Statistical Reliability of 10 Years of Cyber Security User Studies

Thomas Groß^(✉)

School of Computing, Newcastle University, Newcastle upon Tyne, UK
`thomas.gross@newcastle.ac.uk`

Abstract. Background. In recent years, cyber security user studies have been appraised in meta-research, mostly focusing on the completeness of their statistical inferences and the fidelity of their statistical reporting. However, estimates of the field's distribution of statistical power and its publication bias have not received much attention.

Aim. In this study, we aim to estimate the effect sizes and their standard errors present as well as the implications on statistical power and publication bias.

Method. We built upon a published systematic literature review of 146 user studies in cyber security (2006–2016). We took into account 431 statistical inferences including t -, χ^2 -, r -, one-way F -tests, and Z -tests. In addition, we coded the corresponding total sample sizes, group sizes and test families. Given these data, we established the observed effect sizes and evaluated the overall publication bias. We further computed the statistical power vis-à-vis of parametrized population thresholds to gain unbiased estimates of the power distribution.

Results. We obtained a distribution of effect sizes and their conversion into comparable log odds ratios together with their standard errors. We, further, gained funnel-plot estimates of the publication bias present in the sample as well as insights into the power distribution and its consequences.

Conclusions. Through the lenses of power and publication bias, we shed light on the statistical reliability of the studies in the field. The upshot of this introspection is practical recommendations on conducting and evaluating studies to advance the field.

Keywords: User studies · SLR · Cyber security · Effect estimation · Statistical power · Publication bias · Winner's curse

1 Introduction

Cyber security user studies and quantitative studies in socio-technical aspects of security in general often rely on statistical inferences to make their case that

Open Science Framework: osf.io/bcbyte. An extended version of this work is available on arXiv: <https://arxiv.org/abs/2010.02117>.

© The Author(s) 2021

T. Groß and L. Viganò (Eds.): STAST 2020, LNCS 12812, pp. 171–190, 2021.

https://doi.org/10.1007/978-3-030-79318-0_10

observed effects are not down to chance. They are to separate the wheat from the chaff. Indeed, null hypothesis significance testing and p -values indicating statistical significance hold great sway in the community. While the studies in the field have been appraised in recent years on the completeness and fidelity of their statistical reporting, we may still ask how reliable the underlying statistical inferences really are.

“To what extent can we rely on reported effects?” This question can take multiple shapes. First, we may consider the magnitude of observed effects. While a statement of statistical significance is dependent on the sample size at which the inference was obtained, the magnitude of an effect, its *effect size*, informs us whether an effect is practically relevant—or not. While small effects might not make much difference in practice and might not be economical to pursue, large effects estimated with confidence can guide us to the interventions that are likely carrying considerable weight in socio-technical systems.

Indeed, a second dimension of reliability pertains to the confidence we have in observed effects, typically measured with 95% confidence intervals. Here, we are interested how tightly the confidence interval envelops the effect point estimate. The rationale behind such a confidence interval is that if an experiment were repeated many times, we would expect 95% of the observed effect estimates to be within the stated confidence intervals. Wide intervals, thereby, give us little confidence in the accuracy of an estimation procedure.

This consideration is exacerbated if a study conducted many tests in the same test family. Given the risk of multiple comparisons to amplify false-positive rates, we would need to adjust the confidence intervals accounting for the multiplicity and, hence, be prepared to gain even less confidence in the findings.

Third, we may consider statistical power, the likelihood of finding an effect that is present in reality. To put it in precise terms, it is the likelihood of rejecting a null hypothesis when it is, in fact, false—the complement of the false negative rate. At the same time, statistical power also impacts the likelihood that a positive report is actually true, hence further impacts the reliability of a finding. The power distribution, further, offers a first assessment on the statistical reliability of the field.

Finally, we expand on the reliability of the field in terms of evaluating research biases that could undermine results. Two predominant biases of interest are (i) the publication bias [24], and (ii) the related winner’s curse [3].

The publication bias, on the one hand, refers to the phenomenon that the outcome of a study determines the decision to publish. Hence, statistically significant positive results are more likely to be published, than null results—even if null results live up to the same scientific rigor and possibly carry more information for falsification. Furthermore, researchers might be incentivized to engage in research practices that ensure reporting of statistically significant results, introducing biases towards questionable research practices.

The winner’s curse, on the other hand, refers to the phenomenon that under-powered studies tend to report more extreme effects with statistical significance, hence tend to introducing a bias in the mean effect estimates in the field.

To the best of our knowledge, these questions on reliability of statistical inferences in cyber security user studies have not been systematically answered, to date. Coopamootoo and Groß [6] offered a manual coding of syntactic completeness indicators on studies sampled in a systematic literature review (SLR) of 10 years of cyber security user studies, while also commenting on *post-hoc* power estimates for a small sub-sample. Groß [14] investigated the fidelity of statistical test reporting along with an overview of multiple-comparison corrections and the identification of computation and decision errors. While we chose to base our analysis on the same published SLR sample, we close the research gap by creating a sound empirical foundation to estimate effect sizes, their standard errors and confidence intervals, by establishing power simulations vs. typical effect size thresholds, by investigating publication bias and winner's curse.

