

Investigating Research Practice and Methods in Usable Security and Privacy

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Anna-Marie Ortloff

aus
Freising

Bonn, August 2025

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter/Betreuer: Prof. Dr. Matthew Smith
Gutachter: Prof. Dr. Michael Meier

Tag der Promotion: 22.12.2025
Erscheinungsjahr: 2026

Abstract

Usable Security and Privacy (USP) is a fairly young discipline, with unique challenges in conducting user studies because its research is situated at the intersection of Human Computer Interaction (HCI) and IT security. This makes meta-scientific work in this area especially beneficial. There have been individual examples of meta-scientific work in USP for some time but only recently has meta-science explicitly started being included in calls for papers. This work contributes to meta-science in USP in three areas: Qualitative research, quantitative research, and replicability. For the first two areas, it also considers the broader area of HCI research.

For qualitative research, this thesis explores the influence of different qualitative coders and data with differing degrees of complexity on the results of the qualitative analysis process and expectations in the review process, surrounding the number of coders used. The work describes a series of studies with two types of data and coders with different amounts of experience, as well as a survey with reviewers. Results were more similar for simple data than complex, and the results of more experienced researchers were more abstract and focused on the research questions than those of students with less experience. Reviewers had varied opinions on how many coders were appropriate and many expected clear justifications of methods rather than following set rules.

This work investigates two quantitative topics: The use of power analysis to determine necessary sample size prior to conducting studies, and the reporting, understanding and interpretation of effect sizes. Both topics are situated in the null hypothesis significance testing framework of statistical analysis. A literature analysis in a specialized field (Developer-Centered Usable Security) showed that power analysis is rarely conducted and often, the data necessary to conduct a power analysis based on literature is not reported there. Extending this analysis to HCI consolidates issues with reporting. The resulting data sets can be used to inform future power analyses.

Regarding effect sizes, the analysis of HCI literature suggests that the size of effects varies between fields. The thesis uses a subset of this sample to derive approaches to interpret effect sizes. Further, interview and survey studies with researchers in USP and HCI were used to identify influence factors on researchers' size and importance judgments of effect sizes. The results from these studies also suggest that there are misconceptions around the investigated effect sizes.

On the topic of replicability, this work discusses developments in recruitment and analysis but also in the area of study, using the third published iteration of a study on experts' and non-experts' security-related behaviors and advice as an example.

From the results from these three areas of studies, the thesis draws recommendations on how methods and reporting in USP can be improved.

Zusammenfassung

Usable Security and Privacy (USP, dt. benutzbare IT Sicherheit und Privatsphäre) ist ein relativ junges Forschungsfeld, das durch seine Position an der Schnittstelle zwischen der Mensch-Computer-Interaktion (MCI) und der system-orientierteren IT Sicherheit spezifische Herausforderungen in der Durchführung von Nutzerstudien mit sich bringt. Das macht Metaforschung besonders hilfreich für diesen Forschungsbereich. Schon seit einiger Zeit gibt es einzelne Beispiele von metawissenschaftlichen Arbeiten im Forschungsbereich der USP. Metaforschung wird aber erst seit kurzem explizit in Aufrufen zur Einreichung von wissenschaftlichen Arbeiten benannt. Der Beitrag dieser Arbeit zu Metaforschung in der USP teilt sich in drei Bereiche auf: Qualitative Forschung, quantitative Forschung und Replizierbarkeit. Für die ersten beiden Bereiche wird auch das umfassendere Forschungsfeld der MCI betrachtet.

Im Bereich der qualitativen Forschung untersucht diese Arbeit den Einfluss von unterschiedlichen Kodierern und unterschiedlichen Komplexitätsgraden von Daten auf die Ergebnisse des qualitativen Analyseprozesses, sowie Erwartungen im Reviewprozess an die involvierte Anzahl von Kodierern. Die Arbeit beschreibt eine Reihe von Studien mit zwei unterschiedlichen Arten von Daten und Kodierern mit einer unterschiedliche Menge an Erfahrung, sowie eine Umfrage mit Reviewern. Die Ergebnisse des Analyseprozesses waren für einfache Daten ähnlicher als für komplexere Daten und die Ergebnisse der erfahreneren Forscher waren abstrakter und stärker auf die Forschungsfragen fokussiert als die Ergebnisse der studentischen Teilnehmer mit weniger Erfahrung. Reviewer haben unterschiedliche Meinungen dazu geäußert, wie viele Kodierer angemessen sind und viele von ihnen erwarteten Begründungen für die verwendete Methode statt des Befolgens strikter Regeln zur Anzahl der Kodierer.

