Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies

Thomas Groß(✉)

Newcastle University, Newcastle upon Tyne, UK
thomas.gross@newcastle.ac.uk

Abstract. Studies in socio-technical aspects of security often rely on user studies and statistical inferences on investigated relations to make their case. They, thereby, enable practitioners and scientists alike to judge on the validity and reliability of the research undertaken.

To ascertain this capacity, we investigated the reporting fidelity of security user studies.

Based on a systematic literature review of 114 user studies in cyber security from selected venues in the 10 years 2006–2016, we evaluated fidelity of the reporting of 1775 statistical inferences using the R package statcheck. We conducted a systematic classification of incomplete reporting, reporting inconsistencies and decision errors, leading to multinomial logistic regression (MLR) on the impact of publication venue/year as well as a comparison to a compatible field of psychology.

We found that half the cyber security user studies considered reported incomplete results, in stark difference to comparable results in a field of psychology. Our MLR on analysis outcomes yielded a slight increase of likelihood of incomplete tests over time, while SOUPS yielded a few percent greater likelihood to report statistics correctly than other venues.

In this study, we offer the first fully quantitative analysis of the state-of-play of socio-technical studies in security. While we highlight the impact and prevalence of incomplete reporting, we also offer fine-grained diagnostics and recommendations on how to respond to the situation.

Keywords: User studies · SLR · Cyber security · Statistical reporting

1 Introduction

Statistical inference is the predominant method to ascertain that effects observed in socio-technical aspects of security are no mere random flukes, but considered to be "the real McCoy."

In general, statistical inference sets out to evaluate a statistical hypothesis stated *a priori*. It employs observations made in studies to establish the likelihood as extreme as or more extreme than the observations made, assuming the statistical hypothesis not to be true. This likelihood is colloquially referred to as

Preregistered at the Open Science Framework: osf.io/549qn/.

T. Groß and T. Tryfonas (Eds.): STAST 2019, LNCS 11739, pp. 3–26, 2021.
https://doi.org/10.1007/978-3-030-55958-8_1

a p-value. Alternatively to Null Hypothesis Significance Testing (NHST)—and often used complementarily—studies may estimate the magnitude of effects in reality and confidence intervals thereon [5].

The onus of proof is generally on the authors of a study. There are numerous factors influencing whether a study's results can be trusted—a) sound research questions and hypotheses, b) vetted and reliable constructs and instruments, c) documentation favoring reproducibility, d) sound experiment design, yielding internal and external validity, e) randomization and blinding, f) systematic structured and standardized reporting—in the end, it is the outcomes of the statistical inference that often render a final verdict.

These outcomes do not only indicate whether an effect is likely present in reality or not. They also yield what magnitude the effect is estimated at. Thereby, they are the raw ingredient for (i) establishing whether an effect is practically relevant, (ii) evaluating its potential for reuse, and (iii) including it further quantitative research synthesis.

While there have been a number of publications in socio-technical aspects of security offering guidance to the community to that end [2,4,11,15,17] as well as proposals in other communities [1,10,12], the evidence of the state-of-play of the field has been largely anecdotal [17] or in human-coded analysis [3]. While this field is arguably quite young, we argue that it would benefit greatly from attention to statistical reporting, from attaining fault tolerance through reporting fidelity and from preparing for research synthesis (cf. Sect. 2.1).

In this study, we aim at systematically evaluating the fidelity of statistical reporting in socio-technical aspects of security. We analyze (i) whether statistical inferences are fault-tolerant, in the sense of their internal consistency being publicly verifiable, and (ii) whether the reported p-values are correct. Through the semi-automated empirical analysis of 114 publications in the field from 2006–2016, we offer a wealth of information including meta-aspects. We compare statistical reporting fidelity of this field with a related field of psychology as well as analyze the trajectory of the field, that is, the trends found over time. We substantiate the these results with qualitative coding of errors observed to elucidate what to watch out for.

Contributions. We are the first to subject our own field to a systematic empirical analysis of statistical reporting fidelity. In that, we offer a well-founded introspection in the field of socio-technical aspects of security that can serve program committees and authors alike to inform their research practice.

2 Background

2.1 Importance and Impact of Statistical Reporting

Null Hypothesis Significance Testing (NHST) establishes statistical inference by stating *a priori* statistical hypotheses, which are then tested based on observations made in studies. Such statistical inference results in a p-value, which gives the conditional probability of finding data as extreme as or more extreme than the observations made, assuming the null hypothesis being true. Many fields combine NHST with point and interval estimation, that is, establishing an estimate of the magnitude of the effect in the population and the confidence interval thereon.

Table 1. Degrees of fidelity in statistical reporting for the same two-tailed independent-samples t-test on a relation with a large effect size (ES). *Note:* ○ = impossible ◑ = can be estimated ● = supported

	Incomplete triplet		Complete triplet	
	Sig.	p-Value	ES inferrable	ES explicit
Example	$p < .05$	$p = .019$	$t(24) = 2.52$, $p = .019$	$t(24) = 2.52, p = .019$, Hedges' $g = 0.96$, CI $[0.14, 1.76]$
p quantifiable	○	●	●	●
Cross-checkable	○	○	●	●
ES quantifiable	○	○	◑	●
Synthesizable	○	○	◑	●

Reporting Fidelity and Fault Tolerance. Different reporting practices yield different degrees of information and fidelity. It goes without saying that a simple comparison with the significance level α, e.g., by stating that $p < .05$, yields the least information and the least fidelity. Reporting the actual p-value observed offers more information as well as a means to quantify the likelihood of the effect.

To gain further reporting fidelity and fault tolerance, one would not only report the exact p-value, but also the chosen test parameters (e.g., independent-samples or one-tailed), the test statistic itself (e.g., the t-value) and the degrees of freedom (df) of the test. We, then, obtain a consistent triplet (test statistic, df, p-value) along with the test parameters. Table 1 exemplifies degrees of fidelity.

The upshot of a diligent reporting procedure including full triplets is that it enables cross-checks on their internal consistency and, thereby, a degree of fault tolerance. Vice versa, if only the p-value or a comparison with a significance level is reported, the capacity to validate inferences is impaired.

Impact on Research Synthesis. Published studies usually do not stand on their own. To learn what relations are actually true in reality and to what degree, we commonly need to synthesize the results of multiple studies investigating the same relations. More mature fields (such as evidence-based medicine or psychology) engage in systematic reviews and meta analyses to that end.

For these down-stream analyses to be viable, the original studies need to contain sufficient data for subsequent meta-analyses. If the original studies omit the actual test statistics and degrees of freedom, the synthesis in meta analyses is hamstringed or rendered impossible altogether.

2.2 Reporting and Methodology Guidelines

Reporting fidelity is usually one of the goals of reporting standards. Given that the field of socio-technical research in cyber security is a young and does not have its own established reporting standards, it is worthwhile to consider ones of other fields. Psychology seems a sound candidate to consider as a guiding example in this study. Other fields, such as behavioral economics, are equally viable.

