SoK: I Have the (Developer) Power! Sample Size Estimation for Fisher's Exact, Chi-Squared, McNemar's, Wilcoxon Rank-Sum, Wilcoxon Signed-Rank and t-tests in Developer-Centered Usable Security

Anna-Marie Ortloff and Christian Tiefenau, *University of Bonn;*
Matthew Smith, *University of Bonn and Fraunhofer FKIE*

https://www.usenix.org/conference/soups2023/presentation/ortloff

This paper is included in the Proceedings of the Nineteenth Symposium on Usable Privacy and Security.

August 7–8, 2023 • Anaheim, CA, USA

978-1-939133-36-6

Open access to the Proceedings of the Nineteenth Symposium on Usable Privacy and Security is sponsored by USENIX.

SoK: I Have the (Developer) Power! Sample Size Estimation for Fisher's Exact, Chi-Squared, McNemar's, Wilcoxon Rank-Sum, Wilcoxon Signed-Rank and t-tests in Developer-Centered Usable Security

Anna-Marie Ortloff
University of Bonn

Christian Tiefenau
University of Bonn

Matthew Smith
University of Bonn, Fraunhofer FKIE

Abstract

A priori power analysis would be very beneficial for researchers in the field of developer-centered usable security since recruiting developers for studies is challenging. Power analysis allows researchers to know how many participants they need to test their null hypotheses. However, most studies in this field do not report having conducted power analysis. We conducted a meta-analysis of 54 top-tier developer study papers and found that many are indeed underpowered even to detect large effects. To aid researchers in conducting a priori power analysis in this challenging field, we conducted a systematization of knowledge to extract and condense the needed information. We extracted information from 467 tests and 413 variables and developed a data structure to systematically represent information about hypothesis tests, involved variables, and study methodology. We then systematized the information for tests with categorical independent variables with two groups, i.e., Fisher's exact, chi-squared, McNemar's, Wilcoxon rank-sum, Wilcoxon signed-rank, and paired and independent t-tests to aid researchers with power analysis for these tests. Additionally, we present overview information on the field of developer-centered usable security and list recommendations for suitable reporting practices to make statistical information for power analysis and interpretation more accessible for researchers.

1 Introduction

A priori power analysis can be used to calculate the necessary sample size for a study to detect an effect with a given probability, a given significance criterion, and a defined effect size [22]. The probability is often set to 80% and the significance criterion to 5% by convention and the effect size needs to be chosen by the researcher based on their research goals [22, 30]. Using a priori power analysis, researchers can avoid running underpowered studies and missing effects that are actually present in the population and thus wasting resources or, worse yet, potentially publishing results that are misinterpreted as stating that there is no effect. It also prevents researchers from using more resources than necessary by running overpowered studies [35]. This is especially problematic for the sub-field of developer centered usable security (DCUS) [48] since developers are often both hard and expensive to recruit. Running underpowered studies in this field is especially undesirable due to the large amount of effort coupled with low chances of finding the desired effects even if they are there. Despite this, power analysis is not common in the field of DCUS.

In the 54 DCUS papers we analyzed for this SoK, only 9,3% contained any form of power analysis. When the power of studies has been assessed in other fields in meta-analyses, these have frequently shown low power, such as in a review of ACM transactions [12] or in psychology [99]. The lack of a priori power analysis is one likely reason for this. Our analysis raises similar concerns of underpowered studies in DCUS. A potential reason for so few studies being planned with power analysis is that performing a power analysis is non-trivial and researchers must know, estimate or guess key population values such as standard deviations or proportions to calculate effect sizes or estimate the effect sizes themselves. This is especially tricky in the fields of Usable Security and Privacy (USP) and DCUS, due to both the heterogeneity of the populations studied, as well as the heterogeneity of variables being measured and the many non-standardized measurement instruments. On top of that, many tests are not published with enough statistical details to use them for estimating power for similar future studies [50].

In this paper, we present a systematization of knowledge in the field of DCUS with the goal to aid researchers with

USENIX Symposium on Usable Privacy and Security (SOUPS) 2023. August 6–8, 2023, Anaheim, CA, USA

power calculations for some common statistical tests used in the field: tests with a single categorical independent variable with two groups, with either a categorical dependent variable: Fisher's exact test, chi-squared test, and McNemar's test, or with a continuous dependent variable: independent and paired t-test, as well as Wilcoxon rank-sum test and Wilcoxon signed-rank test. We collected DCUS papers from the following major conferences published between 2010 and 2021: the Symposium on Usable Privacy and Security (SOUPS), USENIX Security, the IEEE Symposium on Security and Privacy (S&P), the ACM Conference on Computer and Communications Security (CCS), the IEEE/ACM International Conference on Software Engineering (ICSE) and the security and privacy sessions from the ACM Conference on Human Factors in Computing Systems (CHI). We excluded any paper that did not contain an actual user study. This left us with a set of 54 papers. From these, we manually extracted all relevant data, including information about 467 statistical hypothesis tests involving 413 different variables. We developed a data structure to make this information accessible to researchers for specific power analyses. We further systematized the data from the papers by categorizing the involved variables into 13 different groups to make it easier to find proxies for power calculations if a direct match is not available. We also use our categories to calculate average statistics which can help researchers sanity check their estimates. The database including all the data will be made available to the community. Based on our findings, we make recommendations on how to conduct power analysis for developer studies.

2 Background & Related Work

In the following, we give a very brief overview of the research domain of developer centered usable security before discussing the background of power analysis. We also examine the application of power analysis and the practice of reporting effect sizes and conducting meta-analyses since these are closely related to power analysis.

When examining the use of statistical techniques, such as power analysis, meta-analysis, or reporting of effect sizes, in the following, we try to summarize the state of the practice as close to DCUS as possible. However, since this is a relatively new field [48] there is not yet much work providing an overview of the use of such techniques. Instead, we examine related domains, such as USP and Human Computer Interaction (HCI) in general, or when this is not possible, psychology. While these fields are by no means equal, we posit that methods and constraints are at least comparable in that user studies measuring latent variables, and not only directly observable variables, are common in all of the mentioned fields.

2.1 Developer Centered Usable Security

DCUS is a subfield of USP, but with a focus on the challenges and needs of expert users, such as software developers or administrators, instead of end users, who are at the center of typical USP-studies [3, 48]. DCUS extends USP's notion that security mechanisms should be designed with users in mind, to developers, which are themselves users, e.g., of cryptography APIs [1, 48, 73], programming languages [90] and other security tools [e.g. 9, 61, 88]. Tahaei & Vaniea provide an overview of topics addressed and methods used in DCUS [103]. While developers have been the main focus, we also include other expert users such as administrators in this field (e.g., [106].)

2.2 Theoretical Background of Power Analysis

Power analysis as a concept is situated within the null hypothesis significance testing (NHST) paradigm of statistical analysis. Four parameters are relevant to power analysis: Power, i.e., the probability of the test correctly rejecting the null hypothesis, the significance criterion α, the reliability of the sample results, and the effect size [22, 35]. The largest and invariably present influencing factor on reliability is sample size [22] - larger samples produce more consistent and reliable estimates than smaller ones. Consequently, the sample size is often used as a stand-in for reliability in power analysis.

These four parameters are interdependent, such that when three of them are available, it is possible to calculate the fourth. These calculations are referred to as power analysis. In general, there are four different kinds of power analysis, each used to determine one of the parameters from the other three [22]. The focus of this work is on so-called a priori, or prospective power analysis, which is used to calculate the necessary sample size to detect an effect of a desired size with a chosen power and significance criterion, before actually conducting the study. For a more detailed introduction to the four parameters, see Appendix C or Ellis (2010) [35].

To be able to conduct an a priori power analysis, researchers need to know or guess either a standardized effect size they expect to detect or related statistics that can be used to calculate such an effect size. These effect sizes [35] and the resulting power analysis procedures [38] vary depending on the test used. The standardized effect sizes needed for the tests at the focus of our work are ϕ for the chi-squared test, the odds ratio for McNemar's test, Cohen's d for independent samples, and Cohen's d_z for paired samples mean-comparison tests. While effect sizes can be converted between each other, what is generally seen as small, medium, or large differs between the effect size types [22], and this makes the process of guessing more difficult since not every researcher is familiar with the same types of effect sizes. Some procedures require different information, such the Fisher's exact test, where success prob-

abilities for both groups are needed instead of a standardized effect size, or McNemar's test, which requires the proportion of cases, where changes occur in subjects' responses (proportion of discordant pairs) in addition to the odds ratio effect size. Alternatively, researchers can also guess unstandardized effect sizes, such as group means for both groups, i.e., the difference between these means. This can be more intuitive, especially when taking the approach of aiming to detect the smallest practically relevant effect sizes [35]. However, additional statistics are needed to calculate the necessary standardized effect size from these values. For the tests we focus on in this work, these are the standard deviations for the two groups for independent and paired t-tests, Wilcoxon rank-sum tests, and Wilcoxon signed-rank tests. Additionally, the correlation between the two groups is necessary for within-subjects tests with continuous dependent variables. The main aim of this paper is to help and guide researchers to base these guesses on previous work or at least enable sanity checks on estimated values.

A priori power analysis is important, as both under- and overpowered studies are detrimental to the furthering of knowledge. Conducting an underpowered study means that failing to reject the null hypothesis is likely [35]. Since non-significant results are less likely to be published [7, 96], the effort in planning and conducting the underpowered study may be wasted. On the other hand, overpowered studies are wasteful, too [35]. Highly powered tests can detect very small effects so that in extreme cases, it is possible to find a highly statistically significant, albeit very small actual difference, which may be irrelevant in practice. Less power and fewer resources would have been sufficient to detect a practically relevant effect [35]. In addition to waste of resources, only collecting data from as many participants as necessary minimizes the amount of data collected, with positive effects on participants' privacy. Additionally, both underpowered and overpowered studies may be interpreted incorrectly when focusing on p-values. Underpowered non-significant results may be dismissed as irrelevant, even in the case of a large effect, while overpowered significant results representing trivial effects can be posited as important due to the statistical significance [35]. In DCUS, it is especially important to be mindful when recruiting since developers as specialists are usually time-constrained, and payment is often much higher than in end-user studies.

2.3 Application of Power Analysis

The tools to conduct power analysis have evolved from power and sample size tables [22] to online calculators[1], designated computer programs, like G*Power [38] and multiple implementations in programming languages commonly used for statistical analysis, like the `pwr` package, among others in R [91] or the `statsmodels` library in Python [98].

