to establish the relation to publication venue and year.

Anticipated Results. We anticipate descriptive statistics and graphs of the prevalence of statistical misreporting. We further intend to obtain logistic-regression models on the relation of predictors Venue and Year on coded statcheck outcomes in an correlational study.

Anticipated Conclusions. We anticipate a systematic overview of statistical misreporting over time and venues, yielding an evidence-based estimate how how we are performing as a field as well as what our trajectory is.

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 analysis conducted here is based on a Systematic Literature Review (SLR) conducted by another research project. While the SLR sample has already been collected, manual coding been done on it and descriptive statistics computed, no inferential statistics have been computed to date, to the best of our knowledge. At the time of this pre-registration, statcheck data on the SLR sample has already been collected, but again no inferential statistics evaluated.

3 Aims

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

Outcome Definition.

Definition 1 (SC Outcome Categories)**.** We define the outcome categories of statcheck as follows:

1. CorrectNHST: The NHSTs are reported correctly throughout, where "correctly" is defined as matching triplet of test statistic, degrees of freedom and corresponding p-value.
2. Inconsistency: There exists an inconsistency in any reported triplet (test statistic, df, p-value).
3. DecisionError: There exists a gross inconsistency in any reported triplet (test statistic, df, p-value), in which a re-computed p-value leads to a different decision on rejecting the null hypothesis.
4. Unparseable: There are p-values reported, however sufficient data for a correct triplet missing (test statistic, df, p-value). That is, in most cases that test statistic and degrees of freedom were not reported at all and the consistency cannot be verified.

Descriptives. We will analyze and visualize the prevalence of statistical misreporting along the following lines.

RQ 1 (Prevalence)**.** *How many papers report on Null Hypothesis Significance Testing (NHST) and fall into one of the defined SC outcome categories 1. CorrectNHST, 2. Inconsistency, 3. DecisionError, 4. Unparseable.*

Comparison to Other Fields. We intend to compare the statcheck results in this field with analyses that have been conducted in other fields that seem related. We are most interested in fields at the intersection of human behavior and technology, such as HCI. Granted that statcheck surveys have not been that widely conducted yet, we consider the *Journal of Media Psychology (JMP)* [3] as a primary candidate.

RQ 2 (Comparison)**.** *To what extent do the statcheck SCOutcomes differ between our sample in this field and a comparable field in psychology?*

We then have a multinomial/ordinal test of independence under consideration, for which we would consider the comparison field results as expected distribution.
$H_{C,0}$: The distribution of the SCOutcomes in cyber security user studies is the same as the distribution in the comparison field.
$H_{C,1}$: There is a systematic difference of SCOutcome in cyber security user studies to the comparison field.

Statistical Model on Venue&Year. We establish a statistical model on in correlational study on the following question:

RQ 3 (Influence of Venue and Year)**.** *Considering outcome categories SCOutcome from Def. 1 as response variable, what is the influence of predictors publication Venue and Year.*

In that analysis, we consider the following statistical hypotheses:

1. $H_{V,0}$: There is no influence of the publication Venue on the occurrence of the statcheck outcome SCOutcome.
   $H_{V,1}$: There is a systematic influence of the publication Venue on the occurrence of the statcheck outcome SCOutcome.
2. $H_{Y,0}$: There is no influence of the publication Year on the occurrence of the statcheck outcome SCOutcome.
   $H_{Y,1}$: There is a systematic influence of the publication Year on the occurrence of the statcheck outcome SCOutcome.

Exploratory Model on Authorship&Institution. It stands to reason the the authors present on an paper have an influence on the research methodology and statistical reporting used. Similarly, presumably with more variability, we would expect that the institution of the authors predicts the outcomes on statistical reporting. We imagine that the presence of an author or institution with strong affiliation with sound research methods and reporting standards would influence the entire author cohort of a paper to go "the whole nine yards."

RQ 4 (Influence of Author and Institution)**.** *To what extent is there a systematic influence of authors and institutions (principal components) on SCOutcome?*

4 Methods

Give a brief overview of the methods used.

Data Acquisition. We take as input a sample established in an existing Systematic Literature Review (SLR) by Coopamootoo and Groß on user studies in security and privacy, drawn from selected venues and published in the years 2006–2016. The SLR contained a coding of "Completeness Indicators" according to a defined codebook [1], which the authors elaborated on in an explanation of said indicators [2].

We then conduct a dual analysis with the R package statcheck, considering all p-values reported vis-à-vis of the test statistics that could be parsed in standard format. Merging these two analyses, we obtain the response categories established in Def. 1.