In Bezug auf quantitative Forschungsmethoden werden in dieser Arbeit zwei Themen erforscht: Die Verwendung von Poweranalyse, um die benötigte Stichprobengröße vor der Durchführung von Studien zu ermitteln, und das Berichten, das Verständnis und die Interpretation von Effektgrößen durch Wissenschaftler. Beide Themen sind im Bereich des Nullhypothesen-Signifikanztesten (NHST) einzuordnen. Eine Analyse der Literatur im spezialisierten Feld der auf Entwickler fokussierten IT Sicherheit zeigt, dass die in Publikationen berichteten Daten oft nicht ausreichen, um darauf basierend eine Poweranalyse durchzuführen und dass diese Praktik auch nicht weit verbreitet ist. Eine Literaturanalyse im weniger spezialisierten Feld der MCI festigt diese Erkenntnisse zu Problemen beim Berichten von statistischen Ergebnissen. Die resultierenden Datensätze können bei der Schätzung von Effektgrößen für zukünftige Poweranalysen verwendet werden.

Zum Thema Effektgrößen deutet die Analyse der MCI-Literatur darauf hin, dass die Größe von Effekten zwischen den unterschiedlichen Forschungsfeldern variiert. In der Dissertation wurde eine

Teilmenge dieser Stichprobe verwendet um verwendete Interpretationsstrategien für Effektgrößen zu identifizieren. Zusätzlich wurden Interview- und Umfragestudien mit Forschenden im Bereich USP oder MCI durchgeführt, um Einflussfaktoren auf deren Größen- und Relevanzeinschätzungen für Effektgrößen zu ermitteln. Die Ergebnisse dieser Studien deuten darauf hin, dass es falsche Annahmen und Missverständnisse in Bezug auf das Verständnis der erforschten Effektgrößen gibt.

Im Bereich Replizierbarkeit werden am Beispiel einer zweiten Replikation einer Studie über IT Sicherheitsverhalten und -empfehlungen einerseits Entwicklungen im Bereich der Teilnehmerrekrutierung sowie der Datenanalyse diskutiert, andererseits aber auch Entwicklungen auf der inhaltlichen Ebene der Studie.

Basierend auf den Ergebnissen dieser drei Bereiche werden in der Dissertation Empfehlungen ausgesprochen, mit denen Forschungsmethoden und Ergebnisberichte in der USP verbessert werden können.

Acknowledgments

A Phd is not a solo journey and I have many people to thank for making my Phd journey what it was. First of all, thank you to Matthew for advising me throughout my time as a doctoral student, for giving me true academic freedom in my choice of research topic and lots of food for thought and ideas during my time in Bonn. Thank you to Prof. Michael Meier, Prof. Lucie Flek, and Prof. Caja Thimm for agreeing to be on my committee.

Much of my research relies on human participants being willing to take part in the studies I conducted. Thank you to everyone who participated in my research, whether it is part of this thesis or not. This thanks is especially due to all my friends and family who participated in multiple studies during my undergraduate degrees. Research is not something you do alone. Thank you to all my collaborators for your ideas, discussions and contributions. The English in this thesis was very much improved through the help of Kristin Vondruska and the German of the “Zusammenfassung” through proof-reading by my mother Christine Ortloff. Any remaining mistakes are mine.