[1] e.g., powerandsamplesize.com or jakewestfall.org/power

Nevertheless, historically, power analysis has often not been applied [14, 22]. Unfortunately, in many fields, this is still the case according to more recent reviews, such as in psychology, where only 5% of 183 reviewed publications mentioned power analysis [113]. In other fields, often in the medical domain, power analysis is more prevalent, e.g., 43% of studies in a review of obesity interventions in schools [55] and over 60% of reviewed publications in NEJM and Lancet reported prospective power analyses [109]. A possible reason is adherence to submission guidelines [108], which we recommend updating for the field of USP. If there is enough prior information, power analysis is recommendable for grant applications to ensure sufficient funds are planned for recruitment.

In HCI, power analysis is also frequently not applied. E.g., in 2018, only five of 519 experimental papers at CHI used prospective power analysis [34]. In interviews evaluating a prototype of a program to facilitate power analyses, some researchers were explicitly skeptical of power analysis as a research tool [34]. In a more cursory evaluation of terminology used in the CHI proceedings of 2017 - 2019, rather than manual inspection of publications, only between 1.5% and 2.7% of the papers containing the term "experiment" also contained the term "power analysis" [33]. In our own analysis of USP publications at SOUPS and CHI in 2021 and 2022, we found that only 5.4% of SOUPS papers used power analysis in some form and 8.3% of USP CHI papers did so. Over these two years, only ten of 146 (6.8%) USP papers at CHI and SOUPS used power analysis in some way. Two of those papers conducted post hoc power analysis, which is controversial since power and p-values are directly related, and if a result is not significant, i.e., the p-value is high, then power was too low to detect an effect of the size present in the sample. So little is gained by the post hoc power calculation [45, 56].

Consequently, power to detect small effects in published studies is frequently low [35]. Literature shows this to be the case for such diverse research areas as management information systems, including ACM transactions [12], psychology [93, 99], and health professions education [25]. While not a formal review of power, Cockburn et al. examine empirical computer science literature and find evidence of the same practices that contributed to the replication crisis in other domains [21]. At the largest HCI conference, CHI, quantitative studies may even be underpowered to detect large effects [20].

When power analysis is not used to determine appropriate sample size, alternative approaches include recruiting the maximal number of participants possible, based on population, time, and monetary constraints [35], following prior practice and experience in the domain of interest [20, 35] or using rules of thumb, such as 10 or 15 cases of data per predictor [39], $50 + 8 \times k$, where k is the number of predictors in regression for testing the overall model [49], or two subjects per variable to estimate the coefficients in linear regression [8].

However, none of these methods guarantee that studies will have sufficient power [35].

To increase power, especially in fields like DCUS, where recruiting is challenging, researchers should also consider adapting their research design to increase the reliability of measurements and reduce random errors, e.g., by conducting within-subjects research [35]. Another possibility is conducting sequential analyses, whereby a study can be stopped at planned intervals if a large enough effect can be detected at this time, which on average reduces the necessary number of participants [67]. There is some discussion about whether stopping studies early like this introduces an additional bias towards larger effects [13, 72] or not [42, 97].

2.4 Effect sizes and Meta analysis

Reporting of effect sizes and conducting meta-analyses are topics closely related to power and power analysis. Researchers need to determine an expected effect size or a minimum relevant effect size to conduct power analyses. Meta-analyses, in contrast to (systematic) literature reviews, which provide a narrative summary of a research domain, make use of effect sizes to combine findings from different studies [35]. In conducting re-analyses with larger amounts of data, they achieve a higher power to detect effects [35].

A general lack of detail in statistical reporting was admonished in early HCI meta-analyses [79], and lack of effect size reporting is a problem, e.g., in studies investigating software engineering [62]. Groß's analysis of statistical reporting in USP showed that half of the 114 analyzed user studies from 2006–2016 reported incomplete results [50]. This makes both prospective power analysis and meta-analyses more difficult [50, 62]. In general, there are few meta-analyses in HCI [64]. As a domain using diverse analysis methods and tools derived, e.g., from computer science, design science, or psychology, there are not necessarily unified reporting standards within HCI, which makes meta-analyses difficult [104]. When reporting is not sufficient, a workaround is to ask authors for the raw data, but this comes with the drawback that not all authors will respond, e.g., Hornbaek et al. had a response rate of 48% [57]. Nevertheless, there are examples of meta-analyses in HCI, e.g., about human-robot interaction [36], usability measures [57], and typing experiments [81]. While there are certainly literature reviews in USP, e.g., [15, 32] and also in DCUS [103] we did not find a meta-analysis in this domain.

As part of our systematization of knowledge, we conducted a meta-analysis concerning statistical power in the field of DCUS. However, due to incomplete reporting and test-specific limitations, we could only do this meta-analysis for 140 tests in 20 studies.

3 Literature Collection

In the following, we describe the creation of our literature corpus for DCUS. As a catch-all, we will refer to this sort of literature as a developer paper in the remainder of this work. We define a developer paper as literature including a user study in some form, in which the participants are software developers, software testers, administrators, other people responsible for planning, developing, testing, or managing software, or proxies for such people. An example of proxies would be computer science students, which are commonly used as a stand-in, e.g., for software developers [77]. We focus on the domain of usable security and privacy, which means that the studies should be focused on privacy or security problems and technology. We exclude any papers which do not include a study with actual users.

We started collection of literature in early 2021 and collected developer papers from four major conferences about security and privacy, which were published between 2010 and 2020: SOUPS, USENIX Security, S&P, CCS, and additionally ICSE and the USP tracks of ACM SIGCHI. Abstracts were used to determine whether a paper fits our definition of a developer paper, and in case of uncertainty, the method section of the paper was additionally used to clarify. We updated our literature basis in March 2022. Our final sample consists of 54 papers. Of those, 20 were published at SOUPS, 11 at CCS, 8 at ICSE, 7 at USENIX Security, 5 at S&P and 3 at CHI. The list can be found in Appendix A.

4 Systematizing Study and Statistical Test Data

Our goal was to collect and systematize information on studies and statistical tests from the domain of DCUS, focusing on what is necessary to conduct power analyses for user studies in this domain. We also wanted to add general information on the data collection process and the types of participants, since this might also be relevant when planning a new study. To aid researchers in planning new studies we created a data structure of our systematization and will offer this to the research community as a database. The database can be queried via a companion website for the relevant information to conduct power analysis [2]. An excerpt of two entries can be found in table 2. Further entries are on the companion website. The entries are categorized to help researchers query the database and find similar studies, on which they can base their power calculation. For those cases where no directly similar previous study exists, we have created aggregated data based on our systematization that researchers can use as rough guides.

[2] https://powerdb.info

4.1 Systematization Process

Based on a sample of the literature we had collected, we first analyzed papers from a methodological point of view and collected information on data collection, data analysis, as well as meta information that served to clearly identify and reference the paper. We identified similarities in the type of collected data and iteratively developed a data structure to represent this information.

In addition to the papers themselves, two sources further informed our structure: We made sure to represent information necessary to conduct power analysis, based on the G*Power-software [38], since G*Power is a commonly used and very powerful tool for power analysis.

To help guide our work, we created a set of hypothetical developer studies, for which we would want to run a priori power calculations, e.g., *Do Freelancers recruited from Freelancer.com and Upwork differ in their self-assessment of their reverse engineering skill and in their performance while completing a short reverse engineering task?*. Our aim was to have a mix of hypothetical studies which were closely related to previous work as well as some that had no relations. This was done to a) ensure that it would be easy to find very specific data from closely related work as well as to b) ensure that our categories were useful to guide researchers in uncharted territory as best possible. Based on the case studies, we added features to categorize variables. Finally, after laying the theoretical foundation, we implemented a database and started to enter information from the collected literature.

4.2 Data Structure

In the following we describe how we systematized the data we collected, providing a general overview of the structure, as well as details on those topics specifically relevant to power analysis and finding the right data.

4.2.1 Overview

A general overview of the data structure can be seen in the entity relationship diagram (ERD) in the appendix B.

For each paper in our set, we first collected meta information about it. Each paper can have multiple studies assigned to it. A *study* is a self-contained unit of a combination of data collection and analysis, which is often presented in a separate section in a paper. We separate *data collection descriptions* from the *participant samples* involved. To support filtering and generalization, *Participant sample types*, instances of which could be "student", "security expert", "freelance developer", and *data collection methods*, where instances are, e.g., "interview", "survey" or "experiment/task-based evaluation", are represented as separate entities. These are more easily reused across multiple studies and multiple papers. The results of qualitative analyses cannot directly be used for power analysis or to support meta-analysis since effect sizes are not calculated. However, because insights from qualitative analysis often help inform further quantitative work, we also collected some information on the *qualitative analysis methods* used. Reporting of NHST-type analysis methods all share certain properties, which we collected for all *quantitative analysis methods*, i.e., the name of the hypothesis test, the p-values, and the dependent and independent variables. For more information on the representation of variables, see Section 4.2.2. We categorized the different hypothesis tests used in our sample according to the number and type of dependent and independent variables, the study design, and, in the case of categorical variables, the number of levels in the variable, see Table 1. In the following, we focus on those hypothesis tests with one categorical independent variable with two levels. We collected additional test-specific data for these tests, see Section 4.2.3.

4.2.2 Representation of Variables

For *variables*, we noted the name and a description of how the variable was measured based on the publication and collected information on several additional facets of the variables. For example, the *variable type* attribute represents whether a variable is continuous or categorical. *Variable levels*, i.e., groups, are then provided for categorical variables.

Finally, a *Variable* can be tagged with one or multiple *Variable categories*. These are broad categories of variables, which frequently occur in studies in DCUS, and which serve to ease the search for a specific variable or test and were used to create generic guides for researchers when no specific prior work exists. We generated these categories by open coding all the variables in eight of the papers from our sample, which we chose to cover a broad range of topics and both descriptive and inferential work. One researcher did all the coding alone, and the generated categories and the corresponding variables were frequently discussed together in an iterative process with a second researcher and modified when necessary. Since we were assigning fixed categories to clear units of data, and a second researcher was involved in the generation of the variable categories, we consider this data simple to code and independent recoding to be unnecessary [70, 84]. The eleven categories which emerged from this process are:

usability Measures relating to overall usability, i.e., incorporating effectiveness, efficiency, and user satisfaction according to ISO 9241-11 [59].

security Measures relating to IT security of produced software / artifacts.

functionality Measures relating to the functionality of produced software / artifacts.

participant judgment Measures relating to participants' choices or judgment of something.

