Cyber Security and Privacy Experiments: A Design and Reporting Toolkit

Kovila Coopamootoo, Thomas Gross

Cyber Security & Privacy Experiments: A Design & Reporting Toolkit

Kovila P.L. Coopamootoo and Thomas Groß

Newcastle University,
Newcastle upon Tyne, United Kingdom
{kovila.coopamootoo|thomas.gross}@newcastle.ac.uk

Abstract. With cyber security increasingly flourishing into a scientific discipline, there has been a number of proposals to advance evidence-based research, ranging from introductions of evidence-based methodology [8], proposals to make experiments dependable [30], guidance for experiment design [38,8], to overviews of pitfalls to avoid when writing about experiments [42]. However, one is still given to wonder: What are the best practices in reporting research that act as tell-tale signs of reliable research.

We aim at developing a set of indicators for complete reporting that can drive the quality of experimental research as well as support the reviewing process.

As method, we review literature on key ingredients for sound experiment and studied fallacies and shortcomings in other fields. We draw on lessons learned and infuse them into indicators. We provide definition, reporting examples, importance and impact and guiding steps to be taken for each indicator.

As results, we offer a toolkit with nine systematic indictors for designing and reporting experiments. We report on lessons and challenges from an initial sharing of this toolkit with the community.

The toolkit is a valuable companion for researchers. It incites the consideration of scientific foundations at experiment design and reporting phases. It also supports program committees and reviewers in quality decisions, thereby impacting the state of our field.

1 Introduction

Cyber security and privacy are both exciting fields that weave together methodologies, theories and perspectives from various disciplines: mathematics, engineering, law, psychology and social sciences. As consequence, it gains a collective tapestry, a definite strength that exemplifies inter-disciplinary fields. However, the sharing of expertise and drawing on best practices of each discipline is a challenge. For example, the research area of human factors of cyber security and privacy is inter-disciplinary research area. It clearly benefits from the systematic design and reporting standards characteristic of the rigorous methodology of experimental psychology at its best.

Without guidelines, we rely on researchers to assess the quality of their designs and reporting. We ask program committees and reviewers to make best decisions on submissions to their best judgment. At the same time, these submissions impact the future of the field, may sow uncertainty in the research community and among policy makers alike, especially when they intend to transfer research findings into practice.

Workshop at IFIP Privacy & Identity Management Summerschool 2017. Our workshop on Evidence-Based Methods was intended as a first evaluation of a set of indicators we originally developed for a systematic literature review in the Research Institute in Science of Security (RISCS). We offered a presentation of each of the indicators and their specifications and offered participants a codebook and a marking sheet [9] as well as publications reporting experimental privacy studies.

Contribution. This paper aims to offer support for experimental cyber security and privacy research as scientific discipline. It provides nine clear guidelines to support the design and reporting of experiments and discusses challenges to dissemination from a first encounter with the community.

Outline. In the rest of the paper, we discuss our choice for the set of indicators before detailing each of the nine completeness indicators. For each indicator we proceed with theoretical background, benefits for fulfilling the indicators, outcome of not achieving them, practical steps to take in design and reporting, together with best practice examples. We then provide the lessons learnt from a first connection with the community before providing the discussion and conclusion.

2 Choice of Completeness Indicators

We chose indicators that contribute and build-up towards sound statistical inference. As a consequence, we addressed reproducibility, internal validity, correct statistical reporting and parameter estimation. We deliberately excluded criteria on external validity and ethics, but may consider them in future versions of the toolkit.

Benefits of this first toolkit. The indicators are designed as a toolkit mainly for researchers, providing both a theoretical and a practical component. *First*, it acts as support for the design phase of a user study/experiment and aims to be one-stop resource. We provide a theoretical background with each indicator, substantiated with reasoning in the form of benefits for having the indicators and the outcome of not catering for them. *Second*, it acts as a companion for reporting via the practical steps to take, typical locations in articles and examples of good practice. Further, an additional benefit, is a clear list that can enable program committees to evaluate the reporting of research studies.

3 CI1: Upstream Replication

3.1 Theoretical Background

Similar to Coopamootoo & Groß [8], we call *replication* the attempt to recreate the conditions sufficient to obtaining a previously observed finding, a definition adapted from the Open Science Collaboration [7]. We refer to *upstream replication* when a study replicates existing studies or previously validated methods/instruments.

We note that replication of studies is an important research practice that provides confidence in the findings, where Cumming & Calin-Jageman [12] point out that rarely, if ever can a single finding give definitive answer to a research question, while the *Open Science Collaboration* notes the alarming discovery that a number of widely known and accepted research findings cannot be replicated [12].

We ask: *'Is the study reporting existing studies or methods?'*

Benefits of fulfilling CI1. Researchers engaging in upstream replication are gaining sound foundations for their studies in employing methods whose exact properties are known and well-tested. For instance, for a measurement instrument we expect it to be known, which parameters of the population are measured. We expect of the instrument itself internal validity, repeatability and reproducibility. In the logic of the statistical inference of the given experiment we are then entitled to assume that the properties of the instrument are a given and will be the same for other researchers in the future. As a completeness indicator, CI1 thereby yields evidence whether the foundations of the given reported study are sound.