Our Contributions. We are the first to estimate a large number ($n = 431$) of heterogenous effect sizes from cyber security user studies with their confidence intervals. Based on this estimation, we are able to show that a considerable number of tests executed in the field are underpowered, leaving results in question. This holds especially for small studies which computed a large number of tests at vanishingly low power. Furthermore, we are able to show that the reported effects of underpowered studies are especially susceptible to falter under Multiple-Comparison Corrections (MCC), while adequately powered studies are robust to MCC.

We are the first to quantify empirically that a publication bias is present in the field of cyber security user studies. We can further evidence that the field suffers from the over-estimated effect sizes at low power, the winner's curse. We conclude our study with practical and empirically grounded recommendations for researchers, reviewers and funders.

2 Background

2.1 Statistical Inferences and Null Hypothesis Significance Testing

Based on a—necessarily *a priori*—specified null hypothesis (and alternative hypothesis) and a given significance level α , statistical inference with *null hypothesis significance testing* [18, pp. 163] sets out to establish how surprising an obtained observation D is, assuming the null hypothesis being true. This is facilitated by means of a *test statistic* that relates observations to appropriate probability distributions. It is inherent to the method that the statistical hypothesis *must* be fixed, before the sample is examined.

The *p-value*, then, is the likelihood of obtaining an observation as extreme as or more extreme than D , contingent on the null hypothesis being true, all assumption of the test statistic being fulfilled, the sample being drawn randomly, etc. Indeed, not heeding the assumptions of the test statistic is one of the more subtle ways how the process can fail.

Statistical inferences carry the likelihood of a false positive or *Type I* error [18, pp. 168]. They are impacted, hence, by *multiplicity*, that is, the phenomenon

that computing multiple statistical tests on a test family inflates the family-wise error rate. To mitigate this effect, it is prudent practice to employ *multiple-comparison corrections* (MCC) [18, pp. 415]. The Bonferroni correction we use here is the most conservative one, adjusting the significance level α by dividing it by the number of tests computed in the test family.

2.2 Effect Sizes and Confidence Intervals

We briefly introduce estimation theory [10] as a complement to significance testing and as a key tool for this study. An observed *effect size* (ES) is a *point estimate* of the magnitude of an observed effect. Its *confidence interval* (CI) is the corresponding interval estimate [18, pp. 313]. For instance, if we consider the popular 95% confidence interval on an effect size, it indicates that if an experiment were repeated infinitely many times, we would expect that the point estimate on the population effect were within the respective confidence interval 95% of the cases. The *standard error* of an ES is equally a measure of the effects uncertainty and monotonously related to the width of the corresponding confidence interval.

Notably, confidence intervals are often misused or misinterpreted [17, 22]. For instance, they do *not* assert that the population effect is within a point estimate's CI with 95% likelihood.

However, used correctly, effect sizes and their confidence intervals are useful in establishing the practical relevance of and confidence in an effect [13]. They are, thereby, recommended as minimum requirement for standard reporting, such as by the APA guidelines [1] Whereas a statement of statistical significance or *p*-value largely gives a binary answer, an effect size quantifies the effect observed and, thereby, indicates what its impact in practice might be.

2.3 Statistical Power

In simple terms, *statistical power* ($1 - \beta$) [4] is the probability that a test correctly rejects the null hypothesis, if the null hypothesis is false in reality. Hence, power is the likelihood not to commit a false negative or *Type II* error.

It should not go unnoticed that power also has an impact on the probability whether a positively reported result is actually true in reality, often referred to as *Positive Predictive Value* (PPV) [19]. The lower the statistical power, the less likely a positive report is true in reality. Hence, a field affected by predominately low power is said to suffer from a *power failure* [3].

Statistical power is largely determined by significance level, sample size, and the population effect size θ . *A priori* statistical power of a test statistic is estimated by a power analysis [18, pp. 372] on the sample size employed vis-à-vis of the anticipated effect size, given a significance level α and target power $1 - \beta$.

Post-hoc statistical power [18, p. 391], that is, computed on observed effect sizes after the fact, is not only considered redundant to the *p*-value and confidence intervals on the effect sizes, but also cautioned against as treacherously

misleading: It tends to overestimate the statistical power because it discounts the power lost in the study execution and because it is vulnerable to being inflated by over-estimated observed effect sizes. Hence, especially small under-powered studies with erratic effect size estimates tend to yield a biased post-hoc power. Hence, post-hoc power statements are best disregarded.

We offer a less biased alternative approach in *power simulation*. In that, we specify standard effect size thresholds, that is, we parametrize the analysis on assumed average effect sizes found in a field. We then compute the statistical power of the studies given on their reported sample size against those thresholds. As the true average effect sizes of our field are unknown, we offer power simulations for a range of typical effect size thresholds.