Prevalence. For the descriptives, we will offer descriptive statistics as well as informative plots, such as hierarchical waffle plots on the prevalence of papers fulfilling the outcome categories in Def. 1.

As part of the prevalence discussion, we intend to compare the results in this field with related fields in psychology (RQ 2). In terms of methods, we consider a multinomial contingency table computing a χ^2 of independence to establish the difference of distributions.

Statistical Model on Venue&Year. For the statistical model on RQ 3, we use ordinal and multinomial logistic regression to relate the predictors Venue and Year to the response variable SCOutcome. For an ordinal logistic regression, we assume an ordering of 1. CorrectNHST, 2. Inconsistency, 3. DecisionError, 4. Unparseable. That is, we explicitly declare a unparseable test statistic (or a *p*-value reported without test statistic) as worse than decision errors and mere inconsistencies.

Exploratory Analysis of Author&Institution. For the exploratory model on RQ 4, we can consider the presence or absence of an author or institution as a binomial predictor of the SCOutcome, in principle. Ignoring the order of authorship, we can code the presence/absence of each author/institution as a binomial variable.

However, this modelling will create a logistic regression model with a manifold of predictors (est. 50-80) with few observations ($N = 106$), which would render a logistic regression model unreliable.

While this situation calls for a *dimensionality reduction* on the author/institution variables, we would need to concede that the predictor variables are binomial and not linearly related. Hence, the dataset does not fulfil the assumption of typical techniques such as principal component analysis (PCA) or exploratory factor analysis (EFA). Instead, we will consider non-linear dimensionality-reduction techniques, such as kernel or binary PCA. We intend to use the R packages KernelLab and LogisticPCA as main tools for dimensionality reduction as well as nFactors to support the selection of factors.

Naturally, the staged analysis and choices made on the accepted number of principal components yields considerable researcher degrees of freedom. Specifically, we have at least the following degrees of freedom:

1. Inclusion of institution coding or not (Added information? Variance inflation?),
2. choice of kernel or logistic PCA,
3. choice of kernel type and parameters in the case of the kernel PCA,
4. choice of PC selection method (e.g., based on Cattell-Nelson-Gorsuch (CNG) Scree test of ordered eigenvalues or proportion of accumulated variance),
5. choice of number of selected principal components,
6. choice of selected logistic regression model (based on principal components as predictors).

To mitigate those degrees of freedom, we will strictly conduct the analysis in stages, that is, first obtain sound PCs in decreasing captured variance based on a Scree plot. Once, this decision is made and committed to OSF, we will continue to use the PCs as predictors in a subsequent logistic regression.

Given the researcher degrees of freedom present in this analysis, we declare the analysis *exploratory*.

5 Independent Variables (IVs)

Describe the conditions (for an experimental study) or predictor variables (for a correlational study).

Venue&Year. As primary predictors, we consider publication Venue (nominal) and Year (interval).

We encode these variables with the following levels:

1. Venue: Unordered factor with the levels: 1. CCS, 2. LASER, 3. PETS, 4. S&P, 5. SOUPS, 6. TDSC, 7. TISSEC, 8. USEC, 9. USENIX, 10. WEIS.
2. Year: Ordered factor with the levels: 1. 2006, 2. 2007, 3. 2008, 4. 2009, 5. 2010, 6. 2011, 7. 2012, 8. 2013, 9. 2014, 10. 2015, 11. 2016.

Author&Institution. For the exploratory analyses we have authors and institutions encoded in binomial form.

For each unique author or institution, we create a binary variable with the following encoded levels:

1. *variable* = 0: The corresponding author/institution is absent on the author list of the respective paper.
2. *variable* = 1: The corresponding author/institution is present (explicitly named) on the author list of the respective paper.

Note that the order of authors will not be encoded.

6 Dependent Variables (DVs)

Dependent variables: Describe the key dependent variable(s) specifying how they will be measured.

As primary dependent variable, we consider the outcome of a statcheck analysis per paper, defined as ordinal/multinomial variable SCOutcome defined in Def. 1.

Hence, we have an ordinal variable with four levels: 1. CorrectNHST, 2. Inconsistency, 3. DecisionError, 4. Unparseable.

7 Mediator Variables

Describe any variables you expect to mediate the relationship between your IV's and DV. Specify how they will be measured.

N/A

8 Moderator Variables

Describe any variables you expect to moderate the relationship between your IV's and DV. Specify how they will be measured.

N/A

9 Data Preparation

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

We will check the data for consistency, that is, correct recording of Venue and Year.