Type of IV	# IV	# IV Levels	Type of DV	# DV Levels	Study design	Hypothesis test
Categorical	One	Two	Categorical	Two	Between	Fisher's Exact Test, Chi-Squared Test
Categorical	One	Two	Categorical	Two	Within	McNemar's Test
Categorical	One	Two	Continuous	-	Between	Independent t-test, Wilcoxon Ranksum Test
Categorical	One	Two	Continuous	-	Within	Paired t-test, Wilcoxon Signed Rank Test
Categorical	One +	Two +	Continuous	-	Between	ANOVA, Kruskal-Wallis Test (1 IV)
Categorical	One +	Two +	Continuous	-	Within	Repeated Measures ANOVA, Friedman's ANOVA (1 IV)
Continuous	One	-	Continuous	-	Between	Pearson correlation, Kendall's tau, Spearman's correlation, Polychoric correlation
Any	One +	Any	Categorical	Two	Between	Logistic Regression
Any	One +	Any	Continuous	-	Between	Linear Regression, Poisson Regresion (DV: counts)
Any	One +	Any	Any	Any	Any	Generalized Linear Mixed Model
fixed value	-	-	Categorical	Two	-	Binomial Test
fixed value	-	-	Continuous	-	-	Z-Test

Table 1: Hypothesis tests appearing in DCUS papers. In this paper we focus on tests with single categorical IV and single DV (highlighted in pink) for our assessment of reporting and effect sizes. Remaining tests have a green background. All tests in our sample had a single dependent variable. Three tests did not fit into this categorization: Bernoulli trial, factor analysis and k-means cluster analysis. DV = dependent variable, IV = independent variable

experience Participants' level of experience in something, e.g., programming.

behavior Measures of participants' behavior. Can be either self-reported or objectively measured.

system type Usually an assigned condition in a study, the system / software / prototype, which participants work with or test.

participant type Group to which a participant belongs, e.g., students, freelancing developers.

participant characteristic Any other participant trait, such as the type of company a participant works for, a participant's focus on security, etc.

task related variable Variables related to the tasks in a study, e.g., which task the participants worked on or task order.

study related variable Measures relating to administrative aspects of the study, e.g., drop-outs, prompting, or additional communication with participants, e.g., via email or a support system.

During the data input process, we added one additional category:

artifact-related variable Variables related to artifacts participants produced during or prior to the study, e.g., characteristics of these or types of mistakes encountered in submitted code or other artifacts.

4.2.3 Relevant Test-specific Information for Power Analysis

We specifically focused on collecting information that would be needed to conduct a priori power analysis, as well as additional values, which are typically reported with a hypothesis test according to APA style [5]. For some tests, not all data could be contained in a single entity, e.g., ANOVA, linear regression, and logistic regression. For these, we created multiple related entities in our database, but the meta-analysis will be carried out in future work since very few of these tests were reported with enough detail for a robust analysis at this point.

4.3 Data Input and Checking

Two assistant researchers were hired to aid the main author in entering data into the database. Both had attended at least one university course on statistical hypothesis testing and empirical methods. The main author also has experience teaching empirical methods and statistics.

The main author trained the other two researchers regarding data extraction from the papers and how the data should be entered into the database. The data entry tool provided a checkbox that could be used to mark an entry when feeling unsure, as well as a text field where the issue could be noted. The two assistant researchers received feedback regarding their data entry at set intervals. At the end of the data input process, the main author went over all entered data again to check for any missing data or inconsistencies. Uncertain cases were discussed and resolved together with the co-authors.

5 Meta-Analysis

In the following, we analyze and further systematize the data we gathered. For our meta-analysis of the current state of research in DCUS, we focus on information related to power analysis, e.g., we analyze effect sizes in this field and investigate whether reporting is sufficiently detailed to enable power analysis using the data. We analyzed the data from our database using the Python packages numpy, pandas, and matplotlib and the R tidyverse [91, 119].

We analyzed a total of 54 developer papers, which encompassed 64 individual studies, of which 24 were quantitative, 24 were qualitative, and 16 used both quantitative and qualitative analysis methods. On average, 105 (9 - 330, median=65, SD=99.5) participants took part in quantitative and 14.8 (1 - 49, median=12, SD=12.3) in qualitative studies, and for mixed methods studies on average, 103 (6 - 400, median=44, SD=101.6) participants took part. This is similar to Caine's analysis of papers at CHI 2014, where the sample size was also smaller for qualitative than quantitative work [20].

5.1 Variable Topics

We investigated the distribution of topics of the variables investigated in our literature sample, as represented by the variable categories defined in this work. Multiple categories can apply to a single variable, and this was the case for 138 out of 413 variables.

Of those categories referring to components of usability, e.g., related to either effectiveness, efficiency, or satisfaction [59], *Participant judgment* was the most used variable category (109 times). This may be the case since it applied to all variables representing some sort of participant judgment of an evaluated system or a task, e.g., preference [28], confidence in task correctness [2], or criticality of data [100]. This was followed by *security* (44), as a specific form of effectiveness, which is due to our sample focusing on DCUS. *Functionality* (24) as another form of effectiveness, *usability* overall (14) and *efficiency* (12) appeared less frequently.

Of the other variable categories, which were not related to usability, *participant characteristic* (93), *behavior* (69), and *experience* (47) were the most frequent. Examples for *participant characteristic* are type of participant, e.g., [65, 78], although there is a separate category specifically intended for this, demographics, like state of employment [76] or type of organization [74], and other characteristics relating to participants' opinions or attributes [16, 105]. *Experience*, e.g., with technology like programming languages [e.g., 28, 95], or specific tasks [e.g., 66, 75], often occurs together with *participant characteristic*, as experience can also be considered a defining characteristic of participants. In fact, the most frequent co-occurrence (30) between variable categories was between *experience* and *participant characteristic*. Variables measuring *behavior* included variables tracking participants' behavior during a study, e.g., the number of times they executed a program [54], number of visited websites [66], or lines of code submitted [114], and variables assigned to participants' outcomes retrospectively by researchers [e.g., 114]. Behavior could also be self-reported [e.g., 80, 105]. More specific categories, i.e., *participant type* (9) and *system type* (27) and *artifact related variables* (19) were not as frequent. There was a surprising amount of variables associated with meta-level aspects, such as *task related* (42) and *study related variables* (21), although some of these may be related to task success.

5.2 Use of Hypothesis Tests

We were especially interested in the frequency of hypothesis tests. In the studies in our sample, 7.30 hypothesis tests were conducted per study on average (SD=12.35), and given that a paper could encompass multiple studies, the number of hypothesis tests per paper ranged from 0 (for purely descriptive papers, like [37], or qualitative papers, like [53]) to 74 (M=8.6, SD=13.2). One of the studies, which the authors identified as qualitative, nevertheless contained 26 statistical hypothesis tests. In this comparison of end-users' and administrators' mental models of HTTPS, Fisher's exact tests were used to compare the appearance of various concepts in the mental models of the two participant types [65].

Since some studies included a large number of hypothesis tests, with outliers at 75, 51, and 30 hypothesis tests in a single study, we explored whether p-value correction methods were used in our sample. Figure 1 shows that the majority of studies (31/41) did not use corrections for any of the reported p-values. This likely includes some studies where these corrections were not necessary since the hypotheses tested were about different outcomes or were explicitly considered exploratory [101]. However, all three studies with the largest amount of tests did not include any corrections that we could identify for their hypothesis tests. This means that the results presented may be false positives.

The frequency of each of the different hypothesis tests within our sample is displayed in the left half of Figure 2. The most-used test (123 times) in our sample of papers was the non-parametric Wilcoxon rank-sum (a.k.a. Mann-Whitney U test), followed by two tests used for categorical data, the

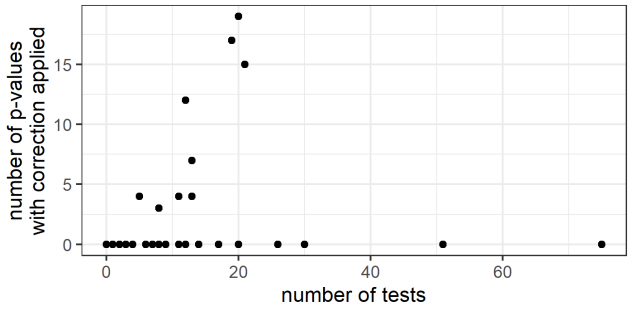

Figure 1: Scatterplot showing the frequency of corrected p-values in relation to non-corrected p-values per study

Fisher's exact and chi-squared Tests, which appeared 59 and 32 times respectively. We categorized the different hypothesis tests used in our sample according to the number and type of dependent and independent variables, the study design, and, in the case of categorical variables, the number of levels in the variable, see Table 1. All of the tests used only one dependent variable. In the remainder of our meta-analysis, we focus on those hypothesis tests with one categorical independent variable with two levels.

5.3 Completeness of Statistical Reporting

Ideally, each of these hypothesis tests would be reported in sufficient detail to be able to conduct power analysis using the data reported in the paper. However, this is not the case, as shown in Figure 3. Of those tests for which we can make this classification, the paired t-test was reported with sufficient information in all cases. However, it was reported only once. For the other tests, the completeness of reporting varied between 11.1% and 74.2% of sufficient reporting (mean=47.5%, sd=28.2%). The most frequently reported test, the Wilcoxon rank-sum test, was reported with sufficient information 38.2% of the time, or in 47 of 123 cases. Overall, this shows that reporting practices, even for these simple tests, are not sufficient to do power analysis using them as a basis about half of the time.

5.3.1 Power Meta-Analysis

To assess whether the field of DCUS suffers from underpowered studies similar to other fields as mentioned in section 2.2, we conducted a power meta-analysis based on Ellis [35, p.74]. This should not be confused with a post hoc power analysis since we did not use the effect sizes or p values reported in the papers. Instead, we only used the reported sample sizes and used G*Power to calculate the power to detect small, medium, and large effects (Cohen's d equivalent of 0.2, 0.5, and 0.8 [22]). This was possible for five of our seven types of hypothesis tests. We excluded Fisher's exact tests and McNemar's tests from this analysis since G*Power required input of effect size based on concrete data from the study, i.e., success proportions for the Fisher's exact test and the total proportion of discordant pairs for McNemar's test. Since post hoc power analysis using values directly from studies like this is not useful [35], we concentrated on the other five types of tests. We set $\alpha = 0.05$ for all analyses and assumed two-tailed tests. For the non-parametric tests, we used the minimal A.R.E. setting in G*Power to get a conservative estimate of the achieved power. Next, we calculated an average over all achieved power values at each level of effect size per included study, and then an overall average [35]. We considered a power of 0.8 to be the lower bound of what should commonly be aimed for [22][3].