2.4 Research Biases

Naturally, even well-executed studies can be impacted by a range of biases on per-study level. In this study, we consider biases of a field, instead. We zero in on two biases, specifically: (i) the publication bias and (ii) the winner's curse.

The *publication bias* [11, 20, 24, 27] refers to the phenomenon that the publication of studies may be contingent on their positive results, hence condemning null and unfavorable results to the file-drawer [24, 25].

The *winner's curse* [3] is a specific kind of publication bias referring to the phenomenon that low-power studies only reach statistically significant results on large effects and, thereby, tend to overestimate the observed effect sizes. They, hence, perpetuate inflated effect estimates in the field.

We chose them as lens for this paper because they both operate on the interrelation between sample size (impacting standard error and power) and effects observed and emphasize different aspects of the overall phenomenon. The publication bias is typically visualized with funnel plots [20], which pit observed effect sizes against their standard errors. We shall follow Sterne and Egger's suggestion [28] on using log odds ratios as best suited x -axis. If no publication bias were present, funnel plots would be symmetrical. Hence, an asymmetry is an indication of bias. This asymmetry is tested with the non-parametric rank correlation coefficient Kendall's τ [2]. We note that funnel-plots as analysis tools can be impacted by the heterogeneity of the effect sizes investigated [29] and, hence, need to be taken with a grain of salt.

3 Related Works

3.1 Appraisal of the Field

Usable security, socio-technical aspects in security, human dimensions of cyber security and evidence-based methods of security are all young fields. The Systematic Literature Review (SLR) by Coopamootoo and Groß [6], hence, zeroed in on cyber security user studies published in the 10 years 2006–2016. The field

has undergone some appraisal and self-reflection. The mentioned Coopamootoo-Groß SLR considered completeness indicators for statistical inference, syntactically codeable from a study's reporting [8]. These were subsequently described in a reporting toolset [9]. The authors found appreciable weaknesses in the field, even if there were cases of studies excelling in their rigor. Operating from the same SLR sample, Groß [14, 15] investigated the fidelity of statistical reporting, on completeness of the reports as well as the correctness of the reported p -values, finding computation and decision errors in published works relatively stable over time with minor differences between venues.

3.2 Guidelines

Over the timeframe covered by the aforementioned SLR, a number of authors offered recommendations for dependable, rigorous experimentation pertaining to this study. Peisert and Bishop [23] considered the scientific design of security experiments. Maxion [21] discussed dependable experimentation, summarizing classical features of sound experiment design. Schechter [26] spoke from experience in the SOUPS program committee, offering recommendations for authors. His considerations on multiple-comparison corrections and adherence to statistical assumptions foreshadow recommendations we will make. Coopamootoo and Groß [7] summarized research methodology, largely focusing on quantitative and evidence-based methods, discussing null hypothesis significance testing, effect size estimation, and statistical power, among other things.

4 Aims

Effect Sizes. As a stepping stone, we intend to estimate *observed* effect sizes and their standard errors in a standardized format (log odds ratios).

RQ 1 (Effect Sizes and their Confidence). *What is the distribution of observed effect sizes and their 95% confidence intervals? How are the confidence intervals affected by multiple-comparison corrections?*

$H_{mcc,0}$: The marginal proportions of tests' statistical significance are equal irrespective of per-study family-wise multiple-comparison corrections.

$H_{mcc,1}$: Per-study family-wise multiple-comparison corrections impact the marginal proportions of tests' statistical significance.

Statistical Power. We inquire about the statistical power of studies independent from their possibly biased observed effect size estimates.

RQ 2 (Statistical Power). *What is the distribution of statistical power vis-à-vis parameterized effect size thresholds? As an upper bound achievable with given sample sizes as well as for the actual tests employed?*

Given the unreliability of post-hoc power analysis, we pit the sample sizes employed by the studies and individual tests against the small, medium, and large effect size thresholds according to Cohen [4]. The actual thresholds will differ depending on the type of the effect size.

Publication Bias. We intend to inspect the relation between effect sizes and standard errors with funnel plots [20], asking the question:

RQ 3 (Publication Bias). *To what extent does the field exhibit signs of publication bias measured in terms of relation between effect sizes and standard errors as well as asymmetry?*

We can test statistically for the presence of asymmetry [2] as indicator of publication bias, yielding the following hypotheses:

$H_{\text{bias},0}$: There is no asymmetry measured as rank correlation between effect sizes and their standard errors.

$H_{\text{bias},1}$: There an asymmetry measured as rank correlation between effect sizes and their standard errors.

The Winner's Curse. We are interested whether low-powered studies exhibit inflated effect sizes and ask:

RQ 4 (Winner's Curse). *What is the relation between simulated statistical power (only dependent on group sizes) and observed effect sizes?*

$H_{\text{wc},0}$: Simulated power and observed effect size are independent.

