A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen's d Effect Sizes

Anna-Marie Ortloff
ortloff@cs.uni-bonn.de
University of Bonn
Bonn, Germany

Julia Angelika Grohs
s6jugroh@uni-bonn.de
University of Bonn
Bonn, Germany

Simon Lenau
lenau@cispa.de
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany

Matthew Smith
smith@cs.uni-bonn.de
University of Bonn
Fraunhofer FKIE
Bonn, Germany

ABSTRACT

Researchers often place a strong focus on statistical significance when reporting the results of statistical tests. However, effect sizes are reported less frequently, and interpretation in the context of the study and the research field is even rarer. These interpretations of effect sizes are, however, necessary to understand the practical importance of a result for the community. To explore how Usable Security & Privacy (USP) and HCI researchers interpret effect sizes and make judgments on practical importance, we conducted survey and interview studies with a total of 63 researchers at CHI and SOUPS 2023. Our studies focused on Cohen's d and odds ratios in two USP and one HCI scenario. We analyzed which artifacts researchers consider when judging effect size, and found misconceptions and variation between the participants, highlighting how difficult judging statistics can be. Based on our findings, we make concrete recommendations for improved reporting practices around effect sizes.

CCS CONCEPTS

• **General and reference** → *Surveys and overviews*; • **Human-centered computing** → **Empirical studies in HCI**; • **Security and privacy** → Human and societal aspects of security and privacy.

KEYWORDS

meta-science, effect size, interpretation, odds ratio, Cohen's d, interviews, survey

ACM Reference Format:
Anna-Marie Ortloff, Julia Angelika Grohs, Simon Lenau, and Matthew Smith. 2025. A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen's d Effect Sizes. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3706598.3714022

CHI '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714022

1 INTRODUCTION

Interpreting effect sizes is a vital aspect of quantitative research, which is often overlooked. It is common for researchers to focus on p-values and statistical significance rather than interpreting effect sizes to judge the practical importance of the results. In many cases effect sizes are not even reported in publications in Human Computer Interaction (HCI) in general [38, 43], or Usable Security and Privacy (USP) [26] specifically. As researchers in this domain, focused on using the statistical tools we have to answer our domain-specific research questions, we understand this issue: Deciding which effect size to use for a statistical test is often not straightforward, with e.g. multiple effectsizes available for a statistical test like an ANOVA: e.g. generalized or partial η^2, $omega^2$, and f^2 [46]. We have encountered this issue in many situations: When deciding which effect size measure to use and how best to report it, when trying to compare findings across studies or when deciding which effect size to use in a power analysis based on prior work.

Interpreting effect sizes can be challenging. There are two major kinds of effect sizes. Simple effect sizes show the size of the effect in the units of the dependent variable [5], e.g., the mean time difference between two groups in seconds. Unitless effect sizes are independent of the units in which a variable is measured. They include standardized effect sizes and risk estimates such as odds ratios where variables are categorical. Standardized effect sizes are a form of unitless effect size normalized with the sample variance [5], e.g., Cohen's d. Simple effect sizes might be more intuitive to interpret but cannot be compared directly between studies using different units of measurement. Unitless effect sizes are theoretically more comparable between studies, but as we show, in practice they are less intuitive. To make matters worse, the scales of the different unitless effect sizes are not uniform, meaning that researchers must develop a feeling for each of the measures separately. This variety makes it harder for researchers to interpret these values [21] and may be a reason why they are not used frequently [26, 38, 58]. Our results show researchers can use the scale of one measure to judge the effect of another by mistake. Even when effect sizes are

reported, they are often not interpreted or merely judged using a set of standard guidelines such as Cohen's small, medium and large ranges for d [15]. A contextual discussion of the importance of an effect for the research subjects or the community, i.e. the practical relevance of the effect, is mostly missing.

We aim to understand how USP and HCI researchers interpret and judge the importance of effects, to improve effect size interpretation in USP and HCI papers. Research in HCI is very diverse, such that tools and methods are adopted from many other fields [64, 65]. USP can be considered similar in the adoption of methods, and is a comparably new field, which faces the additional challenge in empirical work, that security or privacy are often not users' primary task [20]. We chose USP as a subfield of HCI for this study, since we are most familiar with the domain and thus have an overview of methods used and possible research questions in this field. We also believe that HCI researchers in general have some experience with USP topics through overall exposure online. Finally, making wrong decisions in the domain of USP technology leads to obvious and direct harm, which we used in our study design to explore how differing levels of criticality influence effect sizes judgments. In this paper, we present the results of interviews and surveys conducted at the the ACM CHI Conference on Human Factors in Computing Systems (CHI) and the Symposium on Usable Privacy and Security (SOUPS) in 2023 about HCI and USP researchers' interpretations of odds ratio and Cohen's d effect sizes. We picked these two effect size measures because they are relatively common in our research domain and, in our opinion, offer a good balance between complexity and straightforward study design, meaning that we could present them in a reasonable time in a user study. We explicitly decided to use effect sizes which we have seen used in the USP literature, rather than evaluating understanding of effect sizes proposed as easier to understand [11, 29, 50, 68], but not used as much. We started by investigating the following research questions:

RQ1: Which factors influence researchers' judgments of effects
a) regarding the size of effects?
b) regarding the importance of effects?

We use aspects of constructive grounded theory [14], e.g. adapting our survey and interview methods to investigate new questions arising during the data collection. The surveys focused on RQ1 a) and b) and specifically the relationship between effect size and how important an effect is in practice by presenting between one and two scenarios and asking participants for their views on these aspects. In our first interview study, we explored the judgment process in an in-depth way to gather insights into RQ1. In the second interview study, we focused on a new research question:

RQ2: What are researchers' misconceptions about the Cohen's d and odds ratio effect sizes?

We found that researchers' perception of effect size did not always follow common standards for size judgment such as those proposed by Cohen [15]. Various descriptive statistics, p-values, participant numbers, and the effect sizes themselves played a large role when participants interpreted the research results in our vignettes. Beyond these artifacts presented in the vignette, participants used context and the point of view of those affected by study results for their judgment. We find the aforementioned aspects of effect size judgment to be important, but often missing from reporting of results. Conversely, many of the misconceptions about Cohen's d and odds ratios in our studies can be traced back to confusion over the variety of existing effect sizes and differences in their interpretation. We make several recommendations on how effect size reporting can be improved in USP and HCI papers, to lower the risk of misinterpretation.

2 RELATED WORK

We present theoretical background on the concept of effect sizes and their use in HCI and USP and summarize related work on understanding effect sizes.

2.1 Theoretical Background

In quantitative analysis, researchers investigate the relationship between one or more independent, or predictor variables, and one or more dependent or outcome variable(s). Effect sizes measure the strength of the relationship of independent variables with dependent variables [21]. There are many different types of effect size [45]. Simple effect sizes report the size of the effect in the units of the dependent variable [5], or directly derived from these units, e.g. as a percentage difference. In contrast, unitless effect sizes are independent of these units and aim to be comparable across studies. Types of unitless effect sizes are standardized effect sizes, which are normalized with the sample variance [5], e.g. Cohen's d, or risk estimates such as odds ratios where variables are categorical. Another categorization of effect sizes differentiates between the d-family, considering group differences, like Cohen's d, the r-family, considering measures of association, like Pearson's r [67], or risk estimates comparing different groups, like the odds ratio [22]. Especially for risk estimates, the base rate of the risk is important to evaluate the meaningfulness of an effect [22]. Some effect sizes additionally incorporate corrections, e.g. for sampling variability (adjusted R^2) or for shared variance (partial r) [22].

While there are equivalents of effect size indices used in other forms of statistical analysis, such as Bayes Factors or regions of practical equivalence in Bayesian analysis [42, 49], null hypothesis significance testing (NHST) is commonly used in HCI [40]. Effect sizes as discussed in this work are often used in conjunction with NHST. To further contextualize effect sizes, we briefly summarize other concepts from NHST.

The most general form of NHST involves comparing the null hypothesis of no effect, which assumes no relationship between independent and dependent variables, and the alternative hypothesis, which does [21]. A p-value not larger than a specified α threshold serves as the criterion to decide whether the null hypothesis can be rejected. It represents the probability of getting results at least as extreme as those which were observed in the sample if the null hypothesis is correct [21]. Commonly the threshold of α=0.05 (5%) is used for this decision, although there is debate on the topic of p-values in general, and the threshold of 0.05 specifically [69, 81]. While the effect size is a point estimate, confidence intervals (CI) enable communication about the uncertainty of these estimates. Corresponding to the threshold of 5%, a 95% CI is the most commonly used type, including the true population value of

the estimated parameter in 95% of samples selected using a corresponding random sampling design [21]. The more precise an estimate is, the narrower the CI becomes.

Finally, we clarify terminology regarding significance of results. There are two types of significance, statistical significance, commonly associated with $p \leq .05$ in NHST, and practical significance, which refers to the consequences of an effect, e.g. for society as a whole. These two types of significance can, but don't have to correspond, which we elaborate using two example studies. Imagine a census-level study measuring 5$ more income per year, for unionized workers vs non-unionized workers. Due to the large sample size, this effect is statistically significant, but very small and not practically relevant. If this study was replicated by a student sampling their acquaintances, the same results are still not practically relevant, but will likely not be statistically significant either. In contrast, inspired by Kirk [44], imagine an early study on a new Alzheimer's treatment, where 8 out of 10 participants in the treatment group remember at least 50% more names of their friends and family compared to the placebo group. This effect improves quality of life for participants and thus is practically relevant but due to the small sample, the effect is not statistically significant. If this study is repeated with similar results in a larger sample, the effect will be both statistically significant and practically relevant.

In this work, we also use the related term importance when we refer to judgments of practical relevance because we deemed importance to be more subjective and wanted participants to include their frame of reference. This also avoids overloading the term significance and drawing participants' attention to the dichotomy of the two significances.