Overall, our database contained 20 studies that reported enough statistical data to do this meta-analysis for at least one test. We found that nine of these had sufficient mean and median power to detect large effects, one had sufficient mean, and two had sufficient median power to detect medium effects and none of the studies had sufficient mean or median power to detect small effects. Conversely, eleven of the studies did not have 0.8 power to detect even large effects. However, over all the studies, the mean power to detect large effects was only slightly lower than the 0.8 we considered sufficient (mean=0.743, median=0.773). The mean power to detect medium effects was 0.455 (median=0.396), and for small effects, it was 0.132 (median=0.104). This again highlights the importance of a priori power analysis since developer studies are complex and resource intensive to run, and many do not have sufficient power under the common assumptions of $\alpha = 0.05$ and power of 0.8.

6 Systematic A Priori Power Analysis

Conducting an a priori power analysis requires researchers to know or guess the standardized effect sizes (e.g., Cohen's d) they expect or want to detect. Alternatively, they can also know or guess a non-standardized effect size (such as 1 point on a 7-point scale) and additional information, such as the standard deviation of the groups. In a mature field with many studies examining similar variables (e.g., blood pressure), expected effect sizes might be common knowledge. However, in the absence of closely related prior work, as is often the case in relatively new fields, such as DCUS, a more realistic approach is for the researcher to decide what the smallest practically relevant effect size is, that they want to be able to detect [35], also called the smallest effect size of interest. This can be determined by theoretical considerations but also by juxtaposing the benefit of the desired outcome with the costs to achieve this outcome or by considering practical limitations, such as the number of available participants [67]. In any case, researchers need experience and deep knowledge of their field

[3] Although deviations from this are perfectly fine when done consciously.

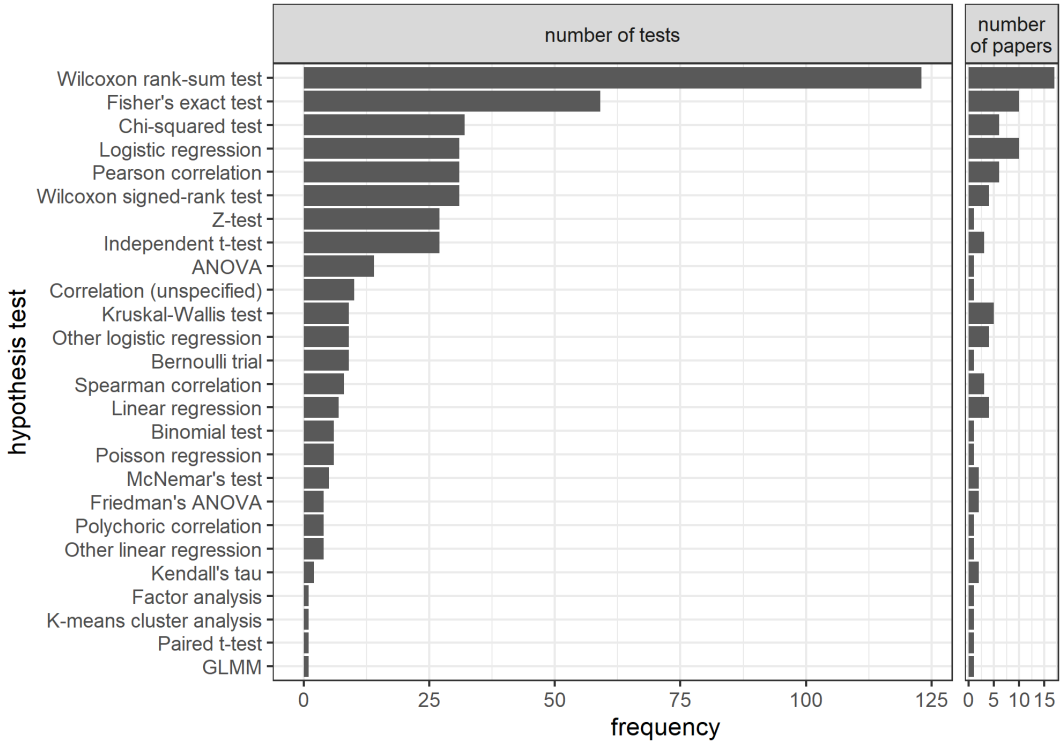

Figure 2: Left side: Frequency of type of hypothesis test in the sample, Right side: Number of papers using this hypothesis test in the sample

DVs in test	IVs in test	Participants	Test	ES	Descriptive stats	Paper
Attempted security (yes, no) security, behavior "participants who attempted to store user passwords securely, but struggled and then deleted their attempts from their solutions (this was coded as attempted but failed, or ABF). [...]"	Priming (priming, non-priming) study-related variable manipulated in experiment. "Priming - Participants were explicitly told to store the user passwords securely in the Introductory Text and in the Task Description."	computer science students (N=40)	FET	OR=19.02; d=1.62	Proportion p1=0.7 (priming); Proportion p2=0.1 (non-priming)	[76]
Secure (secure, insecure) security "In addition, we used a binary variable called secure which was given if participants used at least a hash function in their final solutions and thus did not store the passwords in plain text."	Warning displayed (yes, no) system type, study related variable whether a warning was displayed (could only happen in PyCrypto patch condition). "PyCrypto control condition, or the PyCrypto patch condition, where we tested our security warning"	Python developers (N=53)	FET	OR=56; d=2.22	Proportion p1=0.727 (yes); Proportion p2=0.0455 (no)	[47]

Table 2: Examples of information from the database which can be used for power analysis. The two examples are tests, where there was sufficient data to conduct power analysis. ES=effect size, FET=Fisher's exact test

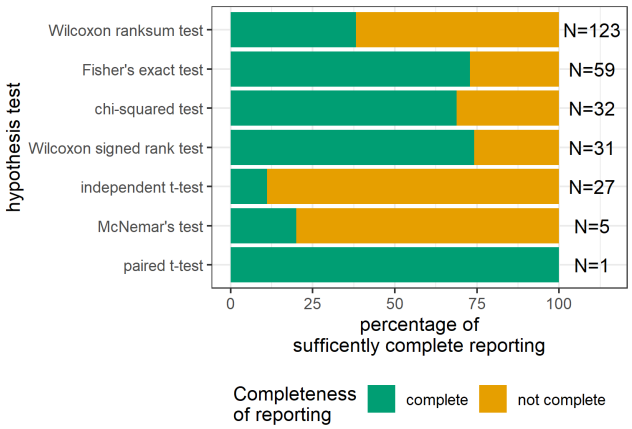

Figure 3: Stacked Barchart depicting the proportions sufficiently complete reporting for those tests at the focus of our analysis

to estimate effect sizes and this is one aspect that makes power analysis difficult [23]. In addition, used effect sizes [35] and power analysis procedures vary between different types of statistical tests [38].

Our systematization of knowledge aims to ease this process. Ideally, related studies have already been published, and standardized effect sizes are reported or can be calculated. In this case, the researcher needs to be able to find them. For this, they can query our database introduced in Section 4. Some types of information systematized in the database which are helpful for the search include variable categories, variables themselves and participant sample type. The participant sample type can help researchers identify prior work with a similar demographic to their planned study.

Table 2 shows an excerpt of the data which can be returned by the database. The first row contains all the data needed to perform a power analysis for a Fisher's exact test. A researcher could find it by any of the keywords or categories listed. The second row contains all the information to perform a power calculation for a different Fisher's exact test. The full data set is available on our companion website [4]. For each test, variables, effect sizes, relevant information to calculate them, the participant sample, and the source paper are listed and sorted by variable category.

Ideally, there will be several similar studies in the database, on which researchers can then base their effect size estimates on. However, this is currently unlikely since the field is still very young and diverse. But even if this is not the case, finding results for some of the variables of interest or a specific demographic can help refine effect size estimates.

[4] https://powerdb.info

6.1 Effectsize Meta-Analysis

As a final step in our systematization, we conducted a meta-analysis of the standardized effect sizes from tests where we had enough information for power analysis. With this, we aimed at providing a broad overview of effect sizes in DCUS which can be used to sanity check power analyses. While it is preferable to find exact or at least close matches, we also want to support researchers where this is not possible. Without related work, researchers basically have to guess standardized effect sizes or things like expected proportions or standard deviations. To at least give a frame of reference against which to judge these guesses, we examined the range of effect sizes present in the field of DCUS. We used the categories from Section 5.1 to aggregate the data for our guide.

To enable comparison and aggregation, we used the effect sizes directly reported in the paper where possible and converted them to Cohen's d, as this is one of the most widely used effect sizes in our sample. In other cases, we first used the provided data to calculate an effect size, which we then converted to Cohen's d. For converting d to odds ratio (OR), we used the formula from Haddock et al. [51], and a correction factor of 1.09, which is an average over the correction factors Poom and af Wåhlberg recommend for sample sizes between 20 and 100 [89], since developer studies mostly feature smaller sample sizes. To convert between d and ϕ, we used Rosenthal's formula [92] as described by Burns et al. [18]. Finally, as described above, sufficient data was not reported for many tests, and we exclude such tests from our analysis.

When judging the size of effects, effect sizes of Cohen's d=0.2 are generally regarded as small, d=0.5 as medium, and d=0.8 as large effect sizes [22]. When converted to other effect sizes, this yields OR=1.37, 2.21, 3.57 as small, medium, and large effect sizes displayed as odds ratios, ϕ=0.10, 0.24, 0.37 for the effect size ϕ used with χ^2-tests.

While we did not encounter the correlation coefficient r used as an effect size in this analysis, we nevertheless note for researchers encountering r, that effect sizes of r=0.1, 0.3, and 0.5 are considered small, medium and large effects respectively [22].