The publication guidelines of the American Psychology Association (APA) [1] require that inferences are reported with their full test statistics and degrees of freedom. Exact p-values are preferred. The APA guidelines require to report appropriate effect sizes and their confidence intervals.

Of course, there are also methodological guidelines that go far beyond reporting statistical tests. For instance, the CONSORT guidelines [12] cover reporting for randomized trials. Furthermore, recently LeBel et al. [10] proposed a unified framework to quantify and support credibility of scientific findings.

Even though socio-technical aspects of security is a young field, there have been initiatives to advance research methodology, considered in chronological order: (i) In 2007, Peisert and Bishop [15] offered a short position paper scientific design of security experiments. (ii) Maxion [11] focused on making experiments dependable, focusing on the hallmarks of good experiments with an eye on validity. (iii) In 2013, Schechter [17] considered common pitfalls seen in SOUPS submissions and made recommendation on avoiding them, incl. statistical reporting and multiple-comparison corrections. (iv) Coopamootoo and Groß proposed an introduction for evidence-based methods [4], incl. sound statistical inference and structured reporting. (v) The same authors published an experiment design and reporting toolset [2], considering nine areas with reporting considerations, incl. test statistics and effect sizes.

2.3 Analysis of Statistical Reporting

We analyze statistical reporting of publications with the R package statcheck [7]. The statcheck tool extracts Strings of the form $ts(df) = x, p \text{ op } y$, where ts is the test statistic, df the degrees of freedom, and op a infix relation, such as, $<$. It recognizes t, F, r, χ^2, and z as test statistics and recomputes the corresponding p-values from them. It, hence, enables a consistency check of reported triplets of test statistic, degrees of freedom and p-values.

In this analysis, statcheck recognizes one-tailed tests to some extent from searching keywords and computing if a test were valid if considered one-tailed. It adheres to the rounding guidelines of the American Psychology Association (APA) [1]. Nuijten et al. [13] concede that statcheck does not recognize p-values adjusted for multiple-comparison corrections.

While the creators of statcheck have argued for its validity and reliability [13,14], the tool faced scrutiny and controversy [18] over its false positive and false negative rates. Schmidt [18], for example, criticized that statcheck's inability to recognize corrected p-values, such as from Greenhouse-Geisser corrections. Lakens [9] found reported errors typically to be minor.

For this study, we prepare to mitigate possible statcheck mis-classifications by manually checking and coding its outcomes.

2.4 Related Works

In 2016/17 Coopamootoo and Groß [3] conducted a Systematic Literature Review (SLR) on cyber security user studies published in the years between 2006–2016. This research was first presented at a 2017 community meeting of the UK Research Institute in the Science of Cyber Security (RISCS). Their study contained three parts: (i) the SLR itself, yielding a sample of 146 cyber security papers, (ii) a qualitative coding of nine "completeness indicators," based on an *a priori* codebook. (iii) a quantitative analysis on a sub-sample using parametric tests on differences between means (e.g., t-tests).

While this study uses the same set of papers as a sample to enable a comparison of results, this study takes an entirely different approach to the analysis: (i) Instead of manual coding of reporting completeness, we focus on the automated analysis reporting fidelity on extracted p-values, (ii) we evaluate quantitative properties on inconsistencies and decision errors of a large part of the sample, and (iii) we obtain a fine-grained understanding of "things going wrong" through grounded coding,

3 Aims

We define the classes of statcheck outcomes for test statistics and papers.

Definition 1 (SC Outcome Categories)

Individual Tests: *SCOutcome has the following cases for individual tests:*

1. *CorrectNHST: The NHST is reported with its test statistic triplet. The given triplet is correct, where "correct" is defined as matching triplet of test statistic, degrees of freedom and corresponding re-computed p-value.*
2. *Inconsistency: The reported triplet (test statistic, df, p-value) is inconsistent.*
3. *DecisionError: The reported triplet (test statistic, df, p-value) is grossly inconsistent, that is, the re-computed p-value leads to a different decision on rejecting the null hypothesis.*
4. *Incomplete: A p-values is reported without sufficient data for an evaluation of the triplet (test statistic, df, p-value).*

Entire Papers: *SCOutcome has the following cases for aggregated over papers:*

1. *CorrectNHST: There exist one or more NHSTs reported with correct test statistic triplets. The given complete triplets are correct throughout, where "correct" is defined as matching triplet of test statistic, degrees of freedom and corresponding re-computed p-value. A paper can be classified as CorrectNHST even if there exist incomplete test statistics.*

2. *Inconsistency: There exists an inconsistent 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. *Incomplete: For all p-values reported, it holds that there is insufficient data for a correct triplet (test statistic, df, p-value). For a paper classified as Incomplete, there is not a single p-value with complete test statistic found.*

We call Complete the complement of Incomplete.

RQ 1 (Prevalence). *How many papers report on Null Hypothesis Significance Testing (NHST) and fall into one of the defined SC outcome categories according to Definition 1 1. CorrectNHST, 2. Inconsistency, 3. DecisionError, 4. Incomplete. Which papers use 1. multiple-comparison corrections (MCC), 2. effect sizes.*

While we originally investigated the use of Amazon Mechanical Turk (AMT) and similar recruiting services, we have declared this aim out of scope for this publication. MCCs and effect sizes are also relevant in relation to power and Positive Predictive Value (PPV) of the studies in question, however, we will consider these inquiries in future work.

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)* [6] as a primary candidate. This choice is made because of similarities

(i) media psychology is concerned with human subjects and socio-technical aspects,
(ii) media psychology includes topics that might also have been published in user studies in cyber security, such as adversarial behavior (e.g., violence) vis-à-vis of HCI, cyber bullying, behavior on social media,
(iii) media psychology is a relatively young field, JMP having been founded in 1989 and gained its current name 2008.

The distinct difference we are interested in is that JMP is subject to reporting standards (APA). We note that the selection of JMP as comparison sample may be controversial and that—at the same time—comparisons to further fields are easily done, yet out of the scope for this study.

RQ 2 (Comparison). *To what extent do the statcheck SCOutcomes differ between our sample in this field and a comparable field in psychology?*
$\mathsf{H}_{\mathsf{C},0}$*: The distribution of the SCOutcomes in cyber security user studies is the same as the distribution in the comparison field.* $\mathsf{H}_{\mathsf{C},1}$*: There is a systematic difference of SCOutcome in cyber security user studies to the comparison field.*

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

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

As an exploratory inquiry, we employ the statcheck analysis to the submissions of STAST 2019, testing its usefulness in supporting PC members.