2.2 Effect Size Use

Even though organizations like the American Psychological Association (APA) [3] make recommendations regarding effect size reporting, these are not always applied in practice. In HCI, Nielsen and Levery were unable to conduct many formal analyses in their early meta-analysis on comparing usability measures, since the statistical reporting in their literature basis lacked sufficient detail [53]. Reviews of HCI publications and a meta-analysis in software engineering also noted issues in statistical reporting [12, 38]. Of 49 award-winning papers at CHI'20 which included hypothesis tests, only 8 (16%) reported either non-standardized or standardized effect sizes [43]. Similarly Salehzadeh Niksirat et al. found that reporting of descriptive statistics, clearly stating the test procedure, and reporting of test statistics and p-values was largely sufficient at CHI, but reporting of effect sizes and even more so, CI, was lacking [70].

Prior work on completeness of statistical reporting in USP from 2006 to 2016 found that many of the reports were incomplete, as well as not APA-compliant [26]. In a survey of work in developer-centered usable security from 2010 to 2021, even simple hypothesis tests were often not reported in sufficient detail to be able to conduct power analysis [58].

Guidelines for interpreting effect sizes exist [15, 66] but are criticized as being insufficiently context-focused or not taking into account other influencing factors like the study method [22, 76]. Even Cohen himself cautioned against using the guidelines indiscriminately when proposing them [15](p.532). This warning has largely gone unheeded. There have been proposals for adjusted guidelines based on actual effect sizes common in specific research areas [10, 55, 61] but not yet in USP. Plonsky and Oswald describe considerations when interpreting effect sizes, like the methodological quality of the study, the maturity of the research domain and practical significance [61].

2.3 Metastudies about Effect Sizes

Metascience analyzes research methods and current practice and aims to improve them [36]. Le Pochat and Joosen provide an overview of metaresearch conducted in security [48]. Communication of results is discussed in this work but more in the context of vulnerability disclosure and publication bias [48].

Prior work investigated understanding and misconceptions about various statistical concepts in different domains [13], such as p-values [28, 80, 82] and effect sizes in medicine [82], and power [17] and confidence intervals [18, 32] in psychology. Participants in these studies were undergraduate or graduate students [13, 18, 32], researchers [17, 32], statistics teachers [28] or doctors [80, 82].

Focusing on effect sizes, Hanel et al. asked their layperson participants for their subjective rating of informativeness of five different effect size measures, the Bayes factor, and standard significance statements and found that Cohen's U3, which is the proportion of the second group, which has smaller values than the median of the first group, was rated most informative effect size, followed by the probability of superiority [29]. Probably due to habituation, significance reporting without an effect size was also rated very informative [29].

Understanding of effect sizes is often studied through visualizations, since they are a more accessible form of communicating results than inferential statistics. Only presenting averages without estimates of uncertainty can lead to participants overestimating the effectiveness of interventions [37, 43]. Variability measures aim to improve this, and those that could be explained in an accessible way performed better, although prior work does not agree which variability measure is best [33, 37, 43]. Hofman et al. found that making the variation in individual outcomes apparent, in the form of an animation of multiple possible samples from an underlying distribution led to the best estimates [33]. On the other hand, Kale et al. found that showing densities, without explicitly specifying the means was most effective [37], while in Kim et al's study, the probability of superiority performed best [43]. Kim et al. also found that presenting results through simple analogies with well-known phenomena (like height differences by age) was able to reduce misperceptions [43]. Prior work comparing two tasks involving effect sizes, size estimation and decision-making, found that participants performed differently at these tasks, with low scores in the first task not necessarily corresponding with low scores in the second [37]. This suggests that effect size interpretation differs depending on the task [37].

While prior work focused on specific statistical concepts in isolation, or researchers' and laypersons' misperception of effects, we provide insight into how researchers in USP and HCI judge effect sizes, when they are presented in the context of a result section, and we compare a security-related research question to a non-security related one.

3 EFFECT SIZE REPORTING AT SOUPS AND CHI'23

To provide a baseline of effect size reporting in the communities where we conducted our research, we analyzed effect size reporting in 33 SOUPS and 47 CHI USP papers published in 2023. We focused on USP publications because our vignettes also originate from this context. We identified the CHI USP papers by searching the conference proceedings for sessions and papers including including the keywords "security" and "privacy". For retrieved sessions, we included all papers in our sample, individual papers were included if they contained the keywords in the title or if we judged that the abstract showed that privacy or security were a main topic of the work, i.e. part of the research question. We excluded types of content other than papers, i.e. we did not consider workshops, case studies, journal presentations or similar.

In this sample, 43 papers (CHI:28, SOUPS:15) reported a hypothesis test, with 30 (CHI:22, SOUPS:8) reporting at least one unitless effect size. Of these, 13 (CHI: 10, SOUPS: 3) provide some size judgment for the effect. For example Kanei et al. [39] reference a methodological publication when they write that they "considered *phi* or Cramer's V >.10 as small effects, >.30 a medium effect, and >.50 a large effect [14]"(citation in quote from original paper). Sometimes effect size judgments are present without such references to the scale used. Depending on the readers' familiarity with effect sizes, they can then determine that most frequently and as in the quote above, Cohen's guidelines for effect size judgment [15] were used.

Only 4 CHI papers offered further interpretation of the effect size or a justification of the size judgment. For example, Borgert et al. [9] discussed the impact of their significant age effect in depth: "Keeping in mind that the underlying regression including every demographic factor only could explain 5.1% of variance in CySESH scores and our high sample size likely was the only factor that enabled detection of this very small effect, we argue that the age effect is not large enough to imply generalizability problems."

We consider our findings in this sample an upper bound on the proportion of effect size reporting. While we assessed whether papers reported or explained unitless effect sizes at all, this does not necessarily apply to all the analyses within the publication. Effect size reporting in our samples is in a similar range to [70] in 2023 and suggests an improvement from 2020 [43]. However, in our sample, deeper discussion of effect sizes and size judgement justification was sadly not common.

4 STUDIES

We conducted two iterations of our study (two surveys (**-s**) and two interviews (**-i**) per study) to get multiple angles of insight into how USP and HCI researchers interpret and understand effect sizes, specifically odds ratio (OR), and in the second iteration, Cohen's d. We chose OR because we believed it to be one of the common and more simple unitless measures since it is based on amounts and percentages, however many participants in our first iteration had trouble with it, so we added Cohen's d as another commonly used effect size measure in our second iteration. Table 1 has an overview of the study iterations and the research questions at the focus of each and Figure 1 gives an overview of the process of the studies.

We recruited HCI and USP researchers at CHI for the first study iterations S-CHI-i and S-CHI-s, and SOUPS for our second iterations S-SOUPS-i and S-SOUPS-s in 2023. Our participants saw a vignette with information about a research question, a corresponding study, and results, including a description of the effect size. After reading the information, we asked the participants about their interpretation of the presented effect, its size and importance. After analyzing the results from these studies, we conducted an interview with a statistics consultant to consolidate our findings. Due to continued contributions after the interview, the consultant later joined us as a co-author on this work. In the following, we describe the design and development of our study and our findings from multiple iterations.

4.1 Positionality

Researchers' experience, knowledge, and beliefs contribute to the interpretation and planning of research [23, 57, 71]. In describing our methods, we try to make it clear when our background influenced research decisions. R1 and R2 conducted and analyzed the interviews. R1, R2 and R3 planned the studies and interpreted the results. R4 (P-E as a participant) contributed to the theoretical background and recommendations, after joining as a co-author. R1, R2 and R3 are part of a research group in usable security and privacy, i.e. at the intersection between the HCI and security communities. In this group, we use both qualitative and quantitative methods in our research, although R2 has more experience with qualitative research. R1 and R3 are also involved in teaching an undergraduate course including a crash course in inferential statistics, and, R1's research focuses on the use of research methods in USP, including effect sizes and power analysis. We consider ourselves as part of the demographic studied in this work and have fallen into many of the traps discussed in this paper. R4 works as a statistics consultant at a research organization, working with researchers like our participants. His scientific background is in sample planning and data quality for surveys.

4.2 Ethics

Our project was approved by the ethics review board at one of our institutions. We complied with the General Data Protection Regulation. Data was collected anonymously and we informed our participants about the study and our data collection process before the start of their participation.

Among the participants of each of the two recruitment phases, we raffled three times 100€. Participants could opt-in to submit their e-mail address to a raffle, which was not connected to the rest of our data collection, after finalizing their participation. They could enter in the raffle once per vignette they worked on. Interview participants entering the raffle were added twice to reflect their larger time commitment. For S-SOUPS-i, participants could also receive a small cuddly toy for their participation.

4.3 Measurement

We were interested in several concepts[1]. We asked participants for a **judgment of importance** about the presented effect, distinguishing whether it was *important* or *not important*. We also measured

[1]We mark important concepts in bold font in this section.

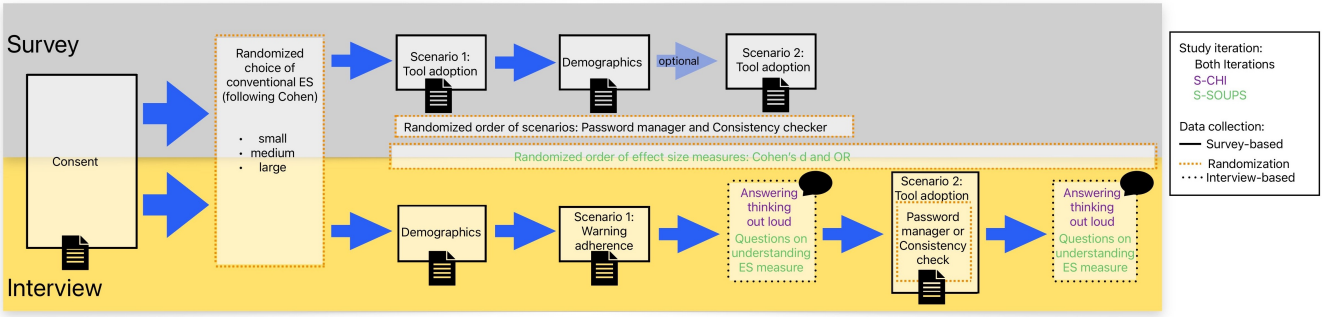


Figure 1: Process of survey and interview studies

Study	Recruitment	Target	Data collection	Research question	ES	Vignette
S-CHI-i	CHI'23	HCI researchers	interview	RQ1 general	OR	warning adherence & (password manager or consistency checker)
S-CHI-s	CHI'23, associated social media	HCI researchers	survey	RQ1 a) + b)	OR	password manager, consistency checker
S-SOUPS-i	SOUPS'23	USP researchers	interview	RQ2	OR, Cohen's d	warning adherence & (password manager or consistency checker)
S-SOUPS-s	SOUPS'23, associated social media	USP researchers	survey	RQ1 a) + b)	OR, Cohen's d	password manager, consistency checker

Table 1: Overview of the studies presented in this work

participants' **judgment of size** of the effects presented in the vignette, and its relation to the judgment of importance. Participants could judge the effect to be *small*, *medium*, or *large*. We honed in on this by asking for a numerical threshold where participants would start considering an effect to be important, and where they saw the borders of the three effect size categories. To ensure that participants had similar knowledge, we provided a basic definition of the effect size measure before they submitted their thresholds. To measure **confidence** in their assessment, we asked participants how sure they were about their judgment of the size of the effect, using a fully labeled six-point scale question.