Outcome for not fulfilling CI1. Should evidence towards CI1 be missing, we would need to assume that the study did not pay attention to its sound foundations. This, in turn, means that the study is on uncertain footing. For the manipulation instruments, it is not assured that they cause the intended change in the participants reliably. For the measurement instruments, it is not assured that they measure the intended property. Consequently, instruments without evidence of sound *a priori* validation yield sources of errors that can well confound the main experiment and thereby put the overall inference in question.

3.2 Steps to Take

How to achieve CI1. The key principle towards gaining evidence for CI1 is the use of validated tools. We recommend to select manipulation and measurement instruments that come with strong evidence of their validation and properties.

In the experiment design and execution, researchers will employ instruments *exactly* as validated, for example, using the defined scale and scoring sheet as provided. If adaptations are made to the instrument instructions, these will be documented.

In the reporting of the study, researcher will then document the evidence for their exact replication, for instance, by including the exact materials used and by citing the validation study they rely upon.

We note that documentation of instruments also apply to those employed for manipulation checks.

Typical location in articles. CI is reported in the methods section with subsections on measurement apparatus and manipulation apparatus or experimental conditions.

Reporting Example.

> *Example 1 (CI1 - Manipulation Apparatus, with amendments from Nwadike et al. [37]).*
> We induce a happy and sad affect via video stimulus, a mood induction protocol recommended by Westermann's critical review of different methods [47]. For happiness affect we used the restaurant scene from the movie *When Harry meets Sally* [clip length 155 seconds] while for sadness affect we used the dying scene from the movie *The Champ* [clip length 171 seconds]. We refer to Rottenberg et al. [40] to start and end the clips at the exact frames as previously validated.

> *Example 2 (CI1 - Manipulation Check, with amendments from [37]).* We used the 60-item full PANAS- X questionnaire [46] as manipulation check on the induced affect state. We focus on *sadness* and *joviality* as equivalent of happiness. The PANAS-X is scale is based on 5-point Likert-items anchored on 1 - very slightly or not at all, 2 - a little, 3 - moderately, 4 - quite a bit, and 5 - extremely. We anchored PANAS-X for affect "at the present moment."

> *Example 3 (CI1 - Validated Measurement Apparatus).*
> "The State-Trait Anxiety Inventory for Adults (STAI-AD) [43] is a 40-question self-report questionnaire. We use the temporary construct of state anxiety, that is, "how you feel right now." It employs 4-point Likert items anchored on 1 – Not At All, 2 – Somewhat, 3 – Moderately So, and 4 – Very Much So."

3.3 Further Sources

Across sciences, a replication crisis has been observed. Prominently in psychology, a large scale replication endeavor by the Open Science Collaboration [7] of $N = 100$ studies across 3 psychology journals found that only 47% of the original effect sizes were in the 95% confidence interval of the replication effect size.

The Open Science Collaboration makes a case that research claims gain credibility when the supporting evidence undergoes sound replication [7]. We note

that the replication needs to be done deliberately to increase the overall Positive Predictive Value of the results [25,34].

In security literature, Maxion [30] postulates that repeatability, reproducibility and validity are the main criteria differentiating a well designed experiment from those that are not.

4 CI2: Reproducibility

4.1 Theoretical Background

CI2 considers the enablement of *downstream replication*. While downstream replication includes *repeatability*, that is, whether a study can be replicated by the same researchers, CI2 considers especially, whether the study is sufficiently reported to be *reproducible* by *other* researchers. We refer to Maxion [30] for further discussion on repeatability and reproducibility.

CI2 establishes whether the reporting supports reproducibility, defined as the closeness of results obtained on the same test material under "changes of [...] conditions, technicians, apparatus, laboratories and so on" [13]. A key requirement of replicating existing studies is the availability of clear documentation.which ideally would entail a detailed step-by-step experimental protocol, which makes provisions for reproducibility.

The principle for reproducibility is diligent documentation of all variables of the study's lifecycle. We ask *'Is there correct reporting of manipulation apparatus, measurement apparatus, detailed procedure, sample size, demographics, sampling and recruitment method, contributing towards reproducibility?'*

Benefits of fulfilling CI2. Offering sound reporting for reproducibility allows for downstream replication and contributes to the enablement of research synthesis in a field. This is crucial to enable falsification and hence empirical progress. Having a reproducible study at hand means that other researchers can test the theories evaluated in the given study and establish independent evidence on the theories, possibly falsifying the earlier result. Furthermore, replication studies inform the overall positive predictive value for the considered relations and allow for a meta analysis on the effect sizes and their confidence intervals.

Hence, as completeness indicator, CI2 checks whether evaluates whether the theories named in the given study can be empirically scrutinized in subsequent experimentation from the given reporting, and thereby whether the given study makes a sound contribution to empirical sciences.