4 Method

The study has been pre-registered at the Open Science Framework (OSF)[1], which also contains Online Supplementary Materials, such as a summary of the SLR specification and the sample itself. All analyses, graphs and tables are computed directly from the data with the R package knitr, where the statcheck output was cached in csv files.

All statistical tests are computed at a significance level of $\alpha = .05$. The Fisher Exact Tests (FETs) for cases with low expected cell frequency are computed with simulated p-values with 10^5 replicates.

4.1 Ethics

This study followed the guidelines of the ethical boards of its institution. While we make the entire list of analyzed papers available for reproducibility, we decided not to single out individual papers. We are aware that the descriptive statistics presented allow making a link to the respective papers; we accept that residual privacy risk. *Full disclosure:* one of the sample's papers belongs to the author of this study; statcheck flagged it.

4.2 Sample

The target population of this study was cyber security user studies. The sampling frame for this study is derived from a 2016/17 Systematic Literature Review (SLR) conducted by Coopamootoo and Groß [3] whose results were first published at a 2017 Community Meeting of the Research Institute in the Science of Cyber Security (RISCS). This source SLR's search, inclusion and exclusion criteria are reported in Online Supplementary Materials.

We have chosen this sample to gain comparability to earlier qualitative and quantitative analyses on it [3]. This sample restricts the venues considered to retain statistical power for a regression analysis. We stress that the automated the analysis methodology can be easily applied to other samples.

[1] osf.io/549qn/.

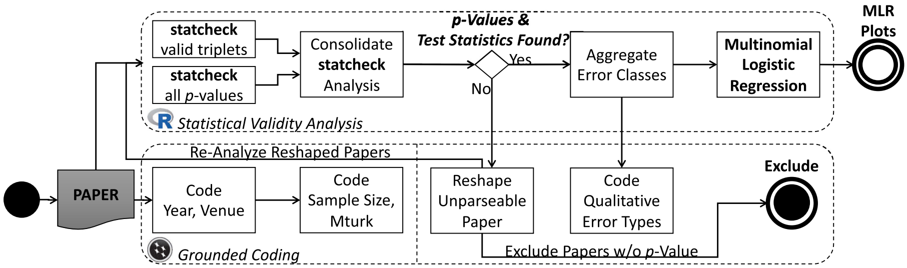


Fig. 1. Flow chart of the study's procedure with two interlinked analyses.

4.3 Procedure

Our procedure, as depicted in Fig. 1, constituted a mixed-methods approach that fusing two interlinked analysis processes: (i) Statistical Validity Analysis and (ii) Grounded Coding of paper properties and errors detected. Our analysis script received as input the PDFs of studies included from the source SLR.

Statistical Validity Analysis. We computed two iterations of statcheck, one only considering statistical statements in standard format and one including all p-values found. The statcheck results were subjected to a manual cross-check, possibly resulting the reshaping of papers that statcheck could not parse out of the box. Subsequently, we merged the results of both analyses and aggregated their events (counting number of correct tests, inconsistencies, decision errors and p-values without parseable test statistics). We, thereby, established the dependent variable SCOutput per statistical test and per paper.

Grounded Coding. We coded paper properties in NVivo. We evaluated the statcheck results in a second lane of grounded coding, classifying errors of statcheck as well as errors committed by authors of the papers.

As a part of this analysis, we "reshape" papers that could not be parsed by statcheck for reasons outside of the research aims of this study. For instance, if a paper embedded statistical tests as image rather than text, we would transcribe the images to text and re-run statcheck on the "reshaped" input.

Once these results are coded, we amend the statcheck outcomes recorded in SCOutcome to ensure that this variable reflects an accurate representation of the sample.

4.4 Grounded Coding

Grounded coding refers to the code being grounded in properties found in the data, instead of being based on an *a priori* codebook.

Paper Properties. We conducted a systematic coding in NVivo with the purpose to establish overall properties of all papers. We were extracting especially: (i) sample size, (ii) use of multiple-comparison corrections, and (iii) use of dependent-samples tests.

Analysis Outcomes. After having run statcheck on the sample, we first conducted a grounded coding of statistical tests marked as inconsistency or decision error. We re-computed the p-values from the test statistics ourselves and interpreted the results in the context of the reporting of the paper. We took into account the formulation around the test as well as overall specification of hypotheses, test parameters (e.g., one-tailed) and multiple-comparison corrections. We include the resulting emergent codebook presented in Table 2.

Secondly, we analyzed the outcomes statcheck marked as neither inconsistency nor decision error. For those results, we compared the raw text with statcheck's parsed version as well as recomputed p-value. We ignored small rounding differences as statcheck as authors rounding test statistics for reporting will naturally cause small differences. In cases of a mismatch between raw text and interpretation (e.g., in degrees of freedom accounted for), we re-computed the statistics manually.

Finally, we coded whether a mistake by statcheck would be considered a FalsePositive or FalseNegative. After this evaluation, we adjusted the SCOutcome to ensure that the subsequent analysis is based on a correct representation of the sample.

Table 2. Codebook of the grounded coding of error types.

Errors of statcheck		Errors of authors	
Code	Definition	Code	Definition
scParsedOK	parsed the PDF correctly	Typo	Likely mis-typed
scCorrect	statcheck result validated	RoundingError	incorrect rounding rules
scMisclassified	misclassified test	OneTailedUS	unspecified one-tailed test
scMissedMC	missed multiple-comparison corrections specified paper	Miscalculation	miscalculated the statistics, wrong p-value for statistic

4.5 Evaluation of statcheck

Appendix A contains the details of the corresponding qualitative coding.

Reshaping of Unparseable Papers. There were eight of papers for which statcheck could neither extract p-values nor test statistics due to encoding issues (e.g., embedding statistics as images). For all of those, we recorded them as unparseable, yet transformed them into parseable text files for further analysis.

Table 3. Confusion matrix for statcheck evaluating tests.

Predicted	Reference	
	Positive	Negative
Positive	29	5
Negative	0	218

Accuracy: .98, 95% CI $[.95, .99]$, $Acc > NIR(.88)$, $< .001$***, Sensitivity = 1.00, Specificity = .98, PPV = .85, $F_1 = .92$

Errors Committed by statcheck. Of the total 252 parsed tests, 34 contained an error, 10 of which a decision error. We compared those outcomes against the grounded coding of results and our re-computation of the statistics.

We found that (i) statcheck parsed papers that were correctly reported without fail, (ii) it misclassified two tests, (iii) it detected one-tailed tests largely correctly, (iv) it treated dependent-samples tests correctly, (v) it did not recognize the specified multiple-comparison corrections in three cases. This leaves us with 5 false positives and no false negatives, marked in Sub-Fig. 8a.

Detection Performance of statcheck. For the analysis of complete test triplets, we analyzed the confusion matrix of statcheck results vs. our coding (Table 3). The Positive Predictive Value (PPV) of 85.3% indicates a decent likelihood of a positive statcheck report being true.