$H_{\text{wc},1}$: There is a negative correlation between simulated power and observed effect size.

5 Method

This study was registered on the Open Science Framework (OSF)¹, before its statistical inferences commenced. An extended version of this work is available on arXiv², including additional analyses and a brief specification of the underlying SLR [16]. Computations of statistics, graphs and tables are done in R with the packages `statcheck`, `metafor`, `esc`, `compute.es`, `pwr`. Their results are woven into this report with `knitr`. Statistics are computed as two-tailed with $\alpha = .05$ as reference significance level. Multiple-comparison corrections are computed with the Bonferroni method, adjusting the significance level used with the number of members of the test family.

5.1 Sample

The sample for this study is based on a 2016/17 Systematic Literature Review (SLR) conducted by Coopamootoo and Groß [6]. This underlying SLR, its search, inclusion and exclusion criteria are reported in short form by Groß [14] are included in this study's OSF Repository. We have chosen this SLR on the one hand, because its search strategy, inclusion and exclusion criteria are explicitly documented supporting its reproducibility and representativeness; the list

¹ <https://osf.io/bcbyte/>.

² <https://arxiv.org/abs/2010.02117>.

of included papers is published. On the other hand, we have chosen it as sample, because there have been related analyses on qualitatively coded completeness indicators as well as statistical reporting fidelity [14] already. Therefore, we extend a common touchstone for the field. The overall SLR sample included $N = 146$ cyber security user studies. Therein, Groß [14] identified 112 studies with valid statistical reporting in the form of triplets of test statistic, degrees of freedom, and p -value. In this study, we extract effect sizes for t -, χ^2 -, r -, one-way F -tests, and Z -tests, complementing automated with manual extraction.

5.2 Procedure

We outlined the overall procedure in Fig. 1 and will describe the analysis stages depicted in dashed rounded boxes in turn.

Fig. 1. Flow chart of the analysis procedure

Automated Test Statistic Extraction. We analyzed the SLR sample with R package *statcheck* proposed by Epskamp and Nuijten [12]. We obtained cases on test statistic type, degrees of freedom, value of the test statistic and p -value along with a correctness analysis. This extraction covered correctly reported test statistics (by APA guidelines) and t -, χ^2 -, r -, F -tests, and Z -tests at that.

Manual Coding. For all papers in the SLR, we coded the overall sample size, use of Amazon Mechanical Turk as sampling platform, and the presence of multiple-comparison corrections. For each statistical test, we also coded group sizes, test statistics, degrees of freedom, p -values, means and standard deviations if applicable as well as test families. For the coding of test families, we distinguished different studies reported in papers, test types as well as dimensions investigated.

Test Exclusion. To put all effect sizes on even footing, we excluded tests violating assumption and ones not constituting one-way comparisons.

Power Simulation. We conducted a power simulation, that is, we specified effect size thresholds for various effect size types according to the classes proposed by Cohen [5]. Table 1 summarizes corresponding thresholds.

Table 1. Effect size thresholds for various statistics and effect size (ES) types [4]

ES type	Statistic	Threshold		
		Small	Medium	Large
Cohen’s d	t	0.20	0.50	0.80
Pearson’s r	r	0.10	0.30	0.50
Cohen’s w	χ^2	0.10	0.30	0.50
Cohen’s f	F	0.10	0.25	0.40

Given the sample sizes obtained in the coding, we then computed the *a priori* power against those thresholds with the R package `pwr`, which is independent from possible over-estimation of observed effect sizes. We further computed power analyses based on group sizes for reported tests, including a power adjusted for multiple comparisons in studies’ test families with a Bonferroni correction. We reported those analyses per test statistic type.

Estimation. We computed a systematic estimation of observed effect sizes, their standard errors and confidence intervals. This estimation was either based on test statistics, their degrees of freedom and group sizes used for the test or on summary statistics such as reported means, standard deviations and group sizes. We conducted the estimation with the R packages `esc` and `compute.es` for cases in which only test statistics were available and with the package `metafor` if we worked with summary statistics (e.g., means and standard deviations). As part of this estimation stage, we also estimated 95% confidence intervals (with and without multiple-comparison corrections).

Publication Bias Analysis. We used the R package `metafor` to compute analyses on the publication bias. In particular, we produced funnel plots on effect sizes and their standard errors [20]. For this analysis, we converted all effect sizes and standard errors irrespective of their origin to log odds ratios as the predominant effect-size form for funnel plots [28]. Following the method of Beggs and Mazumdar [2], we evaluated a rank correlation test to ascertain the presence of asymmetry.

Winner’s Curse Analysis. To analyze for the winner’s curse, we created scatterplots that pitted the simulated power of reported tests against the observed effect sizes extracted from the papers. We applied a Loess smoothing to illustrate the bias in the distribution. Finally, we computed a Kendall’s τ rank correlation to show the relationship between absolute effect size and power. We employed a robust linear regression using an iterated re-weighted least squares (IWLS) fitting to estimate the expected effect size of the field at 100% power.