Outcome for not fulfilling CI2. Should the evaluation for CI2 not offer evidence towards reproducibility, we need to assume that the given study cannot be replicated downstream. First, the lack of reproducibility leaves other researchers with a great ambiguity what was actually done. Second, following Poppers discussion on falsifiability [39], a study that cannot be reproduced does not actually yield strong empirical evidence because other researchers cannot execute the offered experiment to falsify the reported theory, which in turn casts doubt on the study advancing empirical knowledge.

4.2 Steps to Take

How to achieve CI2. Researchers will provide detailed description of experiment design, including the all choices made, possibly supplemented by an experiment diagram, as well as the procedure executed in the experiment itself.

We note that documentation towards reproducibility will often also include planned analyses, which we consider under other CIs. A recommended practice in this case is to pre-commit the experiment and analysis plan at organizations such as the Open Science Framework[1] or AsPredicted[2]. As example, committed analysis plan and analysis report [22] published for password research [17].

Typical location in articles. C2 covers the whole method section including a detailed procedure, sample recruitment and demographics, manipulation and measurement instruments. Planned analysis will be in the analysis or the results section.

Reporting Example.

Example 4 (CI2 - Demographics).
We refer to Table 3 of Kluever and Zanibbi [27] for a detailed demographics report that is relevant to the context of the study reported.

Example 5 (CI2 - Measurement Apparatus precisely referencing sources).
"We administered the NASA Task Load Index in an online form. The form exactly replicated the full NASA TLX questionnaire as specified on in *NASA Task Load Index* (TLX), v. 1.0, Appendix, pp. 13. [24]"

Example 6 (CI2 - Procedure).
"The procedure consisted of (i) pre-task questionnaires for demographics and personality traits, (ii) a manipulation to induce cognitive depletion, (iii) a manipulation check on the level of depletion, (iv) a password entry for a mock-up GMail registration, and (v) a debriefing and memorability check one week after the task with a GMail login mockup. " This was followed with a details of each section.

4.3 Further Sources

First, for reproducibility of the experiment design, which is what this CI mainly focuses on, we refer to experiment design methodology [16,31,33].

Second, for reproducibility of the planned analyses, which involves the documentation of the plan as well as the recording of all the analyses done, we suggest inspiration from reproducibility principles from general computing science

[1] `https://osf.io`
[2] `https://aspredicted.org`

research [41] or more specific sources with focus on computation-supported scientific practice [45]. To render all computations, statistical analyses and graphs reproducible, we suggest the R framework knitr [48] as demonstrated within the analysis report [22].

5 CI3: Internal Validity

5.1 Theoretical Background

CI3 addresses internal validity of the experiment, which refers to the truth that can be ascribed to cause-effect relationships between independent variables (IV) and dependent variables (DV) [3], where the IV is a variable that is induced/manipulated and the DV is the variable that is observed/measured [32].

This CI asks for research questions and hypotheses that provide the foundations for null hypothesis significance testing (NHST) [36]. Operationalization enables systematic and explicit clarification of the predictors or independent variables, and hence the cause and manipulation, while the target variable or dependent variables clarify the effect, hence the measurements. Subject assignment points to whether and how participants were randomly assigned and balanced across experimental conditions.

Manipulation check refers to verification that the manipulation has actually taken effect, hence assuring systematic effects.

We ask *'Is there an explicit and operational specification of the RQs, null and alternative hypotheses, IVs, DVs, subject assignment method and manipulation checks?'*

Benefits of fulfilling CI3. CI3 ensures internal validity and a solid statement of intention for Null Hypothesis Significance Testing (NHST) [36].

Outcome for not fulfilling CI3. Should evidence for CI3 be missing, we would need to assume other possible explanations for the cause-effect relationship investigated, that is that the reported design could involve variables contributing unsystematic effects. This in turn would mean that other researchers could not rely on the results reported.

5.2 Steps to Take

How to achieve CI3. We propose in the first instance that following the step by step exercise we previously detailed [8] on 'An Exercise in Experiment Design' to be beneficial for internal validity. In particular, developing research questions, defining testable hypotheses, operationalizing hypotheses into IVs and DVs. For IVs, reseachers will answer 'What factor is being manipulated and influences the outcome?' For DVs, 'What is being measured?' and how can we measure the outcome of manipulation reliably.

Typical location in articles. The aims section can detail the research questions and hypotheses where as the method to include sub-sections on operationalizing the variables into measures and experimental conditions. The method section will also include subject assignment information.

Reporting Example.

> *Example 7 (CI3 - Research Question, from Cherapau et al. [5]).*
> *"How availability of Touch ID sensor impacts users' selection of unlocking authentication secrets?".*

> *Example 8 (CI3 - Hypotheses, from Cherapau et al. [5]).*
> For null hypotheses H_0: *"Use of Touch ID has no effect on the entropy of passcodes used for iPhone locking."* or *"Availability of Touch ID has no effect on ratio of users who lock their iPhones."*
> For corresponding alternative hypotheses H_1: *"Use of Touch ID affects the entropy of passcodes used for iPhone locking."* or *"Availability of Touch ID increases the ratio of users who lock their iPhones"* [5].