4.6 Multinomial Logistic Regression

We conducted multinomial logistic regressions with the R package nnet [16], relying on Fox's work [8] for visualization. The models were null, year-only, venue-only and year and venue combined. The dependent variables was SCOutput. The independent variables were Year (interval) and Venue (factor).

5 Results

5.1 Sample

We have refined the inputted sample of 146 publications by excluding publications that do neither contain empirical data nor significance tests (p-value), retaining 114 publications for further analysis. We illustrate the sample refinement in Table 4. We include the final sample in the Online Supplementary Materials and outline its distribution by publication venue and year in Table 5. We note that the sample is skewed towards SOUPS and more recent publications. We note that the sample was drawn only from 10 specific venues in an effort to retain power in a logistic regression with venue as a categorical factor.

Table 4. Sample refinement and final composition

Phase	Excluded	Retained sample
Source SLR [3] (Google Scholar)	–	1157
Inclusion/exclusion	1011	146
This study		
Studies with empirical data	24	122
Studies with NHST/p-value	8	114 → **Final sample**

Table 5. Sample composition by venue and year.

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Sum
SOUPS	6	3	4	6	8	4	10	8	13	9	6	77
USEC	0	0	0	0	0	0	0	0	4	0	0	4
CCS	0	0	0	0	0	0	0	0	4	1	3	8
USENIX	0	0	0	1	0	0	4	1	1	0	0	7
PETS	1	0	0	0	0	1	1	0	0	1	2	6
TISSEC	0	0	0	0	0	1	0	0	0	0	2	3
LASER	0	0	0	0	0	0	1	0	0	0	1	2
S&P	0	0	0	0	0	0	0	0	1	1	0	2
TDSC	0	1	0	0	0	0	1	0	1	0	1	4
WEIS	0	0	0	0	0	0	0	0	0	1	0	1
Sum	7	4	4	7	8	6	17	9	24	13	15	114

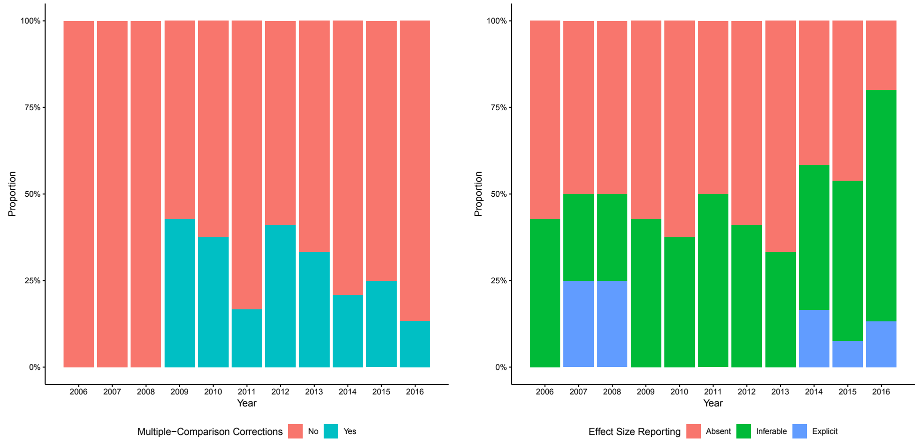
5.2 Exploration of the Distribution

Distribution of Qualitative Properties. We visualize the presence of qualitative properties of papers over time in Fig. 2. We observe (i) Mutliple-Comparison Corrections seeing adoption from 2009 (Fig. 2a), (ii) Effect sizes being on and off over the years (Fig. 2b).

Distribution of p-Values. We analyze the distribution of p-values per paper. Therein we distinguish incomplete and complete triplets including test statistic and degrees of freedom. In Fig. 3, we depict this p-value distribution; Fig. 3a is ordered by number of the tests reported on, distinguishing between complete/incomplete triplets while annotating the presence of multiple comparison corrections (MCC); Fig. 3b is organized by publication year. The included linear regression lines indicate little to no change over time.

5.3 Prevalence of Statistical Misreporting

For RQ1, we compare statistical misreporting by venue and year, considering individual tests as well as entire papers (cf. contingency tables in the Online Supplementary Materials).



(a) Multiple-comparison corrections (MCC)

(b) Effect size reporting

Fig. 2. Properties of SLR papers by year.

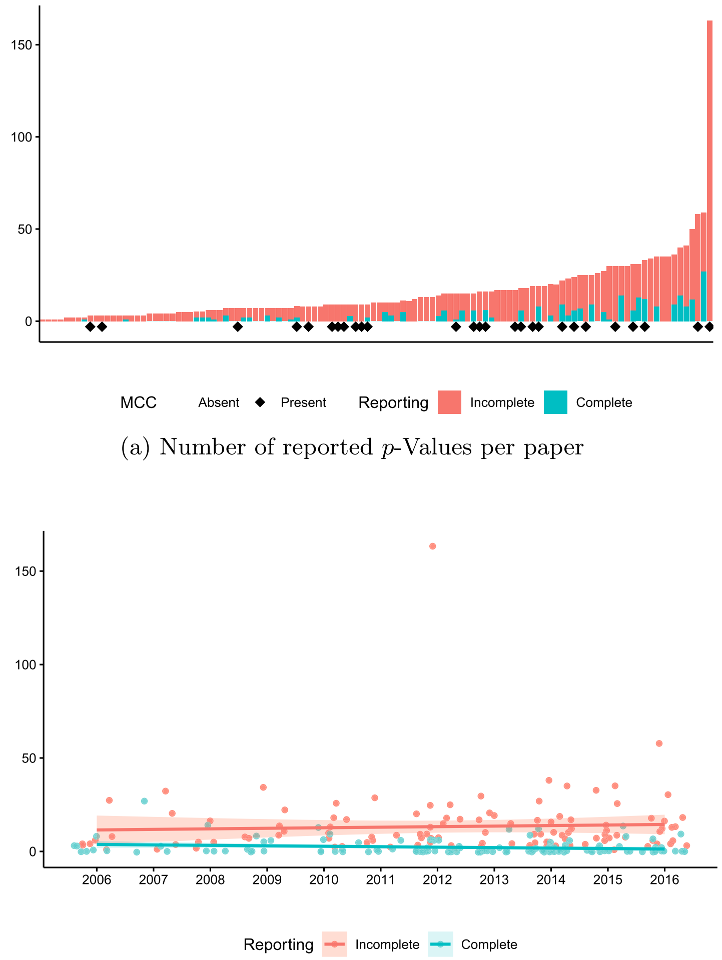
Misreported Tests. For individual tests, there is a statistically significant association between the statcheck outcomes and the publication venue, FET $p = .034$, as well as the publication year, FET $p < .001$. This offers first evidence to reject the null hypotheses $\mathsf{H}_{\mathsf{V},0}$ and $\mathsf{H}_{\mathsf{Y},0}$.