6 Results

6.1 Sample

The sample was refined in multiple stages, first establishing papers that are candidates for effect size extraction, their refinement shown in Table 2 in Appendix A. In total, we retained a sample of $N_{\text{studies}} = 54$ studies suitable for effect size extraction.

Secondly, we set out to extract test statistics and effect sizes with `statcheck` and manual coding. Table 3 in Appendix A gives an overview how these extracted tests were first composed and then pruned in an exclusions process focused on statistical validity. After exclusion of tests that would not yield valid effect sizes, we $N_{\text{es}} = 454$ of usable effect sizes and their standard errors.

We include the descriptives of the complete sample of extracted effect sizes grouped by their tests in Table 4 of Appendix A. The table standardizes all effect sizes as log odds ratios, irrespective of test statistic of origin.

6.2 Effect Size Estimates and Their Confidence

In Fig. 2, we analyze the effect size estimates of our sample with their confidence intervals in a caterpillar plot: estimated effect sizes are plotted with error bars representing their 95% confidence intervals and ordered by effect size. Two thirds of the observed effects did not pass the medium threshold: (i) 37% were trivial, (ii) 28% were small, (iii) 15% were medium, and (iv) 20% were large.

The figure underlays the uncorrected confidence intervals (gray) with the multiple-comparison-corrected confidence intervals in red. While 54% of 431 tests were statistically significant without MCC, only 38% were significant after appropriate MCC were applied.

The multiple-comparison correction significantly impacted the significance of the tests, $FET\ p < .001$, $OR = 0.53$, 95% CI [0.4, 0.7] We, thereby, reject the null hypothesis $H_{\text{mcc},0}$.

6.3 Upper Bounds of Statistical Power

We estimate the upper-bound statistical power studies can achieve had they used their entire sample for a single two-tailed independent-samples t -test versus a given standardized mean difference effect size. Thereby, Fig. 3 offers us a first characterization of the field in a beaded monotonously growing power plot.

Fig. 2. Caterpillar forest plot of $n = 431$ log odds ratios and their 95% confidence intervals, ordered by $\log(\text{OR})$.

Fig. 3. Upper-bound of power against Standardized Mean Difference (SMD) effects and 112 observed study sample sizes N in SLR. (*Note:* Only studies with $N < 1250$ are shown for visual clarity, excluding 14 from the view)

Fig. 4. Histogram-density plot comparing statistical power for all tests in the sample by MCC. *Note:* The histogram is with MCC and square-root transformed.

Let us unpack what we can learn from the graph. Regarding the sample size density on the top of Fig. 3a, we observe that the sample sizes are heavily biased towards small samples ($N < 100$). Considering the ridge power plot in Fig. 3b, the middle ridge of power versus medium effects shows the field to be bipartite: There is there is a peak of studies achieving greater than 80% power against a medium effect. Those studies match the profile of studies with *a priori* power analysis seeking to achieve the recommended 80% power. However, roughly the same density mass is in smaller studies failing this goal. The bottom ridge line tells us that almost no studies achieve recommended power against small effects.

6.4 Power of Actual Tests

Figure 4 illustrates the power distribution of all tests and all ES thresholds investigated taking into account their respective group sizes, comparing between scenarios with and without MCC. Notably, the studies tests designed to have 80% power largely retain their power under MCC. We observe a considerable number of tests with power of approx. 50% which falter under MCC.

Distinguishing further between different test types, we considered independent-samples t - and 2×2 χ^2 -tests as the most prevalent test statistics. Their respective power simulations are included in the extended version of this paper [16].

In both cases, we observe the following phenomena: (i) The density mass is on smaller sample sizes. (ii) The ridge-density plots show characteristic “two-humped” shapes, in exhibiting a peak above 80% power, but also a density mass at considerably lower power. (iii) Both t -tests and χ^2 -tests were largely ill-equipped to detect small effect sizes. Overall, we see a self-similarity of the MCC-corrected power of actual tests vis-à-vis of the upper-bound power considered in the preceding section.

Fig. 5. Funnel plots of $\log(OR)$ effect sizes and their standard errors

6.5 Publication Bias

The funnel plots in Fig. 5 shows the results for 47 papers and a total of 431 statistical tests. For the aggregated plot Fig. 5a, we computed the mean log odds ratio and mean standard error per paper. We observe in both plots that with greater standard errors (that is, smaller samples), the effect sizes become more extreme. Hence, we conjecture that smaller studies which did not find significant effects were not published.

By the Begg-Mazumdar rank-correlation test [2], there is a statistically significant asymmetry showing the publication bias in the per-paper aggregate, Kendall’s $\tau(N = 47) = .349$, $p < .001$, Pearson’s $r = .52$, 95% CI [.52, .52]. We reject null hypothesis $H_{bias,0}$.

6.6 The Winner’s Curse

In Fig. 6 depicts the winner’s curse phenomenon by pitting the simulated power against a threshold medium effect against the observed effect sizes. We observe that at low power, extreme results were more prevalent. At high power, the results were largely clustered closely around the predicted mean log odds.