> *Example 9 (CI3 - Subject Assignment, amended from Bursztein et al. [4]).*
> *"Our task scheduler presented the CAPTCHAs to Turkers in the following way ... Random Order - fully random, where any captcha from any scheme could follow any other."*

We also refer to Example 2 for manipulation checks.

6 CI4: Limitations

6.1 Theoretical Background

CI4 establishes what other factors could affect the cause and effect relationship under investigation and hence limit validity including both internal and external validity. This CI is related to the requirement of controlled variables for experiment design, that is the assurance that an observed change in the dependent variable is a result of a systematic change in the independent variable [32].

We ask *'Was there a discussion on the limitations, possible confounders, biases and assumptions made?'*

Benefits of fulfilling CI4. CI4 provides transparency of validity and assurance that other possible explanations for the stated causal relations, have been considered. This in turn provides confidence in the reported results.

Outcome for not fulfilling CI4. Should the limitations not have been discussed in the experiment report, we would need to assume that the researchers might have failed to control variables that impact the internal validity of the experiment. This puts the reported effects into question.

6.2 Steps to Take

How to achieve CI4. Researchers are (1) to evaluate experimental designs for alternative explanations that could influence the observed effects, such as identifying confounding and controlling for variables, (2) to make explicit the boundaries and of the design, such as whether a convenient sample was used, and (3) acknowledge the limits in interpretations that can be inferred from the findings, such as whether the results are a correct reflection of estimates for the general population.

A discussion of the limits and boundaries of the study, identification of possible confounding variables whose presence affect the relationship under study, and possible assumptions made in setup, are all valuable inputs that strengthen the validity of the experiment.

Typical location in articles. While researchers may report and discuss limitations throughout the article, it is preferred to define a dedicated limitations section, that shows clarity and researcher awareness of the limits of their design.

Reporting Example.

> *Example 10 (CI4 - Sampling bias, from Akhawe & Felt [1]).*
> *"The participants in our field study are not a random population sample. Our study only represents users who opt in to browser telemetry programs. This might present a bias. The users who volunteered might be more likely to click through dialogs and less concerned about privacy. Thus, the clickthrough rates we measure could be higher than population-wide rates."*

7 CI5: Reporting Standard

7.1 Theoretical Background

Statistical reporting guidelines helps the reader, reviewer, policy maker to gain confidence in the reported statistical analysis and results. As example, we propose reporting recommendations of the American Psychology Association (APA) [2] as quality standard.

We ask *'Was the result reported in the APA style?'*

Benefits of fulfilling CI5. Reporting standards provide a degree of comprehensiveness in the information that is reported for empirical investigations. Uniform reporting standards make it easier to generalize within and across fields, to understand implications of individual studies and supports research synthesis. Comprehensive reporting also supports decision makers in policy and practice towards understanding how the research was conducted [2].

Outcome for not fulfilling CI5. The impact of not fulfilling CI5 opens gaps and lead to questioning research quality, reuse and reproducibility.

7.2 Steps to Take

How to achieve CI5. Researchers are to closely adhere to statistical reporting standards such as the APA [2] and reporting statistical inference as recommended whether in paragraphs, tables or figures. This include reporting actual p-values, that is not only whether the p-value is less that α, and effect sizes and confidence intervals.

Typical location in articles. Reporting standards usually focus on the specification of the results section, yet can also indicate the format of a structured abstract or the structure of the overall publication.

Reporting Example.

> *Example 11 (CI5 - with amendments from Coopamootoo et al. [10]).*
> *We computed a one-way ANOVA. "There was a statistically significant effect of the experiment condition on the password strength score,* $F(2, 63) = 6.716$, $p = .002 < .05$. *We measure the effect size* $\ldots \eta^2 = .176, 95\%\ CI\ [0.043, 0.296]$ *[. . .]."*

8 CI6: Test Statistic

8.1 Theoretical Background

The reporting on the test statistic offers a precise interface on the result of the computed statistical analysis. This data allows for a future analysis of *a posteriori* likelihoods, such as in a Positive Predictive Value (PPV) [25]. Simply put, this data helps other researchers to ascertain whether the result could be a false positive or not.

We consider the precise documentation of the outcome of the statistical test. For instance, for a t-test we would expect to learn the t-value as well as the degrees of freedom, along with the exact p-value computed for this t.

We ask *'Did the result statement include test statistic and p-value?'*

Benefits of fulfilling CI6. If the test statistic is fully specified, we gain important data for the subsequent analysis of the result. From the consistency of the reported test statistic and the p-value, we gain confidence in the correct reporting. In addition, the data includes sufficient redundancy that others can validate the presented p-values or use the reporting of the test statistic to compute standardized effect sizes for subsequent meta-analysis.

Outcome for not fulfilling CI6. Should the test static or the p-values not be reported, e.g., by just stating that the result "is statistically significant, $p < .05$, we lose a lot of information. We could neither ascertain the confidence level of the significance nor the internal consistency of the reported test. Hence, the reported result will lack internal credibility and not be particularly trustworthy.