Table 6 contains the corresponding contingency table.

Table 6. Contingency table of individual test statcheck outcomes by venue, FET $p = .034$.

	SOUPS	USEC	CCS	USENIX	PETS	TISSEC	LASER	S&P	TDSC	WEIS
CorrectNHST	170	1	9	4	11	6	5	0	12	0
Inconsistency	19	1	3	0	0	0	1	0	0	0
DecisionError	9	0	0	0	0	0	1	0	0	0
Incomplete	1028	33	122	100	72	71	19	11	60	7

Papers with Misreporting. Sub-Figure 6a on p. 15 shows a hierarchical waffle plot of the statcheck outcomes. For aggregated outcomes per paper displayed in Fig. 4, the associations per venue and year are not statistically significant, FET $p = .963$ and FET $p = .455$ respectively. A likely reason for this result is visible in the histograms of Fig. 5: errors are at times clustered, in that, some papers contain multiple errors.



(a) Number of reported p-Values per paper

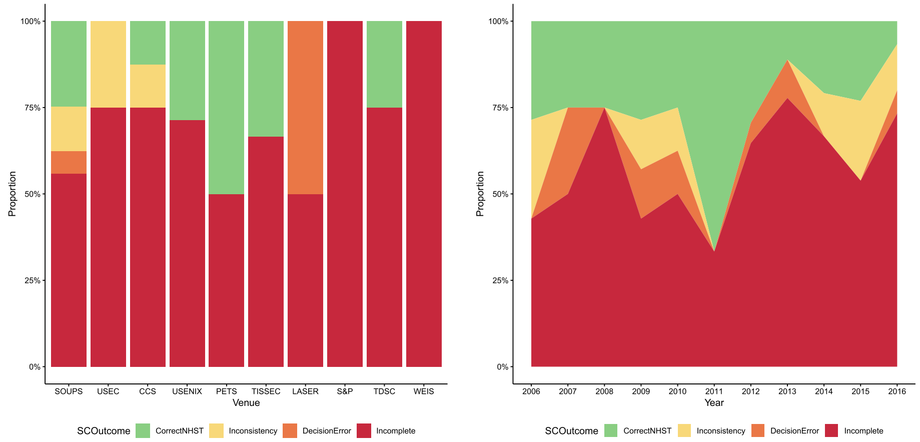
(b) Distribution by Year

Fig. 3. Distribution of statistical reporting of papers, that is, how many p-values per paper are reported Incomplete or Complete. MCC = Multiple-Comparison Corrections.

5.4 Comparison with JMP

With respect to RQ2, the statcheck outcomes of the included SLR and Journal of Media Psychology (JMP) are statistically significantly different, $\chi^2(3) = 88.803, p < .001$. Hence, we reject the null hypothesis $\mathsf{H}_{\mathsf{C},0}$ and conclude that there is a systematic difference between fields. We find an effect of Cramér's $V = 0.646$, 95% CI $[0.503, 0.773]$.

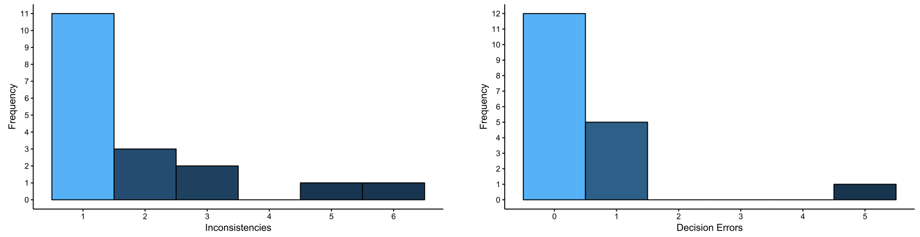
If we restrict the analysis to the papers containing Complete tests and, thereby, exclude papers marked Incomplete, we find that the difference between fields is not statistically significant any longer, $\chi^2(2) = 0.197, p = .906$, Cramér's $V = 0.037$, 95% CI $[0, 0.139]$.



(a) Venues

(b) Years

Fig. 4. Proportions of per-paper aggregated statcheck outcomes by venue and year. The results by year are shown as area plot to highlight development over time.



(a) Frequency of inconsistencies

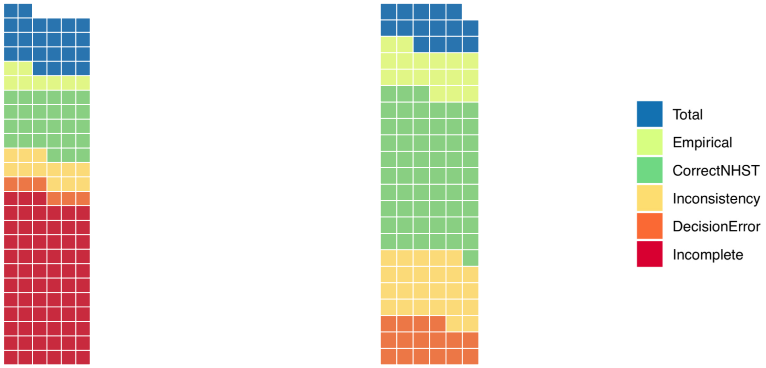
(b) Frequency of decision errors

Fig. 5. Number of errors per paper.

5.5 Reporting Test Outcomes by Venue and Year

While we analyzed tests and aggregated paper SCOutcome by venue and year, we found that these multinomial logistic regressions were not stable. Even if the models were statistically significant, this missing stability was evidenced in extreme odds-ratios, which was likely rooted in the sparsity of the dataset. (We report all MLR conducted in the Online Supplementary Materials for reference). To overcome the sparsity, we chose to collapse the venue factor into SOUPS and OTHER levels, called venue' (and the corresponding null hypothesis $\mathsf{H}_{\mathsf{V}',0}$).

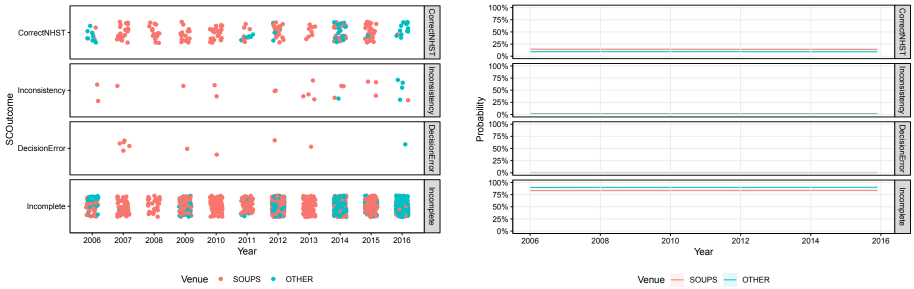
A multinomial logistic regression on individual tests with SCOutcome $\sim$ venue'+year with Incomplete as reference level is statistically significant, LR, $\chi^2(6) = 15.417, p = .017$. Because the model explains McFadden $R^2 = .01$ of the variance, we expect little predictive power.