In our DCUS sample, the reported effect sizes ranged from equivalents of Cohen's d=0.004 to Cohen's d=2.22 (M=0.55, median=0.47, SD=0.42). We excluded two effect sizes from a Fisher's exact test in [76] and a McNemar's test in [106], where the reported success proportion in one of the groups was zero, and thus the effect size approaches infinity. Figure 4 shows the distributions of effect sizes separately for the variable categories of the dependent variables. So if researchers have to make educated guesses for their power analysis, they can compare their values with the violin plots to judge where in the spectrum they lie.

Three variable categories were not assigned to dependent variables with a present effect size: experience, system

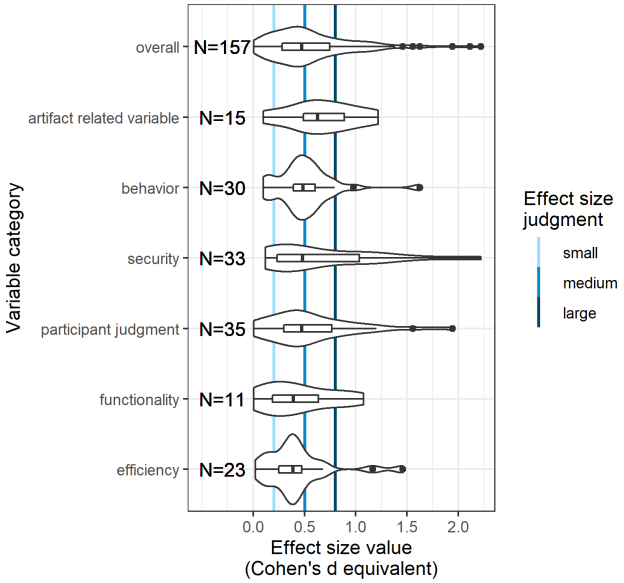

Figure 4: Violinplots for the distribution of effect sizes for tests, faceted by the variable categories of the dependent variable

type, and participant type. These categories were more frequently applied to independent variables. We did not plot data for variable categories where fewer than five effect sizes were reported. This is the case for four variable categories: usability (N=2), participant characteristics (N=1), task related (N=1), and study related variables (N=3). All of the five remaining variable categories exhibit a large variance of effect sizes, which range from negligible and small to large and very large effects. Median effects for tests with artifact-related variables as the dependent variable are in the medium range, and for all other categories and overall, the median effects are in the small range. However, in the cases of behavior, security, and participant judgment as well as overall, they border on medium according to Cohen [22]. We will update the online version when enough tests have been reported with the necessary statistical details.

6.2 Publication Bias Correction

There is one final important warning that needs to be highlighted. Irrespective of whether single entries from the database are used or our aggregation violin plots, researchers must be mindful of the effect of publication bias on reported effect sizes [58]. Since papers with statistically significant results are more likely to be published and studies in DCUS often have fairly poor power it should be expected that reported sample effect sizes are larger than true population effect sizes (see Ellis, p.79ff. [35]). Ideally, our database would also include work that is methodologically sound but did not get published due to statistically non-significant tests. However, since we could not think of any feasible way of including this at scale, we recommend taking this publication/sampling bias into account.

Consequently, when planning a study using effect sizes from related work or from our systematization or even a pre-study, it is recommendable to either correct the acquired effect size estimates [60, 111] or increase the desired sample size to be able to detect slightly smaller effects than ones reported in prior work. While this correction still requires some guesswork, it is a lot easier than having to guess blindly.

7 Power Analysis and Reporting Recommendations

While we hope that our systematization and database will already be a useful aid to researchers, it is still incomplete. A big hindrance in conducting our work was the fact that many tests were not reported with sufficient statistical information.

The American Psychological Association's publication manual contains a very comprehensive list of recommendations on how statistical tests should be reported [5]. Based on the APA and our findings in DCUS we want to highlight some recommendations for statistical reporting that we believe would be particularly helpful for future power analysis.

Report standardized effect sizes Wherever possible, report standardized effect sizes in addition to p values. Ellis provides an overview of effect sizes in Table 1.1 [35], and the guides in the appendix C have notes on effect sizes for seven commonly used hypothesis tests.

Report non-standardized effect sizes and descriptives Since standardized effect sizes can be unintuitive, also report descriptive statistics showing the effect size, e.g., there was a 1-point difference on a 7-point scale so that interested researchers can calculate effect sizes themselves. **Report frequencies** within all groups when comparing nominal variables, as effect size calculations commonly use these. **Report descriptive statistics for each group** of the independent variable(s) when comparing ordinal or interval variables. Report at least means, standard deviations, and group sizes. If this becomes too extensive, move this information to the appendix or supplemental material. **Make sure that the descriptives make sense**, e.g., a mean is not a good summary statistic for a bimodal distribution.

Make fully anonymized data sets available when needed If presenting all the descriptive statistics and frequencies is too extensive, e.g., for regression analyses with multiple independent variables, ideally, the anonymized data can be made available.

Hypothesis confirmation vs exploratory analysis State whether pre-defined hypotheses are being tested or

whether an exploratory analysis is being conducted [107].

8 Limitations

Our paper selection process focuses on the top-tier venues in which developer papers are published. Our sample does not include workshop papers or unpublished work, thus the publication and sampling bias needs to be taken into account as described above.

Additionally, even though descriptive data is useful when assessing effect sizes and conducting power analysis, we only included descriptive data in our database which was directly associated with a hypothesis test. The sheer amount of descriptive data reported in some papers, and the variety of ways it is reported, which included visualizations only partially enabling inference of exact numbers, tables, within the text or in the appendix, makes it hard to find a general structure for storing this type of data. We defer this to future work.

Our work focuses only on a priori power analysis as described in Section 2.2. Power, and as such, power analysis is situated within the NHST framework, and the analyses of the field and recommendations in this work apply within NHST. NHST results in point estimates about coefficients or effect sizes and a priori power analysis serves to determine the number of participants needed to detect such an effect at a specified significance level. The practice of focusing on the value of point estimates, rather than the precision of the estimates has also been criticized [69]. Even when planning a study with adequate power to detect an effect, the confidence intervals around it can be wide, leading to little precision of the effect size. Different analysis methods also exist, e.g., Bayesian analysis [63, 64]. Additionally, some statistical analyses, like regression, are not used merely in NHST to falsify hypotheses, but also to make predictions, and different judgment criteria would apply in these cases [19]. In predictive analysis, i.e., what is often known as machine learning, the influence and explanatory power of individual variables included in the model is not as important as the accuracy of the prediction [17, 19]. Other criteria for what constitutes good reporting may apply in these cases than what we have covered in this work. However, NHST is commonly used in DCUS specifically and HCI in general. We did not in fact encounter any Bayesian analyses in our sample of analyzed DCUS papers and given that recruiting software developers is hard [4], having sufficient data for large-scale predictive analyses is likely rare in this field. In conclusion, we believe that our contributions align with common methods used in DCUS at the time.

Finally, we only conducted our meta-analysis on tests with a single categorical IV and a single DV. Thus, our work is limited to assisting in the power analysis of the following tests: Fisher's exact test, chi-squared test, McNemar's test, Wilcoxon rank-sum tests, Wilcoxon signed-rank tests, and paired and independent t-tests. We will extend this in future work. However, in combination with simplifications even the current data might offer benefits for the assessment of more complex tests[68, 87].

9 Conclusion

In this work, we systematized 467 tests and 413 variables from a data set of 54 DCUS papers published in top-tier venues. We examined their methodology and reporting of statistical results, as well as their power to detect effects of different sizes, which was not sufficient for small effects, and only sufficient for large effects in about half of the papers, where enough information was reported to analyze the power. We provide domain-specific effect size ranges for different categories of variables, which can serve as a fall-back for effect size estimation in a priori power analysis, with effect sizes in DCUS averaging at Cohen's d=0.55, when considering all variable categories. The raw data on which these ranges are based is included in a searchable database of extracted information from these papers, which other researchers can use to reproduce our analyses and facilitate their own power analyses. The brief guides in the appendix can supply further assistance in conducting power analysis for some simple but often-used hypothesis tests. When reporting statistical results, authors should include effect sizes, both standardized and non-standardized, as well as descriptive statistics for each condition as a way to foster the use of power analysis in sample size planning and to enable the re-analysis of results through meta-analysis.

10 Future Work

As stated in the limitations section, we currently only cover a subset of all tests. In future work, we plan to extend this list to include more tests. The current implementation of our database does not have a custom user interface and is operated using SQL or other query tools to access and understand the data within. Ongoing work aims to improve the usability of querying the database. We also plan to extend our approach to the whole field of usable security and privacy. While DCUS faces particular challenges when recruiting participants, we believe end-user studies would also benefit from a priori power analysis. To aid in this extension and general upkeep we plan on developing a public-facing interface to the database so researchers can add their own papers. In the future, the database could also be used to explore other aspects of methodology use, such as the number of coders and use of inter-rater reliability in qualitative work, whether the behavior is measured objectively or subjectively using the categorization of variables, or to conduct sensitivity analysis, i.e., to analyze the power to detect effects.

Acknowledgments

We would like to thank all the students working on paper categorization or the database application in various courses between 2021 and 2023: Ahmad Alaya, Somar Aljabr, Ahmad Assaf, Marwan Jaradat, Tansen Khan, Heng-yi Lin, Florin Martius and Atacan Süder, our colleagues Lisa Geierhaas, Eva Gerlitz, Maximilian Häring, Mischa Meier, Stephan Plöger, and Klaus Tulbure, who aided with paper collection and selection, as well as Simon Bong and Theo Raimbault, who transferred information from collected papers into the database.

References

[1] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. Comparing the Usability of Cryptographic APIs. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 154–171, San Jose, CA, USA, May 2017. IEEE.

[2] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. You Get Where You're Looking for: The Impact of Information Sources on Code Security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 289–305, San Jose, CA, May 2016. IEEE.

[3] Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. You are Not Your Developer, Either: A Research Agenda for Usable Security and Privacy Research Beyond End Users. In *2016 IEEE Cybersecurity Development (SecDev)*, pages 3–8, Boston, MA, USA, November 2016. IEEE.

[4] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L. Mazurek, and Sascha Fahl. Security Developer Studies with {GitHub} Users: Exploring a Convenience Sample. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 81–95, 2017.

[5] American Psychological Association, editor. *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, DC, seventh edition edition, 2020.

[6] Hala Assal and Sonia Chiasson. Security in the Software Development Lifecycle. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 281–296, Baltimore, MD, August 2018. USENIX Association.

[7] Donald R. Atkinson, Michael J. Furlong, and Bruce E. Wampold. Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, 29(2):189–194, 1982.