8.2 Steps to Take

How to achieve CI6. The key principle is to report sufficient data, such that others can cross-check the reported values and use them in further research synthesis. Usually, this involves reporting the test statistic itself, the degrees of freedom vis-à-vis of the sample size, and the exact p-value. When comparisons between conditions are made, then the descriptive statistics for the relevant conditions should be provided (e.g., mean and standard deviation for conditions of a t-test).

Typical location in articles. The test statistics will be specified in the results section of the paper. As a rule of thumb, for each result we claim as being statistically significant, we will provide the test statistic supporting that claim as suffix.

Reporting Example.
We refer to the Example 11 for test statists and p-value reporting.

9 CI7: Assumptions

9.1 Theoretical Background

Statistical tests can easily lead us astray if their assumptions are not fulfilled: they may produce spurious results. Even though some tests have been shown to be somewhat robust against borderline violations of their underlying assumptions, the burden of proof that the assumptions were sufficiently fulfilled is on the researchers who conducted the test.

In general, the exact type of test in a family needs to be specified to inform which assumptions come to bear. For instance, the assumptions of an *independent-samples* t-test will be different from a *dependent-samples* t-test. Similarly, it needs to specified whether the test is "one-tailed" or "two-tailed" to put the reported p-values into perspective.

To ascertain whether the statistical analyses were correctly employed on the data, statistical assumptions need to be made explicit in reporting. For example, the assumptions for parametric tests, in general, are normally distributed data, homogeneity of variance, interval data and independence [15]. Parametric statistical tests often require a systematic treatment of outliers.

We ask *'Were significance level α and test statistics properties and assumptions appropriately stated?'*

Benefits of fulfilling CI7. A precise specification of the test used and explicit documentation of the assumptions checked gives the reader confidence that the statistical tools were appropriately chosen and employed diligently.

Outcome for not fulfilling CI7. Should test properties and assumptions not be documented, we need to assume that researchers did not establish that they could reliably employ the statistical test. Consequently, the reported test statistics and p-values could be off and not be relied upon.

9.2 Steps to Take

How to achieve CI7. One would choose the designated significance level *a priori* and state it explicitly. Similarly, the researchers need to establish whether the test will be one- or two-tailed in advance. Researchers check whether the data meets the assumptions of the planned statistical test and explicitly report whether and how the data met the test assumptions. Decisions on how the data was treated (e.g., outlier management) need to be reported explicitly.

We emphasize that complex statistical models (such as regressions) usually require comprehensive post-hoc model diagnostics to evaluate whether the model is sound.

Typical location in articles. The treatment of assumptions is documented in the results section, either close to the report of the statistical test or in a separate subsection. Often it will support the confidence in the reported results, if a comprehensive analysis report is published alongside the research paper that documents all checks of assumptions and diagnostics, transformations of the data, and decisions made.

Reporting Example.

> *Example 12 (CI7 - Significance level α & test statistics properties, from Groß et al. [23]).*
> "All inferential statistics are computed with two-tailed tests and at an α level of .05"

Example 13 (CI7 - Test statistics assumptions, from Groß et al. [23]).
"The distribution of the Passwordmeter password strength score is measured on interval level and is not significantly different from a normal distribution, Saphiro-Wilk, $D(100) = .99$, $p = .652 > .05$"[a].
"We computed Levene's test for the homogeneity of variances. For the password meter scores, the variances were not significantly unequal."

[a] We note here that numerical normality tests, such as Saphiro-Wilk may have too little sensitivity for small sample sizes and too much sensitivity for large sample sizes. [44]

10 CI8: Confidence Intervals on Effects

10.1 Theoretical Background

An effect size estimates the magnitude of an effect, an unknown parameter of the population, given the observed data of an experiment. Confidence interval procedures on the effect estimate the range of plausible values for the population parameter, if the experiment were repeated independently infinitely many times. We note that this is a frequentist view, in which the confidence level applies to the procedure. For instance, a series of 95% confidence intervals will tend to contain the population parameter on average 95% of the intervals.

Effect sizes and their confidence interval offer an informative view on an experiment's observed effect magnitudes. Consequently, the APA guidelines [2] state that "estimates of appropriate effect sizes and confidence intervals are the minimum expectations." QI8 includes that the effect sizes are reported in a easily human-interpretable form.

We ask *'Were the appropriate the effect sizes and confidence intervals (CI) reported?'*

An effect that is statistically significant is not necessarily scientifically significant or important. To draw conclusions on an effect's importance or practical implications, we consult the magnitude of the effect, its effect size. [6].

In estimation theory, the effect size (ES) provides a *point estimate* of effect in the population, while the confidence interval (CI) provides the interval estimate. While we endorse the use of estimation theory [12,19], we note that interpreting confidence intervals correctly requires diligence [35]. Notably, it is a fallacy to interpret a post-data X% confidence interval to have a X% probability to include the true population parameter.

Benefits of fulfilling CI8. CI8 evaluates the robust reporting of effect magnitudes through parameter and interval estimation, which yields, in turn, the foundation for future meta-analysis and research synthesis.