Fig. 6. ES Bias by Power illustrating the Winner's Curse. *Note:* Entries with more than 1000% bias were removed for visual clarity without impact on the result.

There was a statistically significant negative correlation between power and observed effect size, that is with increasing power the observed effect sizes decrease, Kendall's $\tau(N = 396) = -.338$, $p < .001$, corresponding to an ES of Pearson's $r = -.51$, 95% CI $[-.51, -.51]$ using Kendall's estimate. We reject the winner's curse null hypothesis $H_{wc,0}$.

We evaluated an iterated re-weighted least squares (IWLS) robust linear regression (RLM) on the $ES \sim power$ relation mitigating for outliers, statistically significant at $F(1, 394) = 114.135$, $p < .001$. We obtained an intercept of 1.6 95% CI $[1.44, 1.76]$, $F(1, 394) = 331.619$, $p < .001$. For every 10% of power, the measured effect size decreased by -0.11 ; 95% CI $[-0.13, -0.09]$, $F(1, 394) = 114.135$, $p < .001$. The simulated-power regression explained approximately $R^2 = .08$ of the variance; the standard error of the regression was $S = 0.19$.

We can extrapolate to the expected mean log odds ratio at 100% power $\log(OR) = 0.47$, 95% CI $[0.21, 0.72]$. This corresponds to an SMD estimate in Cohen's $d = 0.26$, 95% CI $[0.12, 0.4]$.

7 Discussion

The Power Distribution is Characteristically Two-Humped. We found empirical evidence that a substantive number of studies and half the tests extracted were adequate for 80% power at a medium target effect size. Hence, it is plausible to conjecture an unspoken assumption in the field that the population effect sizes in

cyber security user studies are medium (e.g., Cohen's $d \geq .50$). The good news here is that studies that were appropriately powered, that is, aiming for 80% power, retained that power also under multiple-comparison corrections. Studies which were under-powered in the first place, got entangled by MCCs and ended up with negligible power retained (cf. Fig. 4, Sect. 6.4).

Having said that, the power distribution for the upper bound as well as for actual tests came in two “humps.” While we consistently observed peaks at greater than 80% power for medium effect sizes, there was a density mass of under-powered tests, where the distribution was roughly split half-half. Typically, tests were altogether too under-powered to detect small effect sizes. Overall, we believe we have evidence to attest a power failure in the field.

Population Effect Sizes May Be Smaller Than We Think. The problem of power failure is aggravated by the mean effect sizes in the SLR having been close to small, shown in the caterpillar forest plot (Fig. 2) and the ES descriptives (Table 4). In fact, our winner's curse analysis estimated a mean Cohen's $d = 0.26$, 95% CI [0.12, 0.4]. Of course, it is best to obtain precise effect size estimates for the effect in question from prior research, ideally from systematic meta-analyses deriving the estimate of population effect size $\hat{\theta}$. Still, the low effect size indicated here should give us pause: aiming for a medium effect size as a rule of thumb might be too optimistic.

Cyber Security User Studies Suffer From a Host of Biases. We showed the presence of an appreciable publication bias (cf. Fig. 5, Sect. 6.5), that is, the phenomenon that the publication of studies was contingent on their positive outcomes, and found evidence of the winner's curse, that is, the phenomenon that under-powered studies yielded exaggerated effect estimates (cf. Fig. 6, Sect. 6.6).

Taken together with the likely close-to-small population effect sizes and the diagnosed power failure, we need to conclude that the field is prone to accept publications that are seemingly “positive” results, while perpetuating biased studies with over-estimated effect sizes. These issues could be resolved with a joint effort by field's stakeholders—authors, gatekeepers and funders: paying greater attention to statistical power, point and interval estimates of effects, and adherence to multiple-comparison corrections.

7.1 Limitations

Generalizability. We observe that we needed to exclude a considerable number of studies and statistical tests. This is consistent with the observations by Coopamootoo and Groß [6] on prevalent reporting completeness, finding that 71% of their SLR sample did not follow standard reporting guidelines and only 31% combinations of actual test statistic, p -value and corresponding descriptives. Similarly, Groß [14] found that 69 papers (60%) did not contain a single completely reported test statistic. Hence, we also observe that meta research is severely hamstrung by the reporting practices found in the field.

We note, further, that we needed to exclude 104 extracted statistical tests and effect sizes due to problems in how these tests were employed, leading to 17 less represented. Studies that inappropriately used independent-sample tests in a dependent-sample research designs or violated other assumptions by, e.g., using difference-between-means test statistics (expecting a t distribution) to test differences between proportions (z -distribution), needed to be excluded to prevent perpetuation of those issues. Finally, we needed to exclude 74 tests because papers reported tests with degrees of freedom $df > 1$ without the summary statistics to establish the effect sizes. Even though those studies contained complete reports the auxiliary data to estimate the effects were missing.