We varied some parts of the vignette to explore possible influencing factors on participants' judgment. **Scenario** refers to the research questions and the study goal presented in the vignettes. We aimed to present vignettes of different criticality and used three different types of scenarios: *password manager adoption*, *consistency checker adoption*, and *TLS warning adherence* (for details see Section 4.4). To see whether participants' views on **criticality** matched our own, we asked them how severe they thought the negative outcome in the vignette would be, using a fully labeled five-point scale question.

We also varied the **conventional size** of effect presented in the vignette between *small*, *medium* and *large* according to Cohen's conventions [15]. We used these despite criticism of the guidelines [41, 73], to examine whether researchers' individual judgments match conventional ones.

The second scenario for the interviews was randomly assigned to either password manager or consistency checker. In the survey, we did not use the warning scenario due to its complexity. Instead, we randomized the order of the password manager and consistency checker scenario. The conventional effect size was randomized across participants but stayed the same within a participant, although we adjusted the numbers slightly so that they were not exactly the same.

We recruited researchers with different **research backgrounds**, either in *USP* or researchers in HCI *without a security background*, to see if these two groups interpreted the security and non-security related effect sizes differently.

4.4 Vignettes

We hypothesized that judgment of effect importance could be affected by the (security) criticality of the scenario, i.e. the harm or good that could befall the individual user, in addition to the effect

size and the number of people affected. Therefore we developed three different scenarios for the vignettes:

password manager concerning an intervention improving password manager adoption.

consistency checker concerning an intervention improving the adoption of a tool for consistency-checking for slide decks.

warning adherence concerning the efficacy of a scarier browser security warning for self-signed certificates.

In our view, the adoption of password managers is more critical than the adoption of a consistency checker since we judged the severity of the consequences of weak passwords to be worse than for potentially badly formatted slides. We added the warning adherence scenario, since the benefits of scaring users into warning adherence are debatable [30, 79], and we wanted to see whether this ambiguity affected participants' judgments.

For each scenario, we created three versions with different conventional effect sizes. The studies presented in the vignettes were not real results and we informed the participants about this before the study. The statistics presented were coherent, so inferential statistics matched descriptive statistics. The OR effect sizes were presented with the 2-way Fisher's exact test. A complete labeled example vignette is in Figure 2 and the full text of all varieties can be found in the supplemental material.

4.5 Conventional Effect Size

We chose values for small, medium and large effect sizes according to Cohen [15], in the lower-middle range for each of these categories. Due to the lack of well-known guidelines for OR, we converted Cohen's d to OR using Haddock's formula [27] with 1.18 as the recommended average correction factor [62].

Effect sizes, p-values and sample size are linked, but we wanted to isolate the effect of the changing effect sizes as much as possible, especially since related work suggests that p-values influence judgment on effect sizes [25]. We chose a number of participants that was high enough so that p would be <.001 for all levels of conventional effect size to avoid different statistical significance levels influencing participants' judgments. We chose the base rate of the positive outcome (technology adoption/warning adherence) so that participant numbers, though high, were still in a realistic range for end-user studies in USP.

4.6 Surveys and Interviews

The study process for survey and interview participants is shown in Figure 1. The survey questions are in the supplemental material.

The interviews are based on the surveys. Due to the interviews being conducted at a busy conference venue, and to reduce possible privacy concerns, we did not record audio during the interviews. Instead, the interviewer was assisted by a second researcher who took notes during the interview. We used a tablet to show participants the vignettes and questions and for them to answer demographic questions, but they answered our interview questions orally. To focus the interviews on the judgment of importance (RQ1b), we removed the threshold questions for the judgment of size and the corresponding confidence questions from the survey.

4.7 Recruitment

Our target demographic was researchers in USP and HCI. S-CHI took place at CHI 2023. We approached attendees for interviews at the conference venue and personally distributed flyers in public spaces. We selected potential participants of different ages, which we hoped would correlate with different career stages, and explicitly combined snowball sampling with approaching researchers we did not know. After the conference, we published our call for participants for the survey on the CHI Discord server with the permission of the Virtual Chair and to get access to more participants. Finally we published our call in a research-oriented Slack channel [78], which one of the authors was a part of. We asked participants to share the survey with their colleagues. We mostly conducted individual interviews, but in two cases (once at each conference) we conducted a joint interview with two participants by request.

4.8 Participants

Information on the participants are in Table 2. We interviewed one additional person in S-CHI-i, but removed their data from the study at their request. Some participants declined to answer part of the questions, and in joint interviews, we could not collect full demographic data, since the demographic questions were included in the survey used to present and randomize the vignettes and as such at most one of the participants in joint interviews could answer them in the survey. Consequently, numbers in the table will not always add up to the full number of participants for a study. Even though CHI is not a USP-specific venue, 9 of 42 participants in S-CHI had a research focus on security and/or privacy and we count them as USP researchers when discussing differences between the two groups. The rest did research in various other topics, with no field dominating.

4.9 Data Analysis

The two authors involved in conducting the interviews jointly analyzed the interview notes. We used a mixture of deductive coding based on our study design and measurement strategy, and inductive coding for other participant statements for S-CHI-i, e.g. when participants discussed approaches to making judgments that we had not considered in our study design. We discussed our findings with colleagues and refined the codebook based on this. The survey data were analyzed graphically and with summary statistics using R [63], and these summarizations were included in our qualitative analysis process. We refrain from making generalizations about the population, instead using the findings from the surveys to corroborate our interview findings. Since the results of our survey and interview studies complement each other, we report them together but specify the data source in the text.

4.10 Changes in S-SOUPS

We provide a brief summary of our findings from S-CHI here to contextualize the changes made to our survey and interview procedure for the second iteration of studies (S-SOUPS). We defer a more in-depth discussion of our findings from both studies to Section 5.

Vignette example	Structure
A study investigated the difference between the adoption of a **formatting tool, which enables consistency checking and adjustment of presentation slides**, in a baseline group that received a general introduction to the tool and an intervention group that additionally was informed that **professional-looking slides had a greater impact.**	Research question and study design
2100 Participants used the **formatting tool** during the study. Two weeks after the end of the study, they were asked whether they were still using the **formatting tool** or not.	
511 out of 1050 (**48.7%**) of those who received the intervention were still using it.	Descriptive statistics
224 out of 1050 (**21.3%**) of those who did not see the intervention were still using it.	
Fisher's exact test showed that this difference was statistically significant (p<0.001, odds ratio = **3.50**, 95% CI=[**2.88, 4.25**]). The effect size (Odds ratio) is **3.50**.	Inferential statistics
This means that the odds for the participants to continue to use the formatting tool in the intervention group were **3.50** times higher than for participants to continue to use the **formatting tool** in the baseline group.	Effect size explanation - only present in S-CHI

Figure 2: Example of the vignette for a conventionally large odds ratio effect size in the consistency checker scenario. The bolded content in the vignette represents variable values which change depending on the conventional effect size or scenario.

study		S-CHI-s	S-CHI-i	S-SOUPS-s	S-SOUPS-i
N participants		30	12	9	12
research method	quantitative	6	2	1	2
	qualitative	5	3	2	3
	both	18	6	6	6
career stage	Master's	1	2	-	-
	Phd	18	2	5	4
	Post-doc	3	1	-	4
	Professor	3	4	4	4
	Industry	4	1	-	-
effect size experience	Cohen's d	16	7	4	6
	Odds ratio	5	3	4	4
	other	8	2	3	3
	none	11	1	1	4
research topic	USP	5	4	9	12

Table 2: demographics of study participants

The size of effect was a prominent factor for interviewees when judging the importance of the effect, but size and importance thresholds varied widely across participants. Participants also took into account other factors.

Since only 8 out of 42 participants in S-CHI had experience with OR, we extended the vignettes to present two effect size measures, Cohen's d or OR in the vignettes, for the second iteration (S-SOUPS). We chose Cohen's d because it was the most frequently named effect size in S-CHI and we were interested in getting more insight into researchers' judgment of a unitless effect size where guidelines for judging size are common.

The vignettes presenting Cohen's d reported a t-test and additionally included the appropriate test statistic and degrees of freedom since this is commonly reported with this type of test. Descriptive statistics for Cohen's d were means and standard deviations instead of amounts and percentages. The order of effect size measure was randomized, and if a participant worked on two vignettes, they were exposed to both measures. We did not include an effect size explanation in the vignette, since Cohen's d does not lend itself to a statement that doesn't already include a full explanation of the effect size, i.e. introducing the standard deviation as the unit of difference between means. To be consistent, we removed the explanation from the vignette for the odds ratios as well. However, we gave explanations before asking participants to specify boundaries for the size and importance categories:

Cohen's d: Cohen's d measures the difference between two means, normalized using the standard deviation, so that a Cohen's d of 1 represents a difference of 1 standard deviation between the two means. Cohen's d of 0 means the means in both groups are the same. The higher Cohen's d is, the larger the difference between the means is.

OR: An Odds Ratio (OR) is the ratio of two odds. An OR=1 means the odds of an outcome are the same in both groups. An OR >1 means the odds are higher in one group than the other. An OR between 0 and 1 means the odds are lower in one group than the other.

During S-CHI-i we also noticed our participants' uncertainty when discussing effect sizes. In the responses to the threshold questions in S-CHI-s, we identified misconceptions about OR, e.g. values that are not in the range of possible values OR can take on, or submitting the minimum (null effect) or maximum as thresholds for a size category, or importance. To explore this in more depth, we added a second research question: *What are researchers' misconceptions about the Cohen's d and odds ratio effect sizes?* and updated our interview procedure.