Outcome for not fulfilling CI8. Without effect size estimate, we only have the significance of the results and $p-$values to go on. However, we will miss out on information on the magnitude of the claimed effects. For example a significant p-value does not say how important the observed effect is: it could well be trivial, and neither contribute much to research nor vouch for changes to practice.

10.2 Steps to Take

How to achieve CI8. To compute effect sizes in experiments together with their confidence intervals and to report these in publications. Literature already provides a number of manuals and research articles on computing the different families of effect sizes [18,29] To also refer to the *New Statistics* [12] for the estimation approach, effect-size and confidence intervals.

Typical location in articles. Effect sizes and their confidence intervals are documented in the results section, either stated as a suffix after the p-value of the corresponding statistical inference or provided in dedicated tables.

Reporting Example.
We refer to the Example 11 for effect size and confidence interval reporting.

10.3 Further Sources

Kirk [26] and Cumming [11] debated that the current research practice of exclusive focusing on a dichotomous reject-nonreject decision strategy of null hypothesis testing that can impeded scientific progress. Rather, they posit, the focus should be on the magnitude of effects, that is the practical significance of effects and the steady accumulation of knowledge. They advise to switch from the much disputed NHST to effect sizes, estimation and cumulation of evidence.

11 CI9: Statistical Inference

11.1 Theoretical Background

CI9 evaluates the overall correctness of the statistical inference, that is, how statements on statistical significance are expressed and what conclusions are drawn from the statement. As such, CI9 relies to some extent on observations made with respect to preceding completeness indicators.

We ask *'Was the significance and hypothesis testing decision interpreted correctly and put in context of effect size and sample size/power?'*

Nickerson [36] offers a comprehensive overview of the controversies around *Null Hypothesis Significance Testing (NHST)*, while Maxwell and Delaney [31, p.48] and Goodman [21] point to p-Value misconceptions, Morey et al. [35] analyze confidence interval fallacies and Ioannidis [25] argues "why most published research findings are false."

The evaluation in our work is founded on Nickerson's review [36] on misconceptions around NHST, which include:

- p misperceived as the probability that the hypothesis be true and $1 - p$ misperceived as the probability that the alternative hypothesis be true,
- a small p considered as evidence that the results be replicable,
- a small value of p misinterpreted as a treatment effect of large magnitude,
- statistical significance considered as theoretical or practical significance,
- significance level α misinterpreted as the probability that a Type I error will be made,
- Type II error rate β considered to mean the probability that the null hypothesis be false,
- failing to reject the null hypothesis misrepresented as equivalent to demonstrating it to be true,
- failure to reject the null hypothesis misinterpreted as evidence of a failed experiment.

While Nickerson's observations are concerned with the correct interpretation of NHST, for us CI9 also includes preparing the ground with population and sampling as well as *a priori* hypothesis specification, and post-hoc concerns such as multiple-comparison corrections.

Benefits of fulfilling CI9. Evidence towards CI9 convinces us of the robustness and diligence of the statistical inference made, because common pitfalls and fallacies have been avoided. The result statement will offer a sound starting point for the interpretation of the findings.

Outcome for not fulfilling CI9. Should there be evidence of incorrect statistical inference or the presence of fallacies, we would need to assume that the researchers interpretation of said results be tainted by the misinterpretations and misrepresentations made. Hence, the overall conclusion of the study would be put into question. We perceive reviews on misconceptions and fallacies as important guard rails [36,31,21,35].

11.2 Steps to Take

How to achieve CI9. To achieve CI9, we recommend to investigate how p-values and confidence intervals can and cannot be interpreted. The key principle here is diligence: The devil is in the details.

Typical location in articles. The correctness of the statistical inference is prepared by the documentation of the *a priori* elements of a study in the methods section, supported by the correct reporting of statistical tests in the results and finally completed by the interpretation of the outcomes in the discussion.

Reporting Example.

> *Example 14 (CI9 - Type I error correction).*
> *"Given the number of comparative* t-tests *computed on the data set, we compute a multiple comparisons correction, where differences marked with a dagger † in Table 1 are statistically significant under Bonferroni-Holm correction for all comparisons made."*

12 Lessons Learnt from the workshop

12.1 Aim

To assess whether and how the set of nine indicators could be applied in practice.

12.2 Method

Procedure. We gave a small presentation of the hallmarks of experiment design (following our 2016 workshop at the same venue) and then presented the nine indicators as a set of 'Quality Indicators', where quality assessment is a stage employed within Systematic Literature Review procedures [14]. These nine indicators were developed as a checklist of factors to be evaluated within experimental studies, as part of a UK Research Institute in Science of Cyber Security (RISCS) funded project, which had the overall aim to evaluate the state of the art in evidence-based methods in cyber security and privacy.

Prior to the workshop we developed a first version of a codebook which specified each of the indicators in terms of sub-criteria and examples and a codesheet providing a marking scheme.

Next, we facilitated open coding with the aim to *extract concepts* from the free-form text. We provided participants with (1) two example research articles reporting user experiments in the context of privacy [20,28], (2) the CI specification as a codebook [9] and (3) the marking as a codesheet [9].