(a) This Study (SLR)

(b) JMP

Fig. 6. Hierarchical Waffle plots comparing user studies (SLR) in cyber security and the Journal of Media Psychology (JMP) (One square represents one paper).



(a) Scatter plot

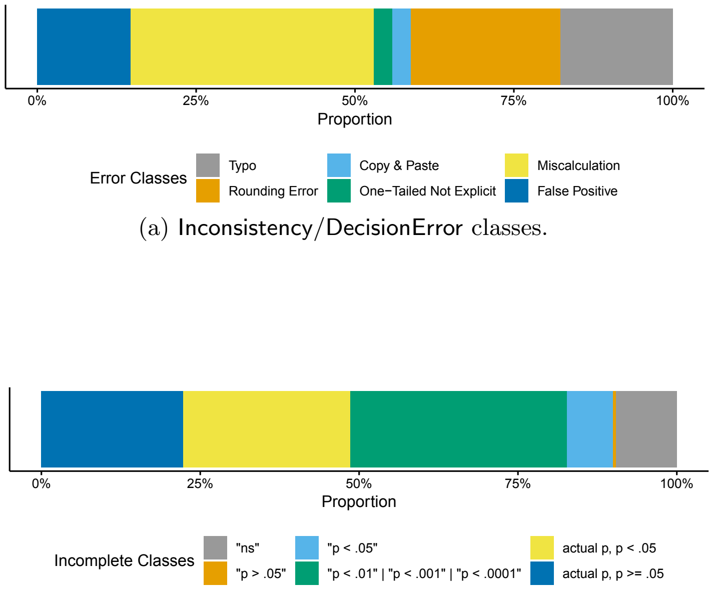
(b) MLR probabilities

Fig. 7. Per-teststatcheck outcomes by venue and year. *Note:* The multinomial logistic regression (MLR) is statistically significant, LR Test, $\chi^2(6) = 15.417, p = .017$.

The corresponding predictors are statistically significant as well. Hence, we reject the null hypotheses $H_{V',0}$ and $H_{Y,0}$. Figure 7 contains an overview of the scatter plot vs. the predicted probabilities from the MLR.

While we find that there is an effect of year in increasing likelihood of Incomplete outcomes, this only accounts for an increase of 0.2% per year, barely perceptible in the graph. Everything else being equal, a transition from venue SOUPS to OTHER yields an increase of likelihood of the Incomplete outcomes, by a factor of roughly 2. However, these changes are dwarfed by the overall intercept of tests being correct (in comparison to Incomplete).

In absolute terms, the expected likelihood of tests being Incomplete is 80%, with OTHER venues having a few percent greater Incomplete likehood. SOUPS exhibits an expected likelihood of 13% of being CorrectNHST, while OTHER venues yield a few percent lower likelihood.



(a) Inconsistency/DecisionError classes.

(b) Incomplete classes.

Fig. 8. Classification of reported statcheck outcomes.

5.6 Qualitative Analysis

We offer a summary of the analysis here, a detailed account is included in Appendix A.

Composition of Incomplete p-Values. Sub-Figure 8b contains an overview of the classes of incompletely reported p-values. Less than half the cases of incomplete triplets contain an actual p-values (half of them, in turn, significant or not significant). 31% of the incomplete cases compared to lower significance bound than $\alpha = .05$. 9% of the tests are simply declared non-significant, another 7% reported as significant wrt. $p < .05$.

Distribution of p-Values. Figure 9 shows the difference between reported and computed p-values. When comparing reported and re-computed p-values, we found that in 22 out of 34 cases, the reported p-value was more significant than the computed one (65%).

5.7 Significance Detection Performance

We analyzed the decision making of authors on statistical significance of reported results vis-à-vis of recomputed p-values (Table 7). We observe a somewhat low specificity of 79.7%. Note that this analysis only refers to a reported significance decision is valid with respect to a corresponding correct p-value, and *not* whether a positive reported result is true.

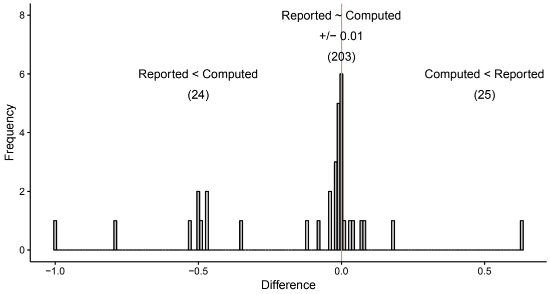


Fig. 9. Histogram of difference reported p-values minus statcheck-computed p-values.

Table 7. Confusion matrix for researchers determining significance.

Predicted	Reference	
	Significant	NS
Significant	191	12
NS	1	47

Accuracy: .95, 95% CI $[.91, .97]$, $Acc > NIR(.76)$, $< .001$***, Sensitivity = .99, Specificity = .80, PPV = .94, F_1 = .97

5.8 Supporting the STAST 2019 PC in Checking Statistics

Aligned with Recommendation 2 in Sect. 7, we offered a statcheck analysis to the STAST PC members to support the workshop's discussion phase. Of 28 submitted papers, 9 papers (32%) included a statistical inference.

Let us consider these 9 papers in detail as an exploratory analysis. One paper contained a major error in terms of statistics being invalid, two papers used the wrong statistical method for the experiment design at hand (e.g., independent-samples statistics in a dependent-samples design). Two of those three papers were also flagged by statcheck. These errors themselves, however, were detected by program committee members, not by the statcheck analysis.

On third of the papers reported statistics in an APA compliant format. 6 papers (66%) reported exact p-values, 4 papers (44%) reported effect sizes as required by the STAST submission guidelines. Of the 9 papers, 7 needed multiple-comparison corrections, which only two provided in their initial submission.

In terms of statcheck evaluation with the methodology of this study, we found 5 papers (56%) to be Incomplete, one paper Inconsistent, three papers (33%) CorrectNHST. This distribution is not significantly different from the SLR sample shown in Fig. 6a, $\chi^2(3) = 0.829, p = .843$, Cramér's $V = 0.082$, 95% CI $[0, 0.188]$.

6 Discussion

Incomplete reporting holds back the field. Nearly two thirds of the papers with p-values did not report a single complete test triplet (cf. Fig. 6a). This impairs the ability to cross-check internal consistency of tests and, thereby, undermines fault-tolerance. Hence, such papers have limited credibility and fidelity of statistical information.

The incomplete reporting observed in this study is in stark contrast to the analysis of the Journal of Media Psychology (JMP), in which not a single paper was Incomplete. Hence, we conclude that mandated reporting standards are an effective tool.

It is further troubling that the likelihood of incomplete reporting did not seem to decrease over time (cf. Fig. 7b).