There were two blocks of questions in our interview guideline for S-SOUPS-i, which is in the supplemental material. The first block focused on participants' understanding of the effect size measure used in the vignette (RQ2), and probed participants to explain it, e.g. by asking about its minimum and maximum value. This part was skipped if the participant indicated that they had no knowledge of the effect size. To debrief participants, we then showed the explanation used in the survey and asked them if there were any changes to their understanding and to summarize the explanation in their own words. The second block of questions was similar to the questions in S-CHI-i about the judgment of importance to get more data for RQ1b. Additionally, we asked for a research question in the participant's research area, where even a small change would be important to find, to get participants thinking about the practical relevance of effects.

Finally, some interviewees from S-CHI-i showed uncertainty about the phrasing in our question measuring criticality, so to alleviate this we changed the phrasing for S-SOUPS.

4.11 Recruitment and Participants in S-SOUPS

We conducted S-SOUPS at SOUPS 2023, and followed a similar approach there, as for S-CHI. SOUPS maintains a Slack channel for communication between conference attendees, which we used to recruit for the survey, in addition to flyers.

General information on the participants from S-SOUPS is also in Table 2. Since we recruited at SOUPS, we consider all participants as researching in the field of USP, but we asked them to specify a sub-field in an open question, and also received a wide variety of answers.

We followed a similar data analysis approach as in S-CHI and re-used our codebook in S-SOUPS where applicable, but due to the different focus of our interviews, we also inductively added new codes.

4.12 Expert Interview

After data analysis for S-CHI and S-SOUPS, we consolidated our findings in an interview with a statistics consultant who works with USP researchers. We introduced him to our study methodology and the results from S-CHI and S-SOUPS and then discussed how the researchers who consult with him handle interpretation of research results, and how their misconceptions and interpretation problems compare to our findings. After he expressed interest in the topic during the interview, we invited him to collaborate with us on this project. He explicitly consented to the de-anonymization that his co-authorship introduces.

Our expert participant has a background in survey statistics, specifically sample planning and data quality, with a Master's degree in statistics, and a PhD in economic and social statistics. He works as a statistics consultant affiliated with a research organization. In this role, he initiates methodological research and advises PhD students and faculty on statistical and methodological questions in their own research. We refer to him as P-E throughout the results and present findings from the expert interview directly with the other results.

5 COMBINED RESULTS

In the following, we present results from both iterations of our study, as well as the expert interview. We explored the relationship between participants' judgments of size and conventional effect size guidelines (RQ1a), factors influencing participants' judgments of the importance of effects (RQ1b) and comprehension of the effect size measures (RQ2). For the analysis of the survey data for RQ1, we filtered out all judgments concerning a particular effect size from participants who had misconceptions about that effect size, since those judgments are likely not reliable. Out of the original 60 vignette responses from 39 participants, we retained 43 responses from 28 participants. We conducted a secondary analysis of the responses from those participants with misconceptions and selectively report interesting findings from this analysis. To analyze misconceptions (RQ2) we retained all 60 responses.

5.1 Understanding of Effect Size (RQ2)

In our survey demographics, we asked participants which effect sizes they used in their own research to get an understanding of their experience with these statistics. We describe the experience our participants had with using OR and Cohen's d in Table 2. For all groups of participants, Cohen's d was used as or more frequently than OR. Some of the other effect sizes used by our participants in S-CHI include variants of η^2, commonly used with ANOVA-type analyses (4 in S-CHI-s, 1 in S-CHI-i), regression coefficients (2 in S-CHI-s, 1 in S-CHI-i), and correlation measures (2 in S-CHI-s, 1 in S-CHI-i). Bayesian effect sizes, phi, R^2, and Cohen's f each were named once. In S-SOUPS, participants also used effect sizes besides the ones our study focuses on, e.g. omnibus effect sizes used in regression, namely R^2 (2 in S-SOUPS-i), correlation measures and regression coefficients. We provided selectable options for the effect sizes in our study, OR and Cohen's d, but all other mentions were self-reported, which could have lead to underreporting for these other effect sizes, especially where it might not have been clear what statistics can be considered as effect sizes. This may be the case, e.g. for correlation coefficients, which are reported as both test statistics and effect sizes in publications. One participant in S-CHI-s also discussed simple effect sizes, but did not see these as effect sizes, answering "none. I mostly use means and sd". On the other hand, a participant in S-SOUPS-s took the contrary view on the value of simple vs. unitless effect sizes, stating they use "direct, real-world measures that aren't 'standardized'". One of the S-SOUPS-i participants also mentioned using non-standardized effect

sizes. Even if participants don't use effect size measures themselves, the encounter them when reading literature, so we believe that understanding of effect size is important regardless.

We found misconceptions and misunderstandings of effect size measures in both interviews and surveys. In the surveys, we gave participants an explanation of the effect size measure before asking them to specify the upper and lower thresholds for what they would consider a medium sized effect. We also asked them to specify from what threshold they would consider an effect important. OR can be expressed as values between 0 and 1 and equivalent values over 1. For better comparability, we asked participants to submit their thresholds as values over 1. Cohen's d can take on values between $-\infty$ and ∞, although often, absolute values are <1. If there is no effect at all, OR=1 and Cohen's d = 0.

Of the three thresholds participants submitted, there were 26 values that indicate misconceptions. These include 5 OR values below 0, which are not valid values for this statistic, six instances of using OR=1 as a threshold, even though OR=1 is the null effect, and six instances of using OR=0 as a threshold, although this is a maximum for odds ratios. Three participants gave OR-thresholds below 1 even though we had instructed them otherwise. While these could have been valid OR thresholds if used consistently, this was not the case. Among all three participants, the importance threshold was submitted as a value smaller than 1, and at least one of the size thresholds were values above 1, so the participants were not consistently considering a different baseline category when submitting the thresholds. Two of these submissions also included other misconceptions. A higher value submitted for the lower threshold of a medium effect than for the upper threshold appeared three times for OR and once for Cohen's d. For Cohen's d, we also judged one occurrence of a threshold d>3 as a misconception since such effect sizes are very uncommon. There were 10 participants with at least one misconception for OR, 1 with a misconception for Cohen's d and 1 with misconceptions about both effect size measures, so 12 out of 39 participants displayed misconceptions and 27 did not. Participants' submission could entail more than one misconception.

We compared subjective measures of how certain participants were in their judgment of effect sizes to the number of misconceptions in their submissions. We calculated an average over the three different instances we asked for a participants' confidence in their answer for Figure 3. It shows that regardless of whether submissions include misconceptions, participants are insecure about their effect size judgments. While comparatively more qualitative researchers are exhibiting high levels of uncertainty, the number of mistakes made is not larger than for researchers using more quantitative methods. Confidence does not appear to be associated with the number of misconceptions as an objective measure of participants' understanding, e.g. one of the participants with the most misconceptions nevertheless was sure of their answers. Many participants in S-CHI-i and S-SOUPS-i also expressed insecurity, e.g. asking questions about effect size or stating that they were merely guessing in their judgments. When asked to explain the meaning of the effect size of the study in their own words in S-SOUPS-i, some participants were not able to state anything concrete, even when prompted for the value in case of a null effect. These participants were then shown an explanation of the respective effect size, instead of later in the interview, to avoid causing frustration.

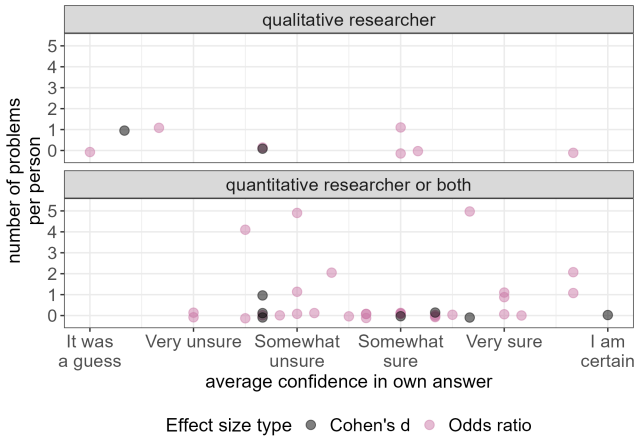


Figure 3: Number of identified problems in submitted Cohen's d or OR thresholds per participant vs. average confidence rating per participant over all submitted thresholds, grouped by participants' research focus. We juxtaposed qualitative researchers with all other research focuses. This includes 23 researchers using both qualitative and quantitative methods, seven purely quantitative, one primarily Bayesian, whom we counted as quantitative and one researcher who did not answer this question. Points are only jittered vertically.

In S-SOUPS-i, we recognized some misconceptions from the surveys when prompting our participants for minimum and maximum possible values of a type of effect size. We identified a broad theme of unclear value ranges, where participants were generally unsure about the maximum possible value for either effect size. For OR the minimum value being 1 was unintuitive for participants, and we saw cases where OR<1 were believed to be smaller than OR=1. P-E hypothesized that some of the confusion was due to mistaking odds ratios for log odds ratios, where "what one is for odds ratio, is zero [for log odds ratio]". Participants also explained OR as a ratio of probabilities, instead of a ratio of odds, referring to "likelihood" in their descriptions, e.g. OR=3 meaning that one group is 3 times more likely to have a specific outcome than the other. However, depending on the base rate of the outcome, the odds ratio and the probability-based risk ratio can differ substantially. Imagine a two-group study with 4 participants per group, where 1 out of 4 in group A receive a negative outcome and 2 out of 4 in group B also do so. The odds for a participant in group A to get a negative outcome is 1/3, while the risk is 1/4. For a participant in group B, the odds to get a negative outcome is 2/2, and the risk is 2/4. Consequently, the OR here is $\frac{2}{2}/\frac{1}{3} = 3$, while the risk ratio is $\frac{2}{4}/\frac{1}{4} = 2$. Higher base rates of outcomes lead to larger differences between odds ratios and risk ratios [59]. This misconception is well-known in medicine and public health research [34, 59]. Several participants confused Cohen's d with Cohen's κ, a measure of inter-rater reliability, e.g. in qualitative analyses. For Cohen's κ, 1 means maximum agreement between coders, with values between 0.8 and 1.0 conventionally discussed as almost perfect agreement [47]. For each effect size measure, at least one participant also saw a direct relationship between