Participants. Participants worked in two groups to review the two articles. $N = 9$ participants attended the workshop, 6 female, 3 male. The 7 participants who provided their age had mean age 31.86 years ($SD = 8.28$). 5 participants were from a usable privacy and security background, while others were from other areas of privacy and security. Participants' first language varied (4 German, 2 English and 1 Tamil, 2 did not answer). With the sole aim to gauge participants' expertise, we offered participants three Likert questions to rate their frequency of use of evidence-based methods (from 1 – 'Never' to 5 – 'A great deal'), their skills (from 1 – 'Poor' to 5 – 'Excellent') and their familiarity (from 1 – 'Not at all familiar' to 5 – 'Extremely familiar') in designing experiments. Participants reported using evidence-based methods such as experiments with a median value of 3, to have a median skill level of 2 and median familiarity in designing experiments of 2.

12.3 Results

We provide results in the form of participant feedback and recommendations.

Practical Requirement. Participants recommended shaping of the indicators as a toolkit that can readily be employed by the community. This involves designing clear sections in the tool set that researchers and committee members can pick up. As a result, following the workshop, we have revised the indicators to match these requirements, as presented through sections CI1 to CI9. In this paper we provided the theoretical underpinnings for each CI together with 'Steps to Take' and 'Examples'.

Design Requirement. Participants noted the time commitment required if one does not know what to look for when applying the toolkit in a reviewing exercise. To address this, we provide clear examples for each CI together with typical sections in research papers that provide support for criteria fulfilling each CI.

Ethical Considerations. Participants suggested to factor in ethical considerations, as aspect of experimental reporting we omitted but foresee its benefits for completeness of reporting.

13 Discussion

Community progress. A toolkit, such as the one we provide here, contributes to a standard to aspire to. It supports the community in developing the skills to design, run and report rigorous experiments in cyber security. At the same time, while the lack of defined best practices requires individual researchers to determine what the standards they adhere to, our toolkit offers a common ground.

It also supports the reviewing process and program committee decisions, by offering syntactic criteria to check for the completeness of scientific reporting. In addition, it contributes to a culture of well designed and reported experiments that can serve as notable examples to follow in the field.

Added value for researchers. We believe this toolkit can be a valuable ingredient for inter-disciplinary security and privacy research. It combines theoretical background and practical guidelines to support foundations in experiment design and reporting. By following the requirements of participants as voiced during the workshop, we provided clear sections that can be picked up by researchers and committee members. It supports both novice and experienced usable security and privacy researchers. While learning a methodology takes time for any novice, we believe that this toolkit may support the learning the nitty-gritty of experimental methodology by being designed as a one-stop resource. For more advanced researchers, it presents itself as a checklist and offers some good practices to follow.

Not exhaustive. We observe that our current toolkit is not exhaustive and foresee that it will grow as discussions advance within the community. In line with this, we plan to facilitate further discussion exercises within the community and to seek ways for engagement.

14 Conclusion

This paper provides a first toolkit for experimental research in cyber security and privacy with a sampler of theoretical foundations and practical guidelines. I can support a study's lifecycle from conception, design, analysis and reporting to replication. It provides a companion for novice researchers as well as reviewers needing a structured checklist. Although the toolkit is certainly not exhaustive, it may still grow with discussions and evidence-based projects within the community. We believe that already in the current form, it can support a culture of robustly designed and reported experiments, thereby contributing to empirical research in the field.

15 Acknowledgment

We are indebted to the the participants of the "Workshop on Evidence-Based Methods" at the 2017 IFIP Summerschool on Privacy and Identity Management for their generous feedback. This work was supported by the UK Research Institute in Science of Cyber Security (RISCS II) project "Scientific Methods in Cyber Security: Systematic Evaluation and Community Knowledge Base for Evidence-Based Methods in Cyber Security." It was in parts funded by the ERC Starting Grant CASCAde (GA n°716980).

References

1. D. Akhawe and A. P. Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *USENIX security symposium*, volume 13, 2013.
2. American Psychological Association (APA). *Publication manual*. American Psychological Association, 6th revised edition, 2009.
3. M. B. Brewer. Research design and issues of validity. *Handbook of research methods in social and personality psychology*, pages 3–16, 2000.
4. E. Bursztein, S. Bethard, C. Fabry, J. C. Mitchell, and D. Jurafsky. How good are humans at solving captchas? a large scale evaluation. In *Security and Privacy (SP), 2010 IEEE Symposium on*, pages 399–413. IEEE, 2010.
5. I. Cherapau, I. Muslukhov, N. Asanka, and K. Beznosov. On the impact of touch id on iphone passcodes. In *SOUPS*, pages 257–276, 2015.
6. J. Cohen. A power primer. *Psychological bulletin*, 112(1):155, 1992.
7. O. S. Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.
8. K. P. Coopamootoo and T. Groß. Evidence-based methods for privacy and identity management. In *Privacy and Identity Management. Facing up to Next Steps*, pages 105–121. Springer, 2016.