[8] Peter C. Austin and Ewout W. Steyerberg. The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, 68(6):627–636, June 2015.

[9] Nathaniel Ayewah and William Pugh. A report on a survey and study of static analysis users. In *Proceedings of the 2008 Workshop on Defects in Large Software Systems*, DEFECTS '08, pages 1–5, New York, NY, USA, July 2008. Association for Computing Machinery.

[10] Thom Baguley. Standardized or simple effect size: What should be reported? *British Journal of Psychology*, 100(3):603–617, 2009.

[11] Khadija Baig, Elisa Kazan, Kalpana Hundlani, Sana Maqsood, and Sonia Chiasson. Replication: Effects of Media on the Mental Models of Technical Users. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 119–138, 2021.

[12] Jack J. Baroudi and Wanda J. Orlikowski. The problem of statistical power in MIS research. *MIS Quarterly*, 13(1):87–106, 1989.

[13] Dirk Bassler, Matthias Briel, Victor M. Montori, Melanie Lane, Paul Glasziou, Qi Zhou, Diane Heels-Ansdell, Stephen D. Walter, Gordon H. Guyatt, and and the STOPIT-2 Study Group. Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects: Systematic Review and Meta-regression Analysis. *JAMA*, 303(12):1180–1187, March 2010.

[14] Scott Bezeau and Roger Graves. Statistical Power and Effect Sizes of Clinical Neuropsychology Research. *Journal of Clinical and Experimental Neuropsychology*, 23(3):399–406, June 2001.

[15] Robert Biddle, Sonia Chiasson, and P.C. Van Oorschot. Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys*, 44(4):19:1–19:41, September 2012.

[16] Larissa Braz, Enrico Fregnan, Gül Çalikli, and Alberto Bacchelli. Why Don't Developers Detect Improper Input Validation? '; DROP TABLE Papers; –. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 499–511, May 2021.

[17] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, August 2001.

[18] Matthew K. Burns, Anne F. Zaslofsky, Rebecca Kanive, and David C. Parker. Meta-Analysis of Incremental Rehearsal Using Phi Coefficients to Compare Single-Case and Group Designs. *Journal of Behavioral Education*, 21(3):185–202, September 2012.

[19] Danilo Bzdok, Denis Engemann, and Bertrand Thirion. Inference and Prediction Diverge in Biomedicine. *Patterns*, 1(8):100119, November 2020.

[20] Kelly Caine. Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 981–992, New York, NY, USA, 2016. Association for Computing Machinery.

[21] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. Threats of a Replication Crisis in Empirical Computer Science. *Communications of the ACM*, 63(8):70–79, 2020.

[22] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, Hillsdale, N.J, 2nd ed edition, 1988.

[23] Jacob Cohen. A power primer. In A. E. Kazdin, editor, *Methodological Issues and Strategies in Clinical Research*, pages 279–284. American Psychological Association, Washington, DC, US, 2016.

[24] Shaanan Cohney, Ross Teixeira, Anne Kohlbrenner, Arvind Narayanan, Mihir Kshirsagar, Yan Shvartzshnaider, and Madelyn Sanfilippo. Virtual Classrooms and Real Harms: Remote Learning at {U.S}. Universities. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 653–674, 2021.

[25] David A. Cook and Rose Hatala. Got power? A systematic review of sample size adequacy in health professions education research. *Advances in Health Sciences Education*, 20(1):73–83, March 2015.

[26] Anastasia Danilova, Alena Naiakshina, Johanna Deuter, and Matthew Smith. Replication: On the Ecological Validity of Online Security Developer Studies: Exploring Deception in a Password-Storage Study with Freelancers. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 165–183. USENIX Association, August 2020.

[27] Anastasia Danilova, Alena Naiakshina, Anna Rasgauski, and Matthew Smith. Code Reviewing as Methodology for Online Security Studies with Developers - A Case Study with Freelancers on Password Storage. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 397–416, 2021.

[28] Anastasia Danilova, Alena Naiakshina, and Matthew Smith. One Size Does Not Fit All: A Grounded Theory and Online Survey Study of Developer Preferences for Security Warning Types. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 136–148, October 2020.

[29] Erik Derr, Sven Bugiel, Sascha Fahl, Yasemin Acar, and Michael Backes. Keep me Updated: An Empirical Study of Third-Party Library Updatability on Android. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 2187–2200, Dallas Texas USA, October 2017. ACM.

[30] Julian di Stephano. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology*, 17(5):707–709, 2003.

[31] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. Investigating System Operators' Perspective on Security Misconfigurations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, pages 1272–1289, New York, NY, USA, October 2018. Association for Computing Machinery.

[32] Verena Distler, Matthias Fassl, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Lorrie Faith Cranor, and Vincent Koenig. A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research. *ACM Transactions on Computer-Human Interaction*, 28(6):43:1–43:50, December 2021.

[33] Alexander Eiselmayer. Supporting the Design and Analysis of HCI Experiments. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, Honolulu HI USA, April 2020. ACM.

[34] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. Touchstone2: An interactive environment for exploring trade-offs in HCI experiment design. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–11, New York, NY, USA, 2019. Association for Computing Machinery.

[35] Paul D. Ellis. *The Essential Guide to Effect Sizes. Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, 2010.

[36] Connor Esterwood, Kyle Essenmacher, Han Yang, Fanpan Zeng, and Lionel Peter Robert. A Meta-Analysis of Human Personality and Robot Acceptance in Human-Robot Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pages 1–18, New York, NY, USA, May 2021. Association for Computing Machinery.

[37] Sascha Fahl, Marian Harbach, Henning Perl, Markus Koetter, and Matthew Smith. Rethinking SSL development in an appified world. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security - CCS '13*, CCS '13, pages 49–60, Berlin, Germany, 2013. ACM Press.

[38] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, May 2007.

[39] Andy Field, Jeremy Miles, and Zoe Field. *Discovering Statistics Using R*. SAGE Publications Ltd, London ; Thousand Oaks, Calif, 1. edition edition, April 2012.

[40] Felix Fischer, Yannick Stachelscheid, and Jens Grossklags. The Effect of Google Search on Software Security: Unobtrusive Security Interventions via Content Re-ranking. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 3070–3084, New York, NY, USA, November 2021. Association for Computing Machinery.

[41] Felix Fischer, Huang Xiao, Ching-Yu Kao, Yannick Stachelscheid, Benjamin Johnson, Danial Razar, Paul Fawkesley, Nat Buckley, Konstantin Bottinger, Paul Muntean, and Jens Grossklags. Stack Overflow Considered Helpful! Deep Learning Security Nudges Towards Stronger Cryptography. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 339–356, Santa Clara, CA, USA, August 2019. USENIX Association.

[42] Boris Freidlin and Edward L Korn. Stopping clinical trials early for benefit: Impact on estimation. *Clinical Trials*, 6(2):119–125, April 2009.

[43] Kelsey R Fulton, Anna Chan, Daniel Votipka, Michael Hicks, and Michelle L Mazurek. Benefits and Drawbacks of Adopting a Secure Programming Language: Rust as a Case Study. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, SOUPS '21, page 20. USENIX Association, August 2021.

[44] Eva Gerlitz, Maximilian Häring, and Matthew Smith. Please do not use !?_ or your License Plate Number: Analyzing Password Policies in German Companies. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 17–36, 2021.

[45] Steven N. Goodman and Jesse A. Berlin. The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results. *Annals of Internal Medicine*, 121(3):200–206, 1994.

[46] Peter Leo Gorski, Yasemin Acar, Luigi Lo Iacono, and Sascha Fahl. Listen to Developers! A Participatory Design Study on Security Warnings for Cryptographic APIs. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA, April 2020. ACM.

[47] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Moeller, Yasemin Acar, and Sascha Fahl. Developers Deserve Security Warnings, Too. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 265–281, Baltimore, MD, August 2018. USENIX Association.

[48] Matthew Green and Matthew Smith. Developers are Not the Enemy!: The Need for Usable Security APIs. *IEEE Security & Privacy*, 14(5):40–46, September 2016.

[49] Samuel B. Green. How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behavioral Research*, 26(3):499–510, July 1991.

[50] Thomas Groß. Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies. In Thomas Groß and Theo Tryfonas, editors, *Socio-Technical Aspects in Security and Trust*, Lecture Notes in Computer Science, pages 3–26, Cham, 2021. Springer International Publishing.

[51] C. Keith Haddock, David Rindskopf, and William R. Shadish. Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods*, 3(3):339–353, 1998.

[52] Joseph Hallett, Nikhil Patnaik, Benjamin Shreeve, and Awais Rashid. “Do this! Do that!, and Nothing will Happen” Do Specifications Lead to Securely Stored Passwords? In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 486–498, May 2021.

[53] Julie M Haney, Mary F Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. "We make it a big deal in the company": Security Mindsets in Organizations that Develop Cryptographic Products. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 357–373, Baltimore, MD, August 2018. USENIX Association.

[54] Norman Hänsch, Andrea Schankin, and Mykolai Protsenko. Programming Experience Might Not Help in Comprehending Obfuscated Source Code Efficiently. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 341–356, Baltimore, MD, August 2018. USENIX Association.

[55] Moonseong Heo, Singh R. Nair, Judith Wylie-Rosett, Myles S. Faith, Angelo Pietrobelli, Nancy R. Glassman, Sarah N. Martin, Stephanie Dickinson, and David B. Allison. Trial Characteristics and Appropriateness of Statistical Methods Applied for Design and Analysis of Randomized School-Based Studies Addressing Weight-Related Issues: A Literature Review. *Journal of Obesity*, 2018:e8767315, June 2018.

[56] John M Hoenig and Dennis M Heisey. The Abuse of Power. *The American Statistician*, 55(1):19–24, February 2001.

[57] Kasper Hornbæk and Effie Lai-Chong Law. Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 617–626, San Jose California USA, April 2007. ACM.

[58] John P. A. Ioannidis. Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19(5):640–648, 2008.

[59] ISO Central Secretary. Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts. Standard ISO 9241-11:2018, International Organization for Standardization, Geneva, CH, 2018.

[60] Andreas Ivarsson, Mark B. Andersen, Urban Johnson, and Magnus Lindwall. To adjust or not adjust: Nonparametric effect sizes, confidence intervals, and real-world meaning. *Psychology of Sport and Exercise*, 14(1):97–102, January 2013.