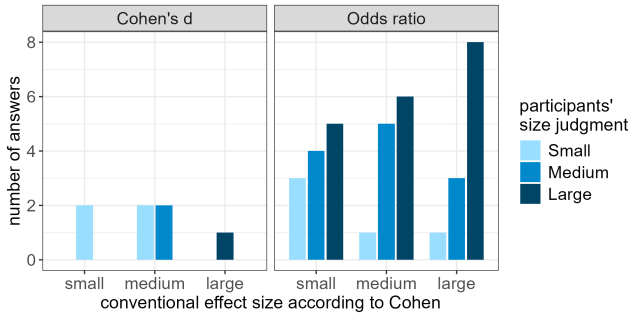


Figure 4: Comparison of two ways to categorize effect size: participants' judgment as selected from the options in the survey (depicted by different color, see legend), and Cohen's conventions (x-axis), converted for OR (right)

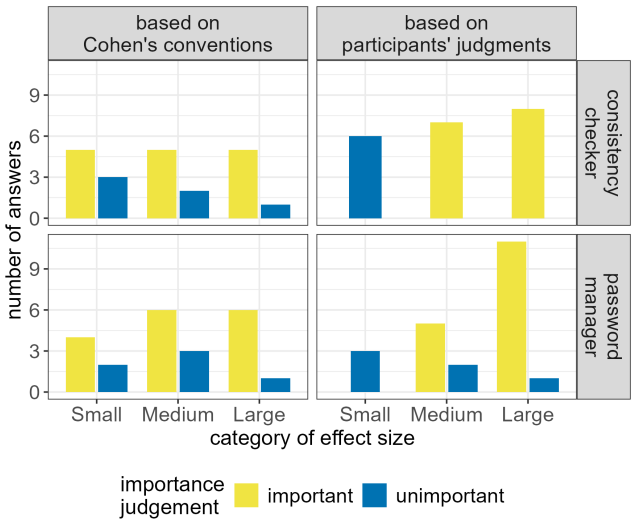


Figure 5: Comparison of judgment of importance based on Cohen's conventions (converted for OR, on the left) and participants' judgment of size (right), grouped by scenario

the sample size and the effect size, which is also a misconception, in that a larger sample size does not change the value of the effect size, but rather the certainty of the estimated effect size, i.e. the confidence intervals around the effect size become smaller.

Our explanation of the effect size statistics helped some of the participants in S-SOUPS-i. Six interviewees who had not been able explain Cohen's d in their own words felt they understood the effect size better after our explanation, and an additional interviewee correctly updated their initial understanding. Understanding of OR improved for 4 interviewees who did not know OR and 1 person who did.

P-E described another misconception, where researchers "equate an effect size in the sample [...] effect sizes which could be found in the population". This is exacerbated by the fact that truly random sampling is rare in the behavioural sciences [31], and especially for specialized populations like software developers, administrators, researchers, etc. [1, 7, 52]. This means that the effects seen in studies may not generalize. P-E suggested using confidence intervals in addition to p-values, to enable readers to gauge uncertainty and counter this misconception.

5.2 Judgment of Size of Effects (RQ1a)

Participants' judgment of size of effects in the surveys is compared to Cohen's effect size conventions [15] in Figure 4. For OR, we converted the guidelines from Cohen's d to OR, so they are probably not familiar to participants. Here, the participants' judgment does not match with the conventions. Instead, all three conventional size categories are represented in all three categories as judged by the participants. For Cohen's d, where the conventions apply directly, our participants' judgment mostly corresponded to them. This suggests that well known conventions for size judgments may override participants' own judgment of effect size. This could be problematic since Cohen himself stated that his thresholds had "no more reliable a basis than [his] own intuition" and should only be used if no other options are feasible [15], p.532.

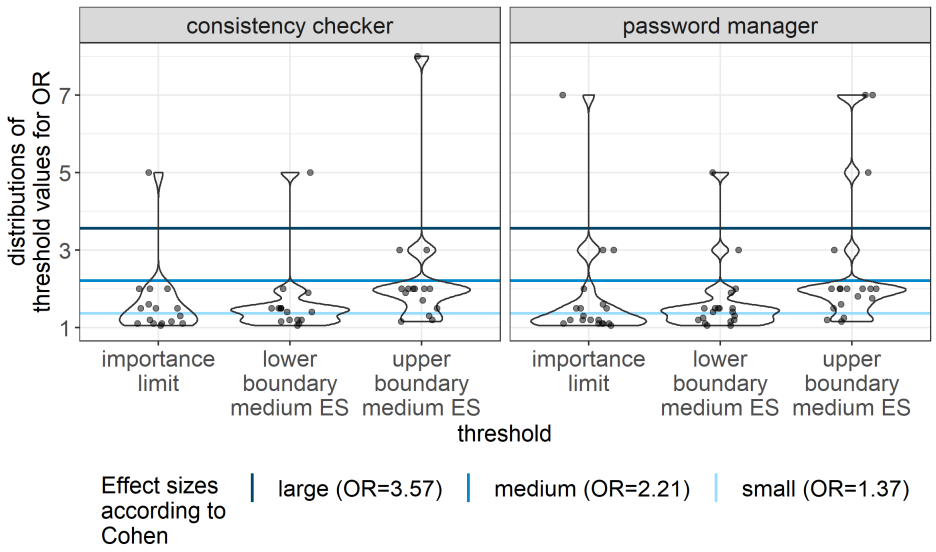
5.3 Judgment of Importance of Effects (RQ1b)

We investigated factors which could influence participants' judgment of importance for our vignettes: size of effect, context (by using the three scenarios), and consequences of participant behaviour, i.e. criticality. Through the interviews, we identified two other factors: point of view, and other numbers in our vignette besides the effect size (e.g. p-values or descriptive statistics).

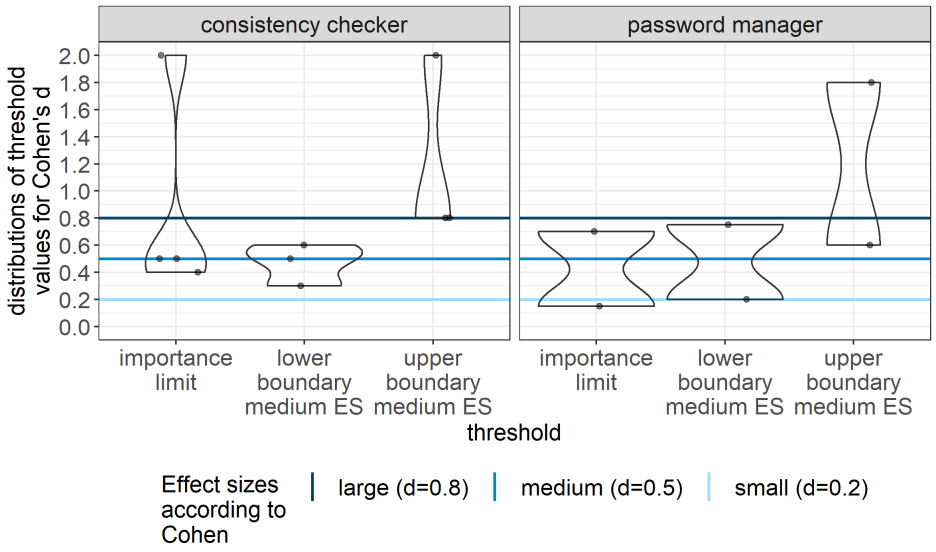
5.3.1 Size. Based on the surveys, the right side of Figure 5 suggests that for their own judgment of sizes, participants largely judge small effects as unimportant and large and medium effects as important for our vignettes. Considering participants with misconceptions (see Figure 7), we find that this observation holds, as effects judged to be large or medium by the participants tend to also be considered important. Since we used OR>1 in our scenarios, the common assumption of larger absolute values being better would hold even for less understanding of the effect size. However, the relationship is not true when applying Cohen's standard. Since participants' judgment of effect size does not always align with Cohen's guidelines, here even small effects (as per Cohen) were considered important. Nevertheless, in the vignettes the tendency towards large effects being judged as more important and small effects as less important is visible on the left side of Figure 5, too. The data from our vignettes suggests that a larger effect size is associated with more importance, but mainly based on participants' personal judgment of sizes instead of Cohen's.

However, when considering the thresholds for important effects participants specified, we see that both for odds ratios (Figure 6a) and Cohen's d (Figure 6b), the threshold for what participants would consider an important effect in the given scenario is sometimes lower than conventional medium or even small effects. In general, importance and size thresholds for Cohen's d, and OR vary widely within our sample.

Since we found this very interesting, in S-SOUPS-i we explicitly asked participants to think of a research question where detecting even a small effects would be important. In the end, 9 of 13



(a) OR effect size, conventions converted from Cohen's d, 4 values of 10 or above excluded from figure



(b) Cohen's d effect size

Figure 6: Distributions of submitted thresholds for importance, lower and upper boundary of medium effect size (ES), for different vignette scenarios and ES measures

participants were able to describe such a scenario. Examples were detecting ads that are inappropriate for minors, increasing confidence in cyber security through games, or noise and real data being similar enough that an attacker cannot distinguish between them. However, three participants stated that in their field, only small sample sizes are possible, so small effects are unlikely to be detected, three participants explicitly stated that small effects could never be important in their field, and four participants had difficulties describing such a scenario. Participants could fall into multiple of these groups.

5.3.2 Context. To investigate the influence of context on judgment of importance, we chose different scenarios for our vignettes, which we thought would affect participants judgment. When constructing the scenarios we judged a small improvement in password manager adoption as an important achievement, while we considered any size of improvement in a slide consistency checker less important. This is due to our research focus on security and privacy. However, comparing the top and bottom row of Figure 5 shows that our participants saw things differently and highlights our personal bias in the choice of scenarios. For both scenarios, six participants

Figure 7: Participants judgments of effect size and importance, grouped by scenario and research background. Sample: Only participants with misconceptions

judged the presented effect to be unimportant, and 15 found the effect in the consistency checker scenario important compared to 16 in the password manager scenario. To our surprise, 5 of the 6 participants considering the effect in our password manager scenario unimportant had a research background in USP.