These exclusions on empirical grounds limit generalizability. The retained sample of 431 tests is focused on the studies that were most diligent in their reporting. This fact, however, makes our investigation more conservative rather than less so.

This Is Not a Meta-analysis. Proper meta-analysis combines effect sizes on *similar* constructs to summary effects. Given that studies operating on the same constructs are few and far between in cyber security user studies, we standardized all effects to log odds ratios to gain a rough overall estimate of the field.

8 Concluding Recommendations

We are the first to evaluate the statistical reliability of this field on empirical grounds. While there is a range of possible explanations of the phenomena we have found—including questionable research practices in, e.g., shirking multiple-comparison corrections in search of significant findings, missing awareness of statistical power and multiplicity, or limited resources to pursue adequately powered studies—we believe the evidence of power failure, possibly close-to-small population effect sizes, and biased findings can lead to empirically underpinned recommendations. We believe that these issues, however, are systemic in nature and that the actions of different stakeholders are, thereby, inter-dependent. Hence, in the following we aim at offering recommendations to different stakeholder, making the assumption that they aim at advancing the knowledge of the field to the best of their ability and resources.

Researchers. The most important recommendation here is: *plan ahead with the end in mind.* That starts with inquiring typical effect sizes for the phenomena investigated. If the reported confidence intervals thereon are wide, it is prudent to choose a conservative estimate. It is tempting to just assume a medium effect size (e.g., Cohen's $d = 0.5$) as aim, but there is no guarantee the population effect sizes are that large. Our study suggests they are not.

While it is a prudent recommendation to conduct an *a priori* power analysis, we go a step further and recommend to anticipate multiple comparisons one might make. Adjusting the target significance level with a Bonferroni correction for that multiplicity can prepare the ground for a study retaining sufficient power

all the way. This kind of foresight is well supported by a practice of spelling out the research aims and intended statistical inferences *a priori* (e.g., in a pre-registration). Taken together these measures aim countering the risk of a power failure.

Speaking from our experience of painstakingly extracting effects from a body of literature, we are compelled to emphasize: One of the main points of strong research is that it is *reusable* by other scientists. This goal is best served by reporting effect sizes and their confidence intervals as well as full triplets of test statistic, degrees of freedom and exact *p*-values, while also offering all summary statistics to enable others to re-compute the estimates. It is worth recalling that *all* tests undertaken should be reported and that rigorous, well-reported studies have intrinsic value, null result or not. This line of recommendations aims at enabling the gatekeepers of the field to do their work efficiently.

Gatekeepers. It bears repeating that the main goal of science is to *advance the knowledge of a field*. With reviewers, program chairs and editors being gatekeepers and the arbiters of this goal, it is worthwhile to consider that the goal is not served well in pursuing shiny significant results or valuing novelty above all else. Such a value system is prone to fail to ward against publication and related biases. A well-powered null result or replication attempt can go a long way in initiating the falsification of a theory in need of debunking. Because empirical epistemology is rooted in falsification and replication, we need the multiple inquiries on the same phenomena. We should strive to include adequately-powered studies of sufficient rigor irrespective of the “positiveness” of the results presented, exercising the cognitive restraint to counter publication bias.

Reviewers can support this by insisting on systematic reporting and on getting to see *a priori* specifications of aims, research designs, tests conducted, as well as sample size determinations, hence creating an incentive to protect against power failure. This recommendation dovetails with the fact that statistical inference is contingent on fulfilling the assumptions of the tests used, where the onus of proof is with the researchers to ascertain that all assumptions were satisfied. Those recommendations are in place to enable the gatekeepers to effectively ascertain the statistical validity and reliability of studies at hand.

Funders. With significant investments being made in putting cyber security user studies on an evidence-based footing, we recall: “*Money talks.*” On the one hand, we see the responsibility with the funders to support studies with sufficient budgets to obtain adequately powered samples—not to speak of adequate sampling procedures and representativeness. On the other hand, the funders are in a strong position to mandate *a priori* power analyses, pre-registrations, strong reporting standards geared towards subsequent research synthesis, published datasets, and open-access reports. They could, furthermore, incentivize and support the creation registered-study databases to counter the file-drawer problem.

Acknowledgment. We would like to thank the anonymous reviewers of STAST 2020 for their comments. Early aspects of this study were 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.” Thomas Groß was funded by the ERC Starting Grant CASCade (GA n°716980).

A Sample Characteristics

We offer a detailed sample refinement on papers in Fig. 2 and extracted statistical tests and corresponding effect sizes in Fig. 3.