[61] Brittany Johnson, Yoonki Song, Emerson Murphy-Hill, and Robert Bowdidge. Why don't software developers use static analysis tools to find bugs? In *2013 35th International Conference on Software Engineering (ICSE)*, pages 672–681, May 2013.

[62] Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag I.K. Sjøberg. A systematic review of effect size in software engineering experiments. *Information and Software Technology*, 49(11):1073–1086, 2007.

[63] Maurits Kaptein and Judy Robertson. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1105–1114, New York, NY, USA, May 2012. Association for Computing Machinery.

[64] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4521–4532, San Jose California USA, May 2016. ACM.

[65] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. "If HTTPS Were Secure, I Wouldn't Need 2FA" - End User and Administrator Mental Models of HTTPS. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 246–263, San Francisco, CA, USA, May 2019. IEEE.

[66] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. “I Have No Idea What I'm Doing” – On the Usability of Deploying HTTPS. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 1339–1356, Vancouver, BC, Canada, 2017. USENIX Association.

[67] Daniël Lakens. Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7):701–710, 2014.

[68] Sean P. Lane and Erin P. Hennes. Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, 35(1):7–31, January 2018.

[69] Scott E. Maxwell, Ken Kelley, and Joseph R. Rausch. Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, 59(1):537–563, January 2008.

[70] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, November 2019.

[71] Abraham H Mhaidli, Yixin Zou, and Florian Schaub. "We Can't Live Without Them!" App Developers' Adoption of Ad Networks and Their Considerations of Consumer Risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, pages 225–244, Santa Clara, CA, August 2019. USENIX Association.

[72] Victor M. Montori, P. J. Devereaux, Neill K. J. Adhikari, Karen E. A. Burns, Christoph H. Eggert, Matthias Briel, Christina Lacchetti, Teresa W. Leung, Elizabeth Darling, Dianne M. Bryant, Heiner C. Bucher, Holger J. Schünemann, Maureen O. Meade, Deborah J. Cook, Patricia J. Erwin, Amit Sood, Richa Sood, Benjamin Lo, Carly A. Thompson, Qi Zhou, Edward Mills, and Gordon H. Guyatt. Randomized Trials Stopped Early for BenefitA Systematic Review. *JAMA*, 294(17):2203–2209, November 2005.

[73] Sarah Nadi, Stefan Krüger, Mira Mezini, and Eric Bodden. Jumping through hoops: Why do Java developers struggle with cryptography APIs? In *Proceedings of the 38th International Conference on Software Engineering*, pages 935–946, Austin Texas, May 2016. ACM.

[74] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. On Conducting Security Developer Studies with CS Students: Examining a Password-Storage Study with CS Students, Freelancers, and Company Developers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA, April 2020. ACM.

[75] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "If you want, I can store the encrypted password": A Password-Storage Field Study with Freelance Developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, Glasgow, Scotland, Uk, May 2019. ACM.

[76] Alena Naiakshina, Anastasia Danilova, and Christian Tiefenau. Deception Task Design in Developer Password Studies: Exploring a Student Sample. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 297–313, Baltimore, MD, August 2018. USENIX Association.

[77] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why Do Developers Get Password Storage Wrong?: A Qualitative Usability Study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 311–328, Dallas Texas USA, October 2017. ACM.

[78] Duc Cuong Nguyen, Dominik Wermke, Yasemin Acar, Michael Backes, Charles Weir, and Sascha Fahl. A Stitch in Time: Supporting Android Developers in Writing Secure Code. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 1065–1077, Dallas Texas USA, October 2017. ACM.

[79] Jakob Nielsen and Jonathan Levy. Measuring usability: Preference vs. performance. *Communications of the ACM*, 37(4):66–75, April 1994.

[80] Tim Nosco, Jared Ziegler, and Zechariah Clark. The Industrial Age of Hacking. In *29th USENIX Security Symposium (USENIX Security 20)*, USENIX Security '20, pages 1129–1146. USENIX Association, August 2020.

[81] Natalia Obukhova. A Meta-Analysis of Effect Sizes of CHI Typing Experiments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–7, New York, NY, USA, May 2021. Association for Computing Machinery.

[82] Daniela Seabra Oliveira, Tian Lin, Muhammad Sajidur Rahman, Rad Akefirad, Donovan Ellis, Eliany Perez, and Rahul Bobhate. API Blindspots: Why Experienced Developers Write Vulnerable Code. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, SOUPS '18, pages 315–328, Baltimore, MD, August 2018. USENIX Association.

[83] Marten Oltrogge, Yasemin Acar, Sergej Dechand, Matthew Smith, and Sascha Fahl. To Pin or Not to Pin Helping App Developers Bullet Proof Their TLS Connections. In *24th 24th USENIX Security Symposium (USENIX Security 15*, pages 239–254, Washington, D.C., USA, August 2015. USENIX Association.

[84] Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. Different researchers, different results? analyzing the influence of researcher experience and data type during qualitative analysis of an interview and survey study on security advice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[85] Hernan Palombo, Armin Ziaie Tabari, Daniel Lende, Jay Ligatti, and Xinming Ou. An Ethnographic Understanding of Software (In)Security and a Co-Creation Model to Improve Secure Software Development. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, SOUPS '20, page 17. USENIX Association, August 2020.

[86] Ivan Pashchenko, Duc-Ly Vu, and Fabio Massacci. A Qualitative Study of Dependency Management and Its Security Implications. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1513–1531, Virtual Event USA, October 2020. ACM.

[87] Marco Perugini, Marcello Gallucci, and Giulio Costantini. A Practical Primer To Power Analysis for Simple Experimental Designs. 31(1):20, July 2018.

[88] Stephan Plöger, Mischa Meier, and Matthew Smith. A Qualitative Usability Evaluation of the Clang Static Analyzer and {libFuzzer} with {CS} Students and {CTF} Players. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 553–572, 2021.

[89] Leo Poom and Anders af Wåhlberg. Accuracy of conversion formula for effect sizes: A Monte Carlo simulation. *Research Synthesis Methods*, 13(4):508–519, 2022.

[90] Lutz Prechelt. Plat_Forms: A Web Development Platform Comparison by an Exploratory Experiment Searching for Emergent Platform Properties. *IEEE Transactions on Software Engineering*, 37(1):95–108, January 2011.

[91] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[92] Robert Rosenthal. Parametric Measures of Effect Size. In Harris Cooper and Larry Hedges, editors, *The Handbook of Research Synthesis*, pages 231–244. Russel Sage Foundation, New York, 1994.

[93] Joseph Rossi. Statistical Power of Psychological Research: What Have We Gained in 20 Years? *Journal of Consulting and Clinical Psychology*, 58(5):646–656, 1990.

[94] Sebastian Roth, Lea Gröber, Michael Backes, Katharina Krombholz, and Ben Stock. 12 Angry Developers - A Qualitative Study on Developers' Struggles with CSP. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 3085–3103, New York, NY, USA, November 2021. Association for Computing Machinery.

[95] Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L. Mazurek, and Piotr Mardziel. Build It, Break It, Fix It: Contesting Secure Development. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 690–703, Vienna Austria, October 2016. ACM.

[96] Roberta W Scherer, Joerg J Meerpohl, Nadine Pfeifer, Christine Schmucker, Guido Schwarzer, and Erik von Elm. Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews*, 2018(11), November 2018.

[97] I. Manjula Schou and Ian C. Marschner. Meta-analysis of clinical trials with early stopping: An investigation of potential bias. *Statistics in Medicine*, 32(28):4859–4874, 2013.

[98] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[99] Peter Sedlmeier and Gerd Gigerenzer. Do studies of statistical power have an effect on the power of studies? In *Methodological Issues & Strategies in Clinical Research*, pages 389–406. American Psychological Association, Washington, DC, US, 1992.

[100] Swapneel Sheth, Gail Kaiser, and Walid Maalej. Us and them: A study of privacy requirements across north america, asia, and europe. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE 2014, pages 859–870, New York, NY, USA, May 2014. Association for Computing Machinery.

[101] David L. Streiner and Geoffrey R. Norman. Correction for Multiple Testing: Is There a Resolution? *CHEST*, 140(1):16–18, July 2011.

[102] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Deciding on Personalized Ads: Nudging Developers About User Privacy. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 573–596. USENIX Association, 2021.

[103] Mohammad Tahaei and Kami Vaniea. A Survey on Developer-Centred Security. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 129–138, June 2019.

[104] Meinald T. Thielsch, Jana Scharfen, Ehsan Masoudi, and Meike Reuter. Visual Aesthetics and Performance: A First Meta-Analysis. In *Proceedings of Mensch Und Computer 2019*, pages 199–210, Hamburg Germany, September 2019. ACM.

[105] Christian Tiefenau, Maximilian Häring, and Katharina Krombholz. Security, Availability, and Multiple Information Sources: Exploring Update Behavior of System Administrators. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, SOUPS '20, pages 239–258. USENIX Association, August 2020.

[106] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. A Usability Evaluation of Let's Encrypt and Certbot: Usable Security Done Right. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pages 1971–1988, London United Kingdom, November 2019. ACM.

[107] Andrew T. Tredennick, Giles Hooker, Stephen P. Ellner, and Peter B. Adler. A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102(6), June 2021.

[108] Patrizio E. Tressoldi and David Giofré. The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology*, 6, 2015.

[109] Patrizio E. Tressoldi, David Giofré, Francesco Sella, and Geoff Cumming. High Impact = High Statistical Standards? Not Necessarily So. *PLOS ONE*, 8(2):e56180, February 2013.

[110] Anwesh Tuladhar, Daniel Lende, Jay Ligatti, and Xinming Ou. An Analysis of the Role of Situated Learning in Starting a Security Culture in a Software Company. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 617–632. USENIX Association, 2021.

[111] Tammi Vacha-Haase and Bruce Thompson. How to Estimate and Interpret Various Effect Sizes. *Journal of Counseling Psychology*, 51(4):473–481, 2004.

[112] Dirk van der Linden, Pauline Anthonysamy, Bashar Nuseibeh, Thein Than Tun, Marian Petre, Mark Levine, John Towse, and Awais Rashid. Schrödinger's Security: Opening the Box on App Developers' Security Rationale. In *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 149–160, October 2020.

[113] Ivan Vankov, Jeffrey Bowers, and Marcus Munafò. On the persistence of low power in psychological science. *Q J Exp Psychol (Hove)*, 67(6):1037–1040, 2014.