Judging by their submitted importance threshold, three of the 15 S-CHI-s and S-SOUPS-s participants exposed to both scenarios, found the effect more important in the consistency checker scenario, and six in the password manager scenario, but six chose the same importance threshold in both scenarios. Interestingly, when only taking into account participants with misconceptions, the effect in the password manager scenario was perceived as important by 6 out of 7 people judging this scenario whereas the effect in the consistency checker scenario, was judged as important by 5 out of 10 participants (see Figure 7). Again, having a research background in USP did not correspond to considering the effect in the password manager scenario to be more important. We hypothesize that participants with misconceptions focused more on context than on numbers presented in the scenario than those with more understanding, since this can be judged without statistical knowledge.

At the same time, participants in the surveys and the interviews mentioned context influencing their judgment. One interviewee in S-CHI-i put it as: "Comparing an elephant to the earth, it's small, but compared to a mouse it is big". Participants, especially in S-CHI mentioned medicine and critical infrastructure as areas where effects would be considered more important: "[Importance] depends on the topic and the population. Using a password manager is not the same as curing cancer." (from S-CHI-s). Comparing the importance thresholds between scenarios in Figure 6, the distribution for the password manager scenario for OR is somewhat wider for lower values, but also has more single higher values than the consistency checker scenario.

5.3.3 Consequences. We asked participants to assess the outcome of the study described in the vignettes to gauge the relationship between the consequences of the research (criticality) and judgment of importance. It is unclear, in that effects may nevertheless be judged as important even where outcomes were not judged as severe or beneficial. However, when effects were judged as unimportant, the outcomes were not judged as very severe or beneficial.

5.3.4 Point of view. "Point of view" was only brought up by participants in S-CHI-i. Participants considered who was affected by an effect. In most cases, the participants first viewed the vignette from their own point of view. This was especially apparent in the browser warning scenario, where they judged the probability of a man-in-the-middle-attack to be low for themselves. Sometimes participants then went on to consider groups more at risk from the consequences of such an attack. A point of view that we had not anticipated was that of companies trying to sell software. When participants took on this point of view, the type of software (password manager vs. consistency checker) did not matter much since, for companies, selling their product is the priority.

5.3.5 Other numbers. Artifacts from our vignette, i.e. other numerical values than the effect size, were a prominent influencing factor in both S-CHI and S-SOUPS, but especially in S-SOUPS-i. Participants commonly used p-values to judge importance, equating statistical significance with practical significance. Related work [25] and P-E also discussed this and P-E hypothesized that this is due to "often needing something significant to be able to publish", which is commonly discussed as publication bias [4, 74]. Furthermore, participants used reference values stemming from their own experience or from a source they could not name at the time of the interview to distinguish between important and unimportant effects. One participant in S-SOUPS-i described this as "intuition". For the browser warning scenario with OR, participants hypothesized about the base rate of true positive warnings to judge the effect of improving warning adherence. A higher base rate means that the consequences of not adhering to warnings affect more people, making effects improving adherence more important. In our vignette, we did not provide a base rate since we wanted to see whether participants would bring it up out of their own accord. Participants used the number of participants to judge the validity of the findings and the quality of the study. Finally, some participants did not specify which numbers from the vignette they used.

5.4 Comparison of HCI and USP researchers

We did not find clear differences between participants doing research in USP and HCI. Participants in S-CHI-i discussed a wider range of factors influencing their judgments than participants in S-SOUPS-i, but this was independent of their research focus. Participants in S-SOUPS-i more often referred to numbers in our vignette for their interpretation, but this may also be a side effect of the interview protocol in S-SOUPS-i, where we asked more questions about the value of the effect size. In the survey, we did not identify any trends or differences between the two groups of researchers.

6 LIMITATIONS

First, due to our recruitment strategy, our sample is not representative of HCI or USP researchers. Approachability and availability played a role in recruitment. We also got recommendations from participants on who else might be interested in taking part in the

study, introducing snowball sampling. While this has some drawbacks it is useful to recruit hard to reach participants [16]. Since the interviews were done in person at the venue, we did not get to interview researchers who participated online. However, online participants were able to participate in the surveys. We tried to balance these limitations by sampling purposefully and successfully recruited individuals working on different research topics and in different stages of their careers. While we had planned to conduct individual interviews, we had to adjust this in two cases when another person joined the location we were at and expressed interest in our study. We were initially unsure whether two joint participants would influence each other's answers, however, their perspective on the scenarios differed and we gained interesting insights through their discussion.

The sample size in our surveys was not sufficient for inferential statistics or to make generalizable statements, so we used the findings from the survey to add context to our interview findings. We only used three vignettes and two effect size measures that cannot cover the full scope of USP and HCI research, especially considering that researchers may be more familiar with different effect sizes than the ones investigated in this work. Finally we are influenced by our biases as USP researchers. Nevertheless, we gained insights on factors influencing participants' judgments in different scenarios.

7 DISCUSSION

In our work, we found that many of our participants misunderstood aspects of unitless effect sizes and statistical reporting in our vignettes. Empirical work on effect size interpretation is rare compared to p-values, where misconceptions are well-documented [8, 28, 80, 82]. In psychology, a pre-print by Schäfer reports on a survey where size judgments were mostly made based on Cohen's conventions, and only about half of the participants requested further information apart from standardized effect sizes [72]. O'Keefe documents misunderstandings in the interventional research literature, where effect sizes are interpreted as the size of effect for a single invention, when they actually result from a comparison of two conditions [56]. Critique of the status quo of effect size reporting and interpretation is common across fields however, e.g. in exercise science [19], psychology [24] and education [6, 77], with recommendations provided that support our findings of influencing factors and existing misconceptions.

We identified a number of factors relevant when judging whether an effect is important, such as the context and researchers' point of view. Thus our main take-away and recommendation is that authors reporting statistics add a discussion of effect sizes and their interpretation of size and importance to help guide readers. Without this there is a high risk of readers misjudging the results. As the authors, they are the most knowledgeable about their study, data and analysis and are best positioned to judge the practical relevance of their results. However, our study suggests that many researchers are currently not familiar with or uncertain about effect sizes, most obviously for the two we investigated in our study. However, standardized effect sizes are not the only way to judge the relevance of results. Unstandardized effect sizes in the units of the studies may be more meaningful and easier to connect to concrete outcomes. Further options for interpretation include those we identified in our study, e.g. considering context or who is affected by the results. Some other possibilities for communicating about effect size include using visualizations for the main hypotheses, although these are not without pitfalls [33], or using effect sizes which are easier to interpret [11, 29]. The binomial effect size display (BESD) [68] and the common language effect size (CLES) [50] have been developed as alternatives that are easier to interpret, especially for people untrained in statistics. CLES can also be described as the probability of superiority. Brooks et al. investigated this by comparing BESD and CLES to the Pearson correlation coefficient r and the coefficient of determination R^2, derived from it [11]. They found that BESD and CLES were indeed perceived as easier to understand, but also as larger than r or R^2, which may be useful for cases where effect sizes are at risk to be misperceived as being smaller than they are in the general population but researchers and practitioners are aware of the clear benefit of a treatment and want to communicate this [11]. Where effects are at risk to be overvalued, such as in research communities generally suffering from publication bias [35], this may be less desirable. Hanel and Mehler conducted a survey to compare subjective informativeness of significance statements without effect sizes to Cohen's U3, the probability of superiority, the overlapping coefficient, Bayes Factors and Cohen's d [29]. Cohen's U3 was rated most informative, but surprisingly still less informative than significance statements without effect size information [29]. The authors attributed this finding to an exposure effect, given that participants were likely more familiar with mere significance statements rather than effect sizes [29]. Considering that the applicability of effect sizes also depends on the analyses conducted, determining which effect sizes should be used to make research results understandable is still an open problem.

Another key finding is that we saw researchers interpret the same effect size numbers presented in a vignette differently. This means that "small" or "large" effects may mean different things to different people, both in different contexts but even within the same context. The size of effects considered important also varies between effect sizes as well as within and between contexts. The variation we found can be due to misinterpretation of statistical values but also legitimate different views. We think it is important to minimize the chance of misinterpretations and support different views. A small effect size can still be important depending on how many people are affected and in what way. It is also possible for a large effect to not be particularly important for the same reasons. Thus while generic guidelines for judging effect size such as those from Cohen [15] can be helpful as an initial frame of reference, our results suggest that there is not one unified scale for USP or HCI.

Considering all this in combination with the fact that currently in USP and HCI it is not common for effect sizes to be reported consistently, it becomes even more likely that results are hard to interpret correctly. However, there are a couple of fairly simple modifications on how results are reported that authors can make which we believe would make it easier for readers to interpret their results.

8 RECOMMENDATIONS

Use reporting guidelines. In fields such as psychology and medicine, multiple guidelines exist on how to report statistical results [2, 3, 75].

Using guidelines, such as CONSORT [75], has been shown to improve reporting of study procedures [60]. While some aspects of these guidelines, e.g. statistical genetics [2], do not apply to USP or HCI, the sections on reporting outcomes offer a good foundation we would like to build upon. The guidelines are in agreement that "the minimally sufficient set of statistics [...] needed to construct the tests" [3] (p. 81) should be reported, although exactly what this data encompasses is different across guidelines, and test-specific [2, 3]. Both discuss effect sizes and variability but differ on how they should be reported [2, 3]. Our summary of what is relevant from these guidelines in the context of our study is as follows: For hypothesis tests, researchers should report - where applicable - the exact sub-group sample size per test, an effect size, statistics necessary to be able to recalculate the test, e.g. test statistics, degrees of freedom etc., and exact p-values or confidence intervals. Based on our study, we propose to extend these guidelines as follows:

Explain effect size measures. We picked OR and then Cohen's d because we believed them to be common and fairly simple, however, as we saw, not all our participants were familiar with them. Thus, providing a short explanation or a reference to used effect sizes and their characteristics in a methods or data-analysis section can make the results more accessible to researchers not familiar with them (RQ2).

*Report both unitless **and** simple effect-sizes.* The reporting guidelines we studied mostly focused on unitless/standardized effects sizes. However, our studies showed that researchers had trouble interpreting these, particularly if they were unfamiliar with the measure. Thus we believe that simple effect-sizes should also always be reported, to aid interpretation. Both forms of effect-size have strengths and weaknesses, so offering both seems beneficial to us.