For the statcheck results, we will cross-check reported inconsistencies and decision errors manually to ascertain that they are indeed genuine and not an artefact of the analysis. For the analysis, we will configure statcheck not to consider reported p-values of .000 as an error even if they are technically impossible and should be reported as $p < .001$. We will check whether inconsistencies came to pass through unrecognized one-tailed tests.

In cases of statcheck results not being able to corroborated with a manual analysis, we will correct the dataset based on the manual analysis and document the change made.

10 Main Analyses

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

We are computing a χ^2 test on the multinomial comparison of the SCOutcome distribution with the chosen comparison field (e.g., publications at JMP).

The primary analysis is an ordinal logistic regression:

$$\mathsf{SCOutcome} \sim \mathsf{Venue} + \mathsf{Year}.$$

11 Secondary Analyses

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

Multinomial Logistic Regression. We conduct as secondary analysis a multinomial logistic regression:

$$\mathsf{SCOutcome} \sim \mathsf{Venue} + \mathsf{Year}.$$

That is, as secondary analysis we drop the order assumption placed on the response variable.

Correlational Influence of Authors&Institutions. We consider an analysis pipeline with a nonlinear (kernel/binary) Principal Component Analysis (PCA) for dimensionality reduction applied to the binomial encodings of authors and/or institutions.

The resulting Principal Components (PCs) are then used as predictors in a ordinal logistic regression on SCOutcome.

Exploratory Analysis on Completeness Indicators. The SLR conducted by Coopamootoo and Groß contained a manual coding of completeness indicators CICode[1..9] on the reporting in the papers in a trinomial/ordinal form: success, partial, and failure.

Naturally the statistical analysis model on SCOutcome (with predictors Venue and Year or authors/institutions) would also equally applicable to completeness indicators as response variables.

Then, the collection of response variables CICode[1..9] replaces the statcheck outcome SCOutcome as response variables in ordinal/multinomial logistic regressions.

12 Validation

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

Regression Diagnostics. We will conduct regression model diagnostics, including distribution of residuals, variance inflation, etc. Largely, we will use the R-package car by John Fox for regression diagnostics.

We will consult the recommendations of Hosmer and Lemeshow *Applied Logistic Regression* [4, Chapter 5] for ordinal/multinomial logistic regressions.

Model Fit. We will consider AIC of the selected model vis-à-vis of the minimal model as well as test Hosmer-Lemeshow to reject the fit.

Accuracy Validation. We intend to compute a 10-fold cross-validation on the same dataset.

13 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 was collected by a different project, which conducted a Systematic Literature Review (SLR) on user studies in security and privacy, from selected venues and published in the years 2006–2016. From this SLR, we obtain the PDFs of the 146 research papers named in the technical report on the SLR.

After exclusion of papers not reporting NHST, we retain $N = 106$ papers as final sample for the statcheck analysis.

For the corresponding logistic regressions, we conducted an *a priori* power analysis. We use G*Power's analysis for logistic regressions as an estimate. Given that the sample size is already fixed because of the SLR sample as input, we conducted a *sensitivity analysis*. With an estimated likelihood of $\mathsf{P}[X = 1|Y = 1]\ \mathsf{H}_0 = .2$, and an estimated R^2 of other X of 0.1, we obtain a sensitivity of $OR = 0.516$ at 80% power.

Note that power analysis for ordinal and multinomial logistic regressions is a complex subject and a matter for discussion in the field. Also, there is little *a priori* information on the H_0 likelihood and R^2 of other variables, rendering the estimates made part of the researcher's degrees of freedom. The G*Power analysis, thereby, only offers a very rough guidance.

14 Exclusion Criteria

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

We are excluding publications that do not conduct Null Hypothesis Significance Testing (NHST), roughly speaking papers that do not contain p-values. This entails an exclusion of position papers and qualitative studies.

15 Exception Handling

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

Exceptions from the study protocol will be recorded explicitly. Unexpected effects observed will be declared as exploratory.

16 Sign-Off

Pre-registration written by (initials): T.G.
Pre-registration reviewed by (initials): T.G.

Change Management

2020-07-27: The pre-registration was amended with author disclosure and project acknowledgment.

2020-11-26: Added arXiv report.

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°716980).

References

[1] 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.

[2] K. P. Coopamootoo and T. Groß. Cyber security and privacy experiments: A design and reporting toolkit. In *IFIP International Summer School on Privacy and Identity Management*, pages 243–262. Springer, 2017.

[3] M. Elson and A. K. Przybylski. The science of technology and human behavior – standards old and new. *Journal of Media Psychology*, 29(1):1–7, 2017.

[4] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.