[114] Daniel Votipka, Kelsey R Fulton, James Parker, Matthew Hou, Michelle L Mazurek, and Michael Hicks. Understanding security mistakes developers make: Qualitative analysis from Build It, Break It, Fix It. In *29th USENIX Security Symposium (USENIX Security 20)*, USENIX Security '20, pages 109–126. USENIX Association, August 2020.

[115] Daniel Votipka, Seth Rabin, Kristopher Micinski, Jeffrey S Foster, and Michelle L Mazurek. An Observational Investigation of Reverse Engineers' Processes. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1875–1892. USENIX Association, August 2020.

[116] Daniel Votipka, Rock Stevens, Elissa Redmiles, Jeremy Hu, and Michelle Mazurek. Hackers vs. Testers: A Comparison of Software Vulnerability Discovery Processes. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 374–391, San Francisco, CA, May 2018. IEEE.

[117] Charles Weir, Ben Hermann, and Sascha Fahl. From Needs to Actions to Secure Apps? The Effect of Requirements and Developer Practices on App Security. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 289–305. USENIX Association, August 2020.

[118] Charles Weir, Awais Rashid, and James Noble. How to improve the security skills of mobile app developers? Comparing and contrasting expert views. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, SOUPS'16, Denver Colorado USA, June 2016. USENIX Association.

[119] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[120] Khaled Yakdan, Sergej Dechand, Elmar Gerhards-Padilla, and Matthew Smith. Helping Johnny to Analyze Malware: A Usability-Optimized Decompiler and Malware Analysis User Study. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 158–177, San Jose, CA, USA, May 2016. IEEE.

[121] Miuyin Yong Wong, Matthew Landen, Manos Antonakakis, Douglas M. Blough, Elissa M. Redmiles, and Mustaque Ahamad. An Inside Look into the Practice of Malware Analysis. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 3053–3069, New York, NY, USA, November 2021. Association for Computing Machinery.

[122] Koen Yskout, Riccardo Scandariato, and Wouter Joosen. Does organizing security patterns focus architectural choices? In *2012 34th International Conference on Software Engineering (ICSE)*, pages 617–627, Zurich, June 2012. IEEE.

[123] Koen Yskout, Riccardo Scandariato, and Wouter Joosen. Do security patterns really help designers? In *Proceedings of the 37th International Conference on Software Engineering - Volume 1*, ICSE '15, pages 292–302, Florence, Italy, May 2015. IEEE Press.

A Included Papers

Acar et al. [2], Acar et al. [1], Acar et al. [4], Assal et al. [6], Baig et al. [11], Braz et al. [16], Cohney et al. [24], Danilova et al. [28], Danilova et al. [26], Danilova et al. [27], Derr et al. [29], Dietrich et al. [31], Fahl et al. [37], Fischer et al. [41], Fischer et al. [40], Fulton et al. [43], Gerlitz et al. [44], Gorski et al. [47], Gorski et al. [46], Hänsch et al. [54], Hallett et al. [52], Haney et al. [53], Krombholz et al. [66], Krombholz et al. [65], van der Linden et al. [112], Mhaidli et al. [71], Nadi et al. [73], Naiakshina et al. [77], Naiakshina et al. [76], Naiakshina et al. [75], Naiakshina et al. [74], Nguyen et al. [78], Nosco et al. [80], Oliveira et al. [82], Oltrogge et al. [83], Palombo et al. [85], Pashchenko et al. [86], Plöger et al. [88], Roth et al. [94], Ruef et al. [95], Sheth et al. [100], Tahaei et al. [102], Tiefenau et al. [106], Tiefenau et al. [105], Tuladhar et al. [110], Votipka et al. [116], Votipka et al. [115], Votipka et al. [114], Weir et al. [118], Weir et al. [117], Yakdan et al. [120], Yong Wong et al. [121], Yskout et al. [122], Yskout et al. [123]

B Database Structure

Figure 5 shows a simplified ERD of the database.

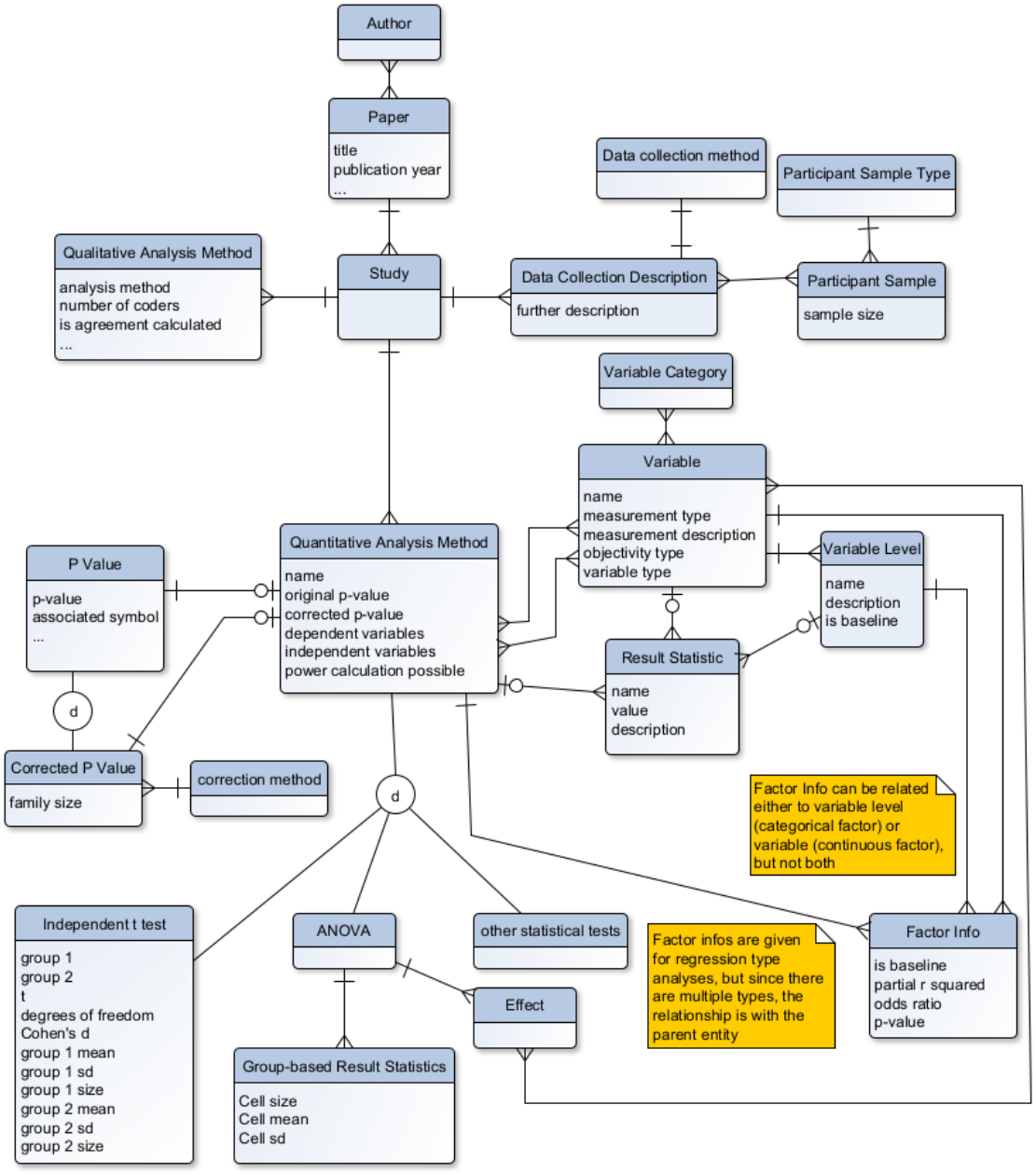

Figure 5: Simplified ERD of the database, not all attributes and entities are depicted.

C A Guide to Power Analysis for Hypothesis Tests with One Categorical Independent Variable with Two Groups

Four parameters are relevant to power analysis: Power, the significance criterion (i.e. the α error level), the reliability of the sample results or sensitivity of the test, and the effect size [22]. The power of a statistical test is the probability of the test correctly rejecting the null hypothesis, i.e. that a statistical test yields a significant result, when the alternative hypothesis is true [35]. Power can also be represented as $1 - \beta$, wherein β is the Type II error, i.e. wrongly rejecting the null hypothesis. This means that if a test has a statistical power of 0.8, as is an often used, acceptable value [22, 30], an actual effect will be detected 80% of the time. The significance criterion or significance level represents the threshold of maximum accepted probability of making a Type I error, i.e. wrongly assuming the alternative hypothesis, detecting an effect, when there actually is none [22]. Using the widely accepted threshold of 0.05 for statistical significance means that only in 5% of cases, an effect is detected in the sample, even though in the population, it does not exist. Reliability refers to how well a sample estimate represents the corresponding population parameter [22]. Reliability is influenced by different factors, depending on the type of estimated parameter, such as the quality of the measurement instrument, and controlling sources of variance in the data, which might distract from the effect you are trying to measure [35]. The largest and invariably present influencing factor, however, is sample size [22], such that larger samples produce more consistent and reliable estimates than smaller ones. Finally, the effect size measures the amount of impact of an independent variable on dependent variables, rather than only judging the presence or absence of an effect [35]. There are generally two types of effect sizes: Non-standardized, or simple effect sizes, which represent the size of effect in the units of the outcome variable, and standardized effect sizes which represent the effect size relative to the variability in the sample or population [10]. When comparing two means, e.g. with a t-test, the difference in mean completion time between two different interface variants represents a simple effect size, measured in units of time, e.g. minutes, while a standardized effect size for this scenario, such as Cohen's d, takes into account the standard deviation in the two groups. Standardized effect sizes are commonly classified as either belonging to the d-family, such as Cohen's d in the example above, or as belonging to the r-family, such as the correlation coefficient Pearson's r [92].

These four parameters are interdependent, such that when three of them are available, it is possible to calculate the fourth. Such calculations are referred to as power analysis. In general, there are four different kinds of power analysis, each used to determine one of the parameters from the other three, although it is also possible to determine both α and power if a ratio for α and β is given together with the other two parameters - this is termed compromise power analysis [38]. The other four flavors are summarized, e.g. by Cohen [22] in Chapter 1.5.

The tutorials on the companion website[5] provide an overview of the data necessary to conduct power analysis for basic hypothesis tests, where to find this data in our database, and how to use it to conduct power analysis using G*Power or R.

[5] powerdb.info/