Interpret findings. While reporting effect sizes is an important first step, we think it is not enough. We recommend that researchers qualify their findings with a short qualitative interpretation. As we saw in our studies, what is considered important, large or small can vary. Since the authors are the subject matter experts, we believe their assessment is very valuable for the reader (RQ1). [2] We also explicitly recommend not to rely on conventional scales such as those by Cohen [15], since our results suggest that there is too much context-based variation.

Interpret Power. As we saw, even small effects can be considered important, but many studies are under-powered [58]. There can be legitimate reasons for this, but it is important to help readers understand if important effects might be detected with a larger study.

There are two additional recommendations we want to put up for discussion which might be more controversial:

Move Test Statistics into Supplemental Material. Test statistics (e.g. t, u, degrees of freedom) were not explicitly used by our interview participants when interpreting effects size. However, they can make results sections with fully reported hypothesis tests cumbersome to read, while adding little to improve interpretability. In our view their main benefit is that they make statistical analyses verifiable, enabling e.g. reviewers to check statistics for coherency [54] and making it a little harder for bad actors to create fake data. To find a middle ground between completeness of reporting and readability, these values could be reported in supplemental material, instead of in the main body of the paper, if they are not explicitly used for interpretation. Ideally this would be done in a machine-readable format.

Consider not reporting p values. There is an on-going debate on the value of reporting p-values, see e.g. Wasserstein et al. [81] for a perspective criticizing p-values and Murtaugh [51] for a proponents' view on p-values. While our study did not look at the effect of p values, since this has already been done in related work, we nonetheless observed some interaction. Despite the fact that we held p values constant in an attempt to not confound our study of effect sizes, participants mentioned using p values in their judgment. We think reporting confidence intervals instead of p values has two main benefits. It avoids biasing the effect size judgment and at the same time highlights the uncertainty/range of possible effect sizes. So we would recommend replacing p values with confidence intervals. The meaning of the interval can and should be contained in the authors interpretation.

9 CONCLUSION

We conducted survey and interview studies, and an expert interview to explore HCI and USP researchers' perception of effect size, specifically Cohen's d and odds ratio. We gathered their views on what they consider small, medium and large effects and what they consider important by using different vignettes. We found that judgments of size and importance varied between researchers, and identified additional factors contributing to these judgments. To improve research practice we suggest connecting statistical results and researchers' domain expertise to interpret research results with a focus on practical importance.

ACKNOWLEDGMENTS

The authors thank Theo Raimbault for his aid in literature research, Laura Abels for her help with visualizing the study process and the researchers in the SOUPS and CHI communities for their time and willingness to participate in this research.

REFERENCES

[1] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. 2016. You Get Where You're Looking for: The Impact of Information Sources on Code Security. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Jose, CA, 289–305. https://doi.org/10.1109/SP.2016.25

[2] Andrew D. Althouse, Jennifer E. Below, Brian L. Claggett, Nancy J. Cox, James A. de Lemos, Rahul C. Deo, Sue Duval, Rory Hachamovitch, Sanjay Kaul, Scott W. Keith, Eric Secemsky, Armando Teixeira-Pinto, Veronique L. Roger, and null null. 2021. Recommendations for Statistical Reporting in Cardiovascular Medicine: A Special Report From the American Heart Association. *Circulation* 144, 4 (July 2021), e70–e91. https://doi.org/10.1161/CIRCULATIONAHA.121.055393

[3] American Psychological Association (Ed.). 2020. *Publication Manual of the American Psychological Association* (seventh edition ed.). American Psychological Association, Washington, DC.

[4] Donald R. Atkinson, Michael J. Furlong, and Bruce E. Wampold. 1982. Statistical Significance, Reviewer Evaluations, and the Scientific Process: Is There a (Statistically) Significant Relationship? *Journal of Counseling Psychology* 29, 2 (1982), 189–194. https://doi.org/10.1037/0022-0167.29.2.189

[2] For tests, where a priori power analysis was conducted, authors should reference decisions about the practical relevance of effect size made in advance, when interpreting the findings.

[5] Thom Baguley. 2009. Standardized or Simple Effect Size: What Should Be Reported? *British Journal of Psychology* 100, 3 (2009), 603–617. https://doi.org/10.1348/000712608X377117

[6] Arthur Bakker, Jinfa Cai, Lyn English, Gabriele Kaiser, Vilma Mesa, and Wim Van Dooren. 2019. Beyond Small, Medium, or Large: Points of Consideration When Interpreting Effect Sizes. *Educational Studies in Mathematics* 102 (2019), 1–8. https://doi.org/10.1007/s10649-019-09908-4

[7] Titus Barik, Justin Smith, Kevin Lubick, Elisabeth Holmes, Jing Feng, Emerson Murphy-Hill, and Chris Parnin. 2017. Do Developers Read Compiler Error Messages?. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE) (ICSE'17)*. IEEE, Buenos Aires, Argentina, 575–585. https://doi.org/10.1109/ICSE.2017.59

[8] Lonni Besançon and Pierre Dragicevic. 2019. The Continued Prevalence of Dichotomous Inferences at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–11. https://doi.org/10.1145/3290607.3310432

[9] Nele Borgert, Oliver D. Reithmaier, Luisa Jansen, Larina Hillemann, Ian Hussey, and Malte Elson. 2023. Home Is Where the Smart Is: Development and Validation of the Cybersecurity Self-Efficacy in Smart Homes (CySESH) Scale. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Chi '23)*. Association for Computing Machinery, New York, NY, USA, Article 507, 15 pages. https://doi.org/10.1145/3544548.3580860

[10] Frank A Bosco, Herman Aguinis, Kulraj Singh, James G Field, and Charles A Pierce. 2015. Correlational Effect Size Benchmarks. *Journal of Applied Psychology* 100, 2 (2015), 431.

[11] Margaret E. Brooks, Dev K. Dalal, and Kevin P. Nolan. 2014. Are Common Language Effect Sizes Easier to Understand than Traditional Effect Sizes? *Journal of Applied Psychology* 99, 2 (2014), 332–340. https://doi.org/10.1037/a0034745

[12] Paul Cairns. 2007. HCI... Not As It Should Be: Inferential Statistics in HCI Research. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK*. British Computer Society, Lancaster, UK, 7 pages. https://doi.org/10.14236/ewic/HCI2007.20

[13] Ana Elisa Castro Sotos, Stijn Vanhoof, Wim Van den Noortgate, and Patrick Onghena. 2007. Students' Misconceptions of Statistical Inference: A Review of the Empirical Evidence from Research on Statistics Education. *Educational Research Review* 2, 2 (2007), 98–113.

[14] Kathy Charmaz. 2014. *Constructing Grounded Theory* (2 ed.). SAGE Publications Ltd, London, UK.

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

[16] Nissim Cohen and Tamar Arieli. 2011. Field Research in Conflict Environments: Methodological Challenges and Snowball Sampling. *Journal of Peace Research* 48, 4 (2011), 423–435. https://doi.org/10.1177/0022343311405698 jstor:23035205

[17] Elizabeth Collins and Roger Watt. 2021. Using and Understanding Power in Psychological Research: A Survey Study. *Collabra: Psychology* 7, 1 (Oct. 2021), 28250.

[18] Noelle M. Crooks, Anna N. Bartel, and Martha W. Alibali. 2019. Conceptual Knowledge of Confidence Intervals in Psychology Undergraduate and Graduate Students. *STATISTICS EDUCATION RESEARCH JOURNAL* 18, 1 (May 2019), 46–62.

[19] Scott J. Dankel, J. Grant Mouser, Kevin T. Mattocks, Brittany R. Counts, Matthew B. Jessee, Samuel L. Buckner, Paul D. Loprinzi, and Jeremy P. Loenneke. 2017. The Widespread Misuse of Effect Sizes. *Journal of Science and Medicine in Sport* 20, 5 (May 2017), 446–450. https://doi.org/10.1016/j.jsams.2016.10.003

[20] Verena Distler, Matthias Fassl, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Lorrie Faith Cranor, and Vincent Koenig. 2021. A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research. *ACM Transactions on Computer-Human Interaction* 28, 6 (Dec. 2021), 43:1–43:50. https://doi.org/10.1145/3469845

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

[22] Christopher J Ferguson. 2009. An Effect Size Primer: A Guide for Clinicians and Researchers. *Professonal Psychology: Research and Practice* 40, 5 (2009), 532–538. https://doi.org/10.1037/a0015808

[23] Nollaig Frost, Sevasti Melissa Nolas, Belinda Brooks-Gordon, Cigdem Esin, Amanda Holt, Leila Mehdizadeh, and Pnina Shinebourne. 2010. Pluralism in Qualitative Research: The Impact of Different Researchers and Qualitative Approaches on the Analysis of Qualitative Data. *Qualitative Research* 10, 4 (2010), 441–460. https://doi.org/10.1177/1468794110366802

[24] David C. Funder and Daniel J. Ozer. 2019. Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science* 2, 2 (June 2019), 156–168. https://doi.org/10.1177/2515245919847202

[25] Steven Goodman. 2008. A Dirty Dozen: Twelve P-value Misconceptions. *Seminars in Hematology* 45, 3 (2008), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

[26] Thomas Groß. 2021. Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies. In *Socio-Technical Aspects in Security and Trust (Lecture Notes in Computer Science)*, Thomas Groß and Theo Tryfonas (Eds.). Springer International Publishing, Cham, 3–26. https://doi.org/10.1007/978-3-030-55958-8_1

[27] C. Keith Haddock, David Rindskopf, and William R. Shadish. 1998. Using Odds Ratios as Effect Sizes for Meta-Analysis of Dichotomous Data: A Primer on Methods and Issues. *Psychological Methods* 3, 3 (1998), 339–353. https://doi.org/10.1037/1082-989X.3.3.339

[28] Heiko Haller and Stefan Krauss. 2002. Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research Online* 7, 1 (2002), 20 pages.