9. K. P. Coopamootoo and T. Groß. A codebook for experimental research: The nifty nine indicators v1.0. Technical Report 1514, Newcastle University, November 2017.
10. K. P. Coopamootoo and T. Groß. An empirical investigation of security fatigue - the case of password choice after solving a captcha. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2017)*. USENIX Association, 2017.
11. G. Cumming. The new statistics: Why and how. *Psychological science*, 25(1):7–29, 2014.
12. G. Cumming and R. Calin-Jageman. *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge, 2016.
13. B. Everitt. *Cambridge dictionary of statistics*. Cambridge University Press, 1998.
14. Evidence-Based Software Engineering (EBSE). Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report EBSE-2007-01, Keele University and University of Durham, July 2007.
15. A. Field. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.
16. A. Field and G. Hole. *How to design and report experiments*. Sage, 2003.
17. T. Fordyce, S. Green, and T. Groß. Investigation of the effect of fear and stress on password choice. In *In proceedings of the 7th ACM Workshop on Socio-Technical Aspects in Security and Trust (STAST'2017)*, 2017.
18. C. O. Fritz, P. E. Morris, and J. J. Richler. Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1):2, 2012.
19. M. J. Gardner and D. G. Altman. Confidence intervals rather than p values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, 292(6522):746–750, 1986.
20. J. Gideon, L. Cranor, S. Egelman, and A. Acquisti. Power strips, prophylactics, and privacy, oh my! In *Proceedings of the second symposium on Usable privacy and security*, pages 133–144. ACM, 2006.
21. S. Goodman. A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology*, volume 45, pages 135–140. Elsevier, 2008.
22. T. Groß. Analysis report – investigation of the effect of fear and stress on password choice. OSF Report `https://osf.io/3cd9h/`, Open Science Framework, 2017.
23. T. Groß, K. Coopamootoo, and A. Al-Jabri. Effect of cognitive depletion on password choice. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2016)*, pages 55–66. USENIX Association, 2016.
24. S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.
25. J. P. Ioannidis. Why most published research findings are false. *PLoS Med*, 2(8):e124, 2005.
26. R. E. Kirk. The importance of effect magnitude. *Handbook of research methods in experimental psychology*, pages 83–105, 2003.
27. K. A. Kluever and R. Zanibbi. Balancing usability and security in a video captcha. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, page 14. ACM, 2009.
28. S. Korff and R. Böhme. Too much choice: End-user privacy decisions in the context of choice proliferation. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 69–87, 2014.
29. D. Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, 4, 2013.
30. R. Maxion. Making experiments dependable. *Dependable and Historic Computing*, pages 344–357, 2011.

31. S. E. Maxwell and H. D. Delaney. *Designing experiments and analyzing data: A model comparison perspective*, volume 1. Psychology Press, 2nd edition, 2004.
32. S. Miller. *Experimental design and statistics*. Routledge, 2005.
33. D. C. Montgomery. *Design and analysis of experiments*. John Wiley & Sons, 8th edition, 2012.
34. R. Moonesinghe, M. J. Khoury, and A. C. J. Janssens. Most published research findings are false—but a little replication goes a long way. *PLoS Med*, 4(2):e28, 2007.
35. R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E.-J. Wagenmakers. The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, 23(1):103–123, 2016.
36. R. S. Nickerson. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, 5(2):241, 2000.
37. U. Nwadike, T. Groß, and K. P. Coopamootoo. Evaluating users' affect states: Towards a study on privacy concerns. In *Privacy and Identity Management. Facing up to Next Steps*, pages 248–262. Springer, 2016.
38. S. Peisert and M. Bishop. How to design computer security experiments. In *Fifth World Conference on Information Security Education*, pages 141–148. Springer, 2007.
39. K. Popper. *The logic of scientific discovery*. Routledge, 2005.
40. J. Rottenberg, R. Ray, and J. Gross. Emotion elicitation using films. handbook of emotion elicitation and assessment. edited by: Coan ja, allen jjb. 2007.
41. G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig. Ten simple rules for reproducible computational research. *PLoS computational biology*, 9(10):e1003285, 2013.
42. S. Schechter. Common pitfalls in writing about security and privacy human subjects experiments, and how to avoid them. *Microsoft, January*, 2013.
43. C. D. Spielberger, R. L. Gorsuch, and R. E. Lushene. Manual for the state-trait anxiety inventory. 1970.
44. L. Statistics. Testing for normality. `https://statistics.laerd.com` [Accessed 2018-01-20].
45. V. Stodden, F. Leisch, and R. D. Peng. *Implementing reproducible research*. CRC Press, 2014.
46. D. Watson and L. A. Clark. The panas-x: Manual for the positive and negative affect schedule-expanded form. 1999.
47. R. Westermann, G. Stahl, and F. Hesse. Relative effectiveness and validity of mood induction procedures: analysis. *European Journal of social psychology*, 26:557–580, 1996.
48. Y. Xie. knitr: a comprehensive tool for reproducible research in r. *Implement Reprod Res*, 1:20, 2014.