Table 2. Sample refinement on papers

Phase	Excluded	Retained
<i>Source SLR [6]</i>		
Search results (Google Scholar)	–	1157
Inclusion/Exclusion	1011	146
<i>Refinement in this study</i>		
Empirical studies	2	144
With sample sizes	21	123
With extractable tests	69	54

Table 3. Sample refinement on extracted effect sizes

Phase	Excluded	Retained
Total effects extracted	0	650
statcheck automated extraction		252
Test statistic manual coding		89
Means & SD manual coding		309
<i>Refinement in this study</i>		
Independent-samples test on dependent sample	46	604
Treated proportion as t -distribution	8	596
Reported dependent-samples test w/o correlation	62	534
Reported χ^2 without df	5	529
χ^2 with $df > 1$ without contingency	72	457
Multi-way F -test	22	435
Yielded infinite ES or variance	3	432
Duplicate of other coded test	1	431

Table 4. Descriptives of the observed absolute effect sizes as log odds ratios.

Statistic	n	Min	Q:25	Mdn	Q:75	Max	IQR	M	SD
Chi2	102	0.07	1.05	1.38	1.74	3.73	0.69	1.40	0.90
F	42	0.04	0.31	0.73	1.57	7.71	1.26	1.43	1.81
r	6	−1.03	1.02	1.33	1.86	2.15	0.84	1.40	0.60
t	317	−4.58	0.21	0.43	0.99	8.35	0.78	0.85	1.18
Z	33	0.07	0.31	0.66	1.59	7.17	1.29	1.37	1.57

References

1. American Psychological Association (ed.): Publication Manual of the American Psychological Association (6th revised ed.). American Psychological Association (2009)
2. Begg, C.B., Mazumdar, M.: Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50**, 1088–1101 (1994)
3. Button, K.S., et al.: Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**(5), 365–376 (2013)
4. Cohen, J.: A power primer. *Psychol. Bull.* **112**(1), 155 (1992)
5. Cohen, J.: Statistical power analysis. *Curr. Dir. Psychol. Sci.* **1**(3), 98–101 (1992)
6. Coopamootoo, K., Groß, T.: Systematic evaluation for evidence-based methods in cyber security. Technical Report TR-1528, Newcastle University (2017)
7. 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
8. Coopamootoo, K.P., Groß, T.: A codebook for experimental research: the nifty nine indicators v1.0. Technical report TR-1514, Newcastle University (November 2017)
9. 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
10. Cumming, G.: *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge (2013)
11. Dickersin, K.: The existence of publication bias and risk factors for its occurrence. *Jama* **263**(10), 1385–1389 (1990)
12. Epskamp, S., Nuijten, M.: *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)
13. Gardner, M.J., Altman, D.G.: Confidence intervals rather than *p* values: estimation rather than hypothesis testing. *Br. Med. J. (Clin. Res. Ed.)* **292**(6522), 746–750 (1986)
14. Groß, T.: Fidelity of statistical reporting in 10 years of cyber security user studies. In: Groß, T., Tryfonas, T. (eds.) *STAST 2019. LNCS*, vol. 11739, pp. 3–26. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-55958-8_1
15. Groß, T.: Fidelity of statistical reporting in 10 years of cyber security user studies [extended version]. Newcastle University (2020). arXiv Report [arXiv:2004.06672](https://arxiv.org/abs/2004.06672)

16. Groß, T.: Statistical reliability of 10 years of cybersecurity user studies [extended version]. Newcastle University (2020). arXiv Report [arXiv:2010.02117](https://arxiv.org/abs/2010.02117)
17. Hoekstra, R., Morey, R.D., Rouder, J.N., Wagenmakers, E.J.: Robust misinterpretation of confidence intervals. *Psychon. Bull. Rev.* **21**(5), 1157–1164 (2014)
18. Howell, D.C.: *Fundamental Statistics for the Behavioral Sciences* (8th international ed). Cengage Learning (2014)
19. Ioannidis, J.P.: Why most published research findings are false. *PLoS Med.* **2**(8), e124 (2005)
20. Light, R.J., Pillemer, D.B.: *Summing Up; The Science of Reviewing Research*. Harvard University Press, Cambridge (1984)
21. 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
22. Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., Wagenmakers, E.J.: The fallacy of placing confidence in confidence intervals. *Psychon. Bull. Rev.* **23**(1), 103–123 (2016)
23. Peisert, S., Bishop, M.: How to design computer security experiments. In: Fletcher, 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
24. Rosenthal, R.: The file drawer problem and tolerance for null results. *Psychol. Bull.* **86**(3), 638 (1979)
25. Scargle, J.D.: Publication bias (the “file-drawer problem”) in scientific inference. arXiv preprint physics/9909033 (1999)
26. 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>
27. Sterling, T.D., Rosenbaum, W.L., Weinkam, J.J.: Publication decisions revisited: the effect of the outcome of statistical tests on the decision to publish and vice versa. *Am. Stat.* **49**(1), 108–112 (1995)
28. Sterne, J.A., Egger, M.: Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *J. Clin. Epidemiol.* **54**(10), 1046–1055 (2001)
29. Terrin, N., Schmid, C.H., Lau, J., Olkin, I.: Adjusting for publication bias in the presence of heterogeneity. *Stat. Med.* **22**(13), 2113–2126 (2003)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.