In terms of research reuse and synthesis, the situation is aggravated, because effect sizes are vastly under-reported in this field. Only a small minority reports them explicitly; one third of the papers allows to infer them (cf. Fig. 2b).

There are three consequences to this phenomenon: (i) It is exceedingly difficult for practitioners to ascertain the magnitude of effects and, thereby, their practical significance. (ii) It is near-impossible to compare research results in meta-analyses and to synthesize well-founded summary effects. (iii) Hence, disputes over differences between original studies and replications are hard to settle satisfactorily.

While some errors are minor, we caution against clustered errors and miscalculations. Of the 44 papers with complete test statistic triplets analyzed, 60% were deemed correct; more than one quarter had at least one inconsistency; 14% had at least one decision error. Of all tests with complete triplets analyzed 14% were erroneous. Here, the socio-technical security sample showed similar error rates as the psychology sample.

Especially the 26 papers with complete test triplets and correct reporting—one quarter of the sample—stand testament to efforts of authors and program committees "get it right."

The errors observed by statcheck were often minor typos and rounding errors that could have been easily avoided, however nearly 40% seemed to be serious miscalculations. We found that these errors were at times clustered: there are a few papers with a number of errors.

There is a dark figure of decision errors lurking in the underuse of multiple-comparison corrections. This study leaves the detailed analysis of power and multiple-comparison corrections (MCCs) to future work. Still, we do not want to withhold insights already apparent from Fig. 3a: There is a Damocles sword hanging over many papers: Multiple-Comparison Corrections (MCCs).

We have seen in Fig. 2a that even though MCCs came in use from year 2009, only about one third of the papers employed them. From Fig. 3a, we observe that there are papers with a considerable number of reported p-values without MCCs. Hence, there may well be a sizable dark figure of papers with decision errors in store once adequate MCCs are employed.

These observations inform Recommendation 3 in that observing studies with many comparisons but without corrections can be an indication of the number of comparisons, multiple-comparison corrections as well as the power needed to sustain them only being considered as an afterthought.

Automated checking of statistical reporting is viable. The statcheck detection rates were very good and comparable to the rates reported by Nuijten et al. [13]. We note, however, that statcheck did not operate completely autonomously, but was complemented with human coding to overcome parsing issues. We find the approach viable for the use in socio-technical aspects of security.

6.1 Limitations

Generalizability. The study is based on an existing SLR sample that largely consists of SOUPS publications and only contains few cases for other venues. Dealing with a sparse matrix, the likelihoods computed for non-SOUPS venues as well as overall logistic regressions suffer from more uncertainty.

Syntactic Validity Checks. While we have made good experiences with statcheck and only found few false positives and negatives, we observe that statcheck results can suffer from hidden errors. While we complemented the automated analysis with a human review and coding of reported errors, we observe that statcheck could have missed or misinterpreted individual tests. However, based our inspection of the 114 analyzed papers, we expect that the number of statcheck errors is small compared to the 1775 tests analyzed. In the end, an automated tool cannot replace the trained eye of a knowledgable reviewer. However, this study is about the overall distribution of errors, which will be hardly skewed by rare false positives or negatives.

Deviations from the Pre-registration. We deviated from the OSF pre-registration by 1. not attempting the exploratory analysis of the impact of authors, 2. not attempting an exploratory logistic regression on completeness indicators, 3. abandoning the planned ordinal logistic regression in favor of the MLR, because SCOutcome did not yield an ordinal scale, 4. merging non-SOUPS venue levels to overcome the sparsity of the dataset, 5. not attempting further cross-validation due to low variance explained.

7 Recommendations

The recommendations made here need to be seen as part of a greater paradigm shift. Instead of focusing on single publications, one may consider that a study does not stand on its own. Truly advancing the knowledge of a field calls for creating robust studies that prepare the ground for systematic replications, reuse and research synthesis.

1. Establish sound reporting standards. Sound and generally accepted reporting standards could greatly improve the credibility of the field. This could either mean developing systematic reporting standards for socio-technical aspects of security or adopting existing standards.

Developing systematic reporting standards would involve a stable coalition of program committee chairs and members as well as journal editors forming a working group to that effect. Such a working group would likely take into account requirements for this field as well as examples of mature reporting standards from other fields.

Given that considerable thought has gone into APA standards [1] and that these standards apply to human dimensions, they are a viable and sufficiently mature candidate, at least when it comes to statistical reporting. Our analysis showed that the majority of papers reporting complete test statistics triplets were actually compliant to APA requirements.

While not perfect, their recommendations on statistical reporting could have considerable benefits for reporting fidelity, research reusability and synthesis. One option in this context would be to only adopt a subset of recommendations directly benefiting reporting fidelity.

In any case, one would consider sound reporting for test statistics themselves, effect sizes and their confidence intervals, as well as essential information on the sample, design and procedure. Again, this field can well take into account more comprehensive initiatives from other fields [10].

2. Support PCs in checking statistics. From our experience researching this study, we can attest that checking statistics can be a tedious affair. Even with all their failings, tools like statcheck can support program committee members in detecting incorrect results. Such an approach certainly requires human mediation to avoid false positives, yet can offer insights at low cost.

As reported in Sect. 5.8, we tested this recommendation on the STAST 2019 program committee. While statcheck correctly identified reporting issues and did not produce a false positive, major errors were discovered by program committee members in the analysis of experiment designs vis-à-vis their statistical inferences. This yields an indication that an automated tool, such as statcheck, will only support but never replace the expert judgment of the reviewers.

There are organizational methods, such as pre-registrations or registered reports, that can support a PC further in ascertaining the integrity of results.

3. Embrace *a priori* power and multiple-comparison corrections. We make this recommendation with a grain of salt, as we have not reported on a dedicated study on power, yet. However, even this study on reporting fidelity shows that this consideration would benefit the community.

Low power and missing adequate MCCs can well undermine the results of a good study and increase the likelihood of a positive result being a false positive. We encourage researchers to plan in advance for the power required, accounting for the MCCs necessary for the planned tests.

8 Conclusion

This study is the first systematic analysis of a large sample of security user studies with respect to their statistical reporting fidelity. For the first time, we offer a comprehensive, quantitative, and empirical analysis of the state-of-play of the field of socio-technical aspects of security. We offer a wealth of different perspectives on the sample, enabling us to obtain a fine-grained analysis as well as broad recommendations for authors and program committees alike.

We stress that the research and reviewing process for security user studies constitutes a socio-technical system in itself that impacts the decision making in security and privacy. Because scientists and practitioners alike seek to reuse research results, the fidelity or uncertainty of those results—especially their statistical inferences—plays a major role in the credibility of the field and the confidence of its audience. Hence, self-reflection of the field will ultimately impact the decision making by users in security and privacy, as well.