[29] Paul HP Hanel and David MA Mehler. 2019. Beyond Reporting Statistical Significance: Identifying Informative Effect Sizes to Improve Scientific Communication. *Public Understanding of Science* 28, 4 (2019), 468–485. https://doi.org/10.1177/0963662519834193

[30] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. 2014. Using Personal Examples to Improve Risk Communication for Security & Privacy Decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Chi '14)*. Association for Computing Machinery, New York, NY, USA, 2647–2656. https://doi.org/10.1145/2556288.2556978

[31] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The Weirdest People in the World? *Behavioral and Brain Sciences* 33, 2-3 (June 2010), 61–83. https://doi.org/10.1017/S0140525X0999152X

[32] Rink Hoekstra, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. 2014. Robust Misinterpretation of Confidence Intervals. *Psychonomic Bulletin & Review* 21, 5 (Oct. 2014), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

[33] Jake M. Hofman, Daniel G. Goldstein, and Jessica Hullman. 2020. How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376454

[34] William L Holcomb, Tinnakorn Chaiworapongsa, Douglas A Luke, and Kevin D Burgdorf. 2001. An Odd Measure of Risk: Use and Misuse of the Odds Ratio. *Obstetrics & Gynecology* 98, 4 (Oct. 2001), 685–688. https://doi.org/10.1016/S0029-7844(01)01488-0

[35] John P. A. Ioannidis. 2008. Why Most Discovered True Associations Are Inflated. *Epidemiology* 19, 5 (2008), 640–648. jstor:25662607

[36] John P. A. Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N. Goodman. 2015. Meta-Research: Evaluation and Improvement of Research Methods and Practices. *PLOS Biology* 13, 10 (Oct. 2015), e1002264. https://doi.org/10.1371/journal.pbio.1002264

[37] Alex Kale, Matthew Kay, and Jessica Hullman. 2021. Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 272–282.

[38] Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag I.K. Sjøberg. 2007. A Systematic Review of Effect Size in Software Engineering Experiments. *Information and Software Technology* 49, 11 (2007), 1073–1086. https://doi.org/10.1016/j.infsof.2007.02.015

[39] Fumihiro Kanei, Ayako A. Hasegawa, Eitaro Shioji, and Mitsuaki Akiyama. 2023. Analyzing the Use of Public and In-House Secure Development Guidelines in U.S. and Japanese Industries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Chi '23)*. Association for Computing Machinery, New York, NY, USA, Article 187, 17 pages. https://doi.org/10.1145/3544548.3580705

[40] Maurits Kaptein and Judy Robertson. 2012. Rethinking Statistical Analysis Methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1105–1114. https://doi.org/10.1145/2207676.2208557

[41] Ken Kelley and Kristopher J. Preacher. 2012. On Effect Size. *Psychological Methods* 17, 2 (2012), 137–152. https://doi.org/10.1037/a0028086

[42] Riko Kelter. 2020. Analysis of Bayesian Posterior Significance and Effect Size Indices for the Two-Sample t-Test to Support Reproducible Medical Research. *BMC Medical Research Methodology* 20 (April 2020), 88. https://doi.org/10.1186/s12874-020-00968-2

[43] Yea-Seul Kim, Jake M Hofman, and Daniel G Goldstein. 2022. Putting Scientific Results in Perspective: Improving the Communication of Standardized Effect Sizes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 625, 14 pages. https://doi.org/10.1145/3491102.3502053

[44] Roger E. Kirk. 1996. Practical Significance: A Concept Whose Time Has Come. *Educational and Psychological Measurement* 56, 5 (Oct. 1996), 746–759. https://doi.org/10.1177/0013164496056005002

[45] Roger E. Kirk. 2003. The Importance of Effect Magnitude. In *Handbook of Research Methods in Experimental Psychology*. Blackwell Publishing Ltd., Malden, MA, USA, 83–105.

[46] Daniel Lakens. 2013. Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs. *Frontiers in Psychology* 4 (2013), 12 pages.

[47] J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. jstor:pdf/2529310.pdf

[48] Victor Le Pochat and Wouter Joosen. 2023. Analyzing Cyber Security Research Practices through a Meta-Research Framework. In *Proceedings of the 16th Cyber Security Experimentation and Test Workshop (CSET '23)*. Association for Computing Machinery, New York, NY, USA, 64–74. https://doi.org/10.1145/3607505.3607523

[49] Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. 2019. Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology* 10 (Dec. 2019), 2767. https://doi.org/10.3389/fpsyg.2019.02767

[50] Kenneth O. McGraw and S. P. Wong. 1992. A Common Language Effect Size Statistic. *Psychological Bulletin* 111, 2 (1992), 361–365. https://doi.org/10.1037/0033-2909.111.2.361

[51] Paul A. Murtaugh. 2014. In Defense of P Values. *Ecology* 95, 3 (2014), 611–617. https://doi.org/10.1890/13-0590.1

[52] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. 2017. 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)*. Association for Computing Machinery, New York, NY, USA, 311–328. https://doi.org/10.1145/3133956.3134082

[53] Jakob Nielsen and Jonathan Levy. 1994. Measuring Usability: Preference vs. Performance. *Commun. ACM* 37, 4 (April 1994), 66–75. https://doi.org/10.1145/175276.175282

[54] Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. 2016. The Prevalence of Statistical Reporting Errors in Psychology (1985–2013). *Behavior Research Methods* 48, 4 (Dec. 2016), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

[55] Natalia Obukhova. 2021. 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)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3411763.3451520

[56] Daniel J O'Keefe. 2017. Misunderstandings of Effect Sizes in Message Effects Research. *Communication Methods and Measures* 11, 3 (2017), 210–219.

[57] Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. 2023. 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)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3580766

[58] Anna-Marie Ortloff, Christian Tiefenau, and Matthew Smith. 2023. 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. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. USENIX Association, Anaheim, CA, 341–359. https://www.usenix.org/conference/soups2023/presentation/ortloff

[59] Alexander Persoskie and Rebecca A. Ferrer. 2017. A Most Odd Ratio:: Interpreting and Describing Odds Ratios. *American Journal of Preventive Medicine* 52, 2 (Feb. 2017), 224–228. https://doi.org/10.1016/j.amepre.2016.07.030

[60] Amy C Plint, David Moher, Andra Morrison, Kenneth Schulz, Douglas G Altman, Catherine Hill, and Isabelle Gaboury. 2006. Does the CONSORT Checklist Improve the Quality of Reports of Randomised Controlled Trials? A Systematic Review. *Medical Journal of Australia* 185, 5 (2006), 263–267. https://doi.org/10.5694/j.1326-5377.2006.tb00557.x

[61] Luke Plonsky and Frederick L Oswald. 2014. How Big Is "Big"? Interpreting Effect Sizes in L2 Research. *Language learning* 64, 4 (2014), 878–912.

[62] Leo Poom and Anders af Wåhlberg. 2022. Accuracy of Conversion Formula for Effect Sizes: A Monte Carlo Simulation. *Research Synthesis Methods* 13, 4 (2022), 508–519. https://doi.org/10.1002/jrsm.1560

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

[64] Stuart Reeves. 2015. Human-Computer Interaction as Science. In *Proceedings of the Fifth Decennial Aarhus Conference on Critical Alternatives (Ca '15)*. Aarhus University Press, Aarhus N, 73–84. https://doi.org/10.7146/aahcc.v1i1.21296

[65] Katja Rogers and Katie Seaborn. 2023. The Systematic Review-lution: A Manifesto to Promote Rigour and Inclusivity in Research Synthesis. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 11. https://doi.org/10.1145/3544549.3582733

[66] James A. Rosenthal. 1996. Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research* 21, 4 (Oct. 1996), 37–59. https://doi.org/10.1300/J079v21n04_02

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

[68] Robert Rosenthal and Donald B Rubin. 1982. A Simple, General Purpose Display of Magnitude of Experimental Effect. *Journal of Educational Psychology* 74, 2 (1982), 166–169. https://doi.org/10.1037/0022-0663.74.2.166

[69] Ralph L Rosnow and Robert Rosenthal. 1989. Statistical Procedures and the Justification of Knowledge in Psychological Science. *American Psychologist* 44, 10 (1989), 1276–1284.

[70] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanotham, and Mauro Cherubini. 2023. Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3544548.3580848

[71] Shruti Sannon and Andrea Forte. 2022. Privacy Research with Marginalized Groups: What We Know, What's Needed, and What's Next. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 455:1–455:33. https://doi.org/10.1145/3555556

[72] Thomas Schäfer. 2023. On the Use and Misuse of Standardized Effect Sizes in Psychological Research. https://doi.org/10.31219/osf.io/x8n3h

[73] Thomas Schäfer and Marcus A. Schwarz. 2019. The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology* 10, Article 813 (2019), 13 pages.

[74] Anne M. Scheel, Mitchell R. M. J. Schijen, and Daniël Lakens. 2021. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science* 4, 2 (April 2021), 12 pages. https://doi.org/10.1177/25152459211007467

[75] Kenneth F. Schulz, Douglas G. Altman, and David Moher. 2010. CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials. *Journal of Pharmacology and Pharmacotherapeutics* 1, 2 (Dec. 2010), 100–107. https://doi.org/10.4103/0976-500X.72352

[76] James P. Shaver. 1993. What Statistical Significance Testing Is, and What It Is Not. *The Journal of Experimental Education* 61, 4 (July 1993), 293–316. https://doi.org/10.1080/00220973.1993.10806592

[77] Adrian Simpson. 2020. On the Misinterpretation of Effect Size. *Educational Studies in Mathematics* 103 (2020), 125–133. https://doi.org/10.1007/s10649-019-09924-4

[78] Slack. n.d.. ACM SIGIR Slack. https://acmsigir.slack.com/.

[79] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. 2009. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *18th USENIX Security Symposium (USENIX Security 09)*. USENIX Association, Montreal, Canada, 399–432.

[80] Chun Wah Michael Tam, Abeer Hasan Khan, Andrew Knight, Joel Rhee, Karen Price, and Katrina McLean. 2018. How Doctors Conceptualise P Values: A Mixed Methods Study. *Australian Journal of General Practice* 47, 10 (Oct. 2018), 705–710.

[81] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. 2019. Moving to a World Beyond " $p < 0.05$". *The American Statistician* 73, sup1 (March 2019), 1–19. https://doi.org/10.1080/00031305.2019.1583913

[82] Donna M Windish, Stephen J Huot, and Michael L Green. 2007. Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature. *JAMA* 298, 9 (2007), 1010–1022.