As future work, we consider expanding the sample, including further venues, such as CHI, as well as offering a dedicated analysis of statistical power and Positive Predictive Value (PPV) present in the field.

Acknowledgment. We would like to thank Malte Elson for the discussions on statcheck, on the corresponding analyses in psychology, and on general research methodology. We thank the anonymous reviewers of STAST 2019 for their discussion and insightful comments, as well as the volume co-editor Theo Tryfonas for offering additional pages to include the requested changes.

This study was in parts funded 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" (Pathway I led by Thomas Groß). The author was in parts funded by the ERC Starting Grant CASCAde (GA n°716980).

A Details on Qualitative Analysis

A.1 Errors Committed by statcheck

Parsing Accuracy. In all 34 error cases, statcheck parsed the PDF file correctly, and its raw test representation corresponded to the PDF. In all but two tests, statcheck recognized the test correctly. In said two cases, it mistook a non-standard-reported Shapiro-Wilk test as χ^2 test, creating two false positives. There was one case in which the statcheck computed p-value for an independent-samples t-test differed slightly from our own calculation, yet only marginally so, presumably because of a unreported Welch correction.

One-Tailed Tests. In seven cases, statcheck recognized one-tailed tests correctly. For three of those tests, the authors framed the hypotheses as one-tailed. In three other tests, the authors used one-tailed test results without declaring their use. There was one additional case in which the authors seemed to have used a one-tailed test, yet the rounding was so far off the one-tailed result that statcheck did not accept it as "valid if one-tailed" any longer. There was one test marked as "one-tail" which statcheck did not recognize as one-tailed, yet that test also suffered from rounding errors.

Dependent-Samples Tests. There were 7 papers using dependent-samples methods (such as matched-pair tests or mixed-methods regressions). We found that statcheck treated the corresponding dependent-samples statistics correctly.

Multiple Comparison Corrections. In three cases, statcheck did not recognize p-values that were correctly Bonferroni-corrected, counting as three false positives. It is an open point, however, how many paper should have employed multiple-comparison corrections, but have not done so, an analysis statcheck does not perform.

A.2 Errors Committed by Authors

Typos. We considered 6 to be typos or transcription errors (18%). Another 1 error seemed to be a copy-paste error (3%)

Rounding Errors. Of all 34 reported errors, we found 8 to be rounding errors (24%).

Miscalculations. We found 13 cases to be erroneous calculations (38%).

A.3 Composition of Incomplete p-Values

Of 1523 incomplete cases, 134 were declared "non-significant" without giving the actual p-value (8.8%). Further, 6 were shown as $p > .05$. (0.394%).

Of the incomplete cases, 102 were reported statistically significant at a .05 significance level (6.7%).

Of the incomplete cases, 477 were reported statistically significant at a lower significance level of .01, .001, or .0001 (31.3%).

Of 1523 incomplete p-values, 680 gave an exact p-value (44.6%). Of those exactly reported p-values, half (367) were claimed statistically significant at a significance level of $\alpha = .05$ (54%). Of those exatly reported p-values, 19 claimed an impossible p-value of $p = 0$ (2.79%).

Online Supplementary Materials

We made the materials of the study (specification of the inputted SLR, included sample, contingency tables) publicly available at its Open Science Framework Repository (see Footnote 1).

References

1. American Psychological Association (ed.): Publication Manual of the American Psychological Association, 6th revised edn. American Psychological Association (2009)
2. Coopamootoo, K.P.L., Groß, T.: Cyber security and privacy experiments: a design and reporting toolkit. In: Hansen, M., Kosta, E., Nai-Fovino, I., Fischer-Hübner, S. (eds.) Privacy and Identity 2017. IAICT, vol. 526, pp. 243–262. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92925-5_17
3. Coopamootoo, K., Groß, T.: Systematic evaluation for evidence-based methods in cyber security. Technical report TR-1528, Newcastle University (2017)
4. Coopamootoo, K.P.L., Groß, T.: Evidence-based methods for privacy and identity management. In: Lehmann, A., Whitehouse, D., Fischer-Hübner, S., Fritsch, L., Raab, C. (eds.) Privacy and Identity 2016. IAICT, vol. 498, pp. 105–121. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55783-0_9
5. Cumming, G.: Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. Routledge, New York (2013)
6. Elson, M., Przybylski, A.K.: The science of technology and human behavior - standards old and new. J. Media Psychol. **29**(1), 1–7 (2017). https://doi.org/10.1027/1864-1105/a000212
7. Epskamp, S., Nuijten, M.B.: statcheck: extract statistics from articles and recompute p values (v1.3.0), May 2018. https://CRAN.R-project.org/package=statcheck
8. Fox, J., Andersen, R.: Effect displays for multinomial and proportional-odds logit models. Sociol. Methodol. **36**(1), 225–255 (2006)
9. Lakens, D.: Checking your stats, and some errors we make, October 2015. http://daniellakens.blogspot.com/2015/10/checking-your-stats-and-some-errors-we.html
10. LeBel, E.P., McCarthy, R.J., Earp, B.D., Elson, M., Vanpaemel, W.: A unified framework to quantify the credibility of scientific findings. Adv. Methods Pract. Psychol. Sci. **1**(3), 389–402 (2018)
11. Maxion, R.: Making experiments dependable. In: Jones, C.B., Lloyd, J.L. (eds.) Dependable and Historic Computing. LNCS, vol. 6875, pp. 344–357. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24541-1_26
12. Moher, D., et al.: CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. J. Clin. Epidemiol. **63**(8), e1–e37 (2010)
13. Nuijten, M.B., van Assen, M.A., Hartgerink, C.H., Epskamp, S., Wicherts, J.: The validity of the tool "statcheck" in discovering statistical reporting inconsistencies (2017). https://psyarxiv.com/tcxaj/
14. Nuijten, M.B., Hartgerink, C.H.J., van Assen, M.A.L.M., Epskamp, S., Wicherts, J.M.: The prevalence of statistical reporting errors in psychology (1985–2013). Behav. Res. Methods **48**(4), 1205–1226 (2015). https://doi.org/10.3758/s13428-015-0664-2

15. Peisert, S., Bishop, M.: How to design computer security experiments. In: Futcher, L., Dodge, R. (eds.) WISE 2007. IAICT, vol. 237, pp. 141–148. Springer, New York (2007). https://doi.org/10.1007/978-0-387-73269-5_19
16. Ripley, B., Venables, W.: nnet: feed-forward neural networks and multinomial log-linear models, February 2016. https://CRAN.R-project.org/package=nnet
17. Schechter, S.: Common pitfalls in writing about security and privacy human subjects experiments, and how to avoid them (2013). https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/commonpitfalls.pdf
18. Schmidt, T.: Sources of false positives and false negatives in the STATCHECK algorithm: reply to Nuijten et al. (2016). https://arxiv.org/abs/1610.01010