My academic journey started earlier than the work on this thesis however. Thank you to Dr. David Elswiler and Dr. Markus Kattenbeck, who kickstarted my love of methodological topics, introduced me to academia, and let me work on usable privacy topics, despite this not being their area of focus. Thank you to my first co-authors, Denis Feth and Svenja Polst, who took a chance on three enthusiastic undergraduates, and to Lydia Güntner and Maximiliane Windl who were great project partners and co-authors throughout my Bachelors and Masters degrees. I was very happy to be able to start a Phd in the Behavioral Security and Privacy group (then Usable Security and Privacy group) in Bonn. Thank you to Chris for helping me get a foot in the door here. Thank you, Alena and Chris, for welcoming me into the fold academically, setting me up with my first contributions to research projects in Bonn, and to Alena specifically for making me feel like a statistics expert for the first time later in my Phd. Thank you Anne and Stephan for making me feel welcome in Bonn on a personal level during a time when personal contact was hard to come by. Thank you Kerrin for saving the day on the administrative/organizational level multiple times. Thank you to Anne, Florin, Leonard, and Laura who were tutors for the course I taught and made the load of preparation and evaluation much lighter. Thank you first to Eva and later on to Arthi, who were both fantastic office room mates to me, sharing groans and laughs, and for the brunt of the plant-care-taking in our office. Lisa, Florin and Arthi, and later Lukas were sources of levity and community even during times of stress. Thanks for letting me win at Tetris at least once. Thank you to Ashi for jumping in on an early project which has been the one constant resubmission during my Phd, for failing together (and trying again) and becoming friends in the process. Thank you Jenny, for being the first person I met in real life (at SOUPS’23) who matched my enthusiasm for statistics.

Finally, life is not all academics. Thank you to my family, my mom and dad, Christine and Stefan, and my brother Anton, who were a big part of how I became the person I am now, who fostered my development into the person I have become, challenged me to grow from the first baby steps up until now, and who believed in me from the start. To Andi, my wonderful partner, who was there for me throughout, took care that I was still eating and sleeping healthily even in times of stress, who took part in my enthusiasm and who shared with me love, patience and an open ear for all the good and the bad. Thank you!

Contents

Abstract	i
Zusammenfassung	iii
Acknowledgments	v
Contents	ix
1 Introduction	1
2 Background and Related Work	5
2.1 Usable Security and Privacy	6
2.1.1 Developer-Centered Usable Security	7
2.1.2 Security Advice	7
2.2 Meta Science	8
2.2.1 Support for Meta-Scientific Research at Venues Publishing USP Papers	8
2.2.2 Meta-Scientific Research in USP and HCI	11
2.3 Qualitative Analysis	14
2.3.1 Coding Practices in USP	15
2.3.2 Quality Criteria for Qualitative Research	15
2.3.3 Meta-Analyses on Qualitative Analysis	17
2.4 Quantitative Analysis	17
2.4.1 Null-Hypothesis Statistical Testing	18
2.4.2 Effect Sizes	19
2.4.3 Power Analysis	20
2.4.4 Reporting of Statistical Methods and Results	23
2.5 Replications	25
2.6 Positionality statement	26

3	Influence Factors on Results from Qualitative Analysis	29
3.1	Motivation	30
3.2	Empirical Evaluation of the Coding Process in Qualitative Analysis	31
3.2.1	Choosing a Suitable Security Topic as a Basis for our Meta-Study	31
3.2.2	Meta-Studies	33
3.2.3	Interview Study (Complex data)	34
3.2.4	Survey Study (Simple Data)	37
3.2.5	Participants	38
3.2.6	Ethics	40
3.3	Results	40
3.3.1	The Influence of Different Data Types and Researcher Experience on the Analysis Outcomes	40
3.3.2	The Influence of Researcher Experience on Mindmap Structure in Complex Data Analysis	45
3.3.3	The Influence of Multiple Coders on the Qualitative Analysis Process for Complex Data	47
3.3.4	The Influence of Different Data Types and Researcher Experience on Calculated Inter-rater Reliability	49
3.3.5	Summary: RQ-Qual-1 and RQ-Qual-2	51
3.4	Survey of SOUPS program committee members	52
3.4.1	Recruitment and Participants	52
3.4.2	Ethics	52
3.4.3	Data Analysis	53
3.4.4	Important Quality Criteria	53
3.4.5	Acceptable Coding Practices	54
3.4.6	Acceptable Levels of Calculated Inter-Coder Agreement	55
3.5	Limitations	56
3.6	Discussion	58
3.7	Conclusion and Future Work	59

4	Power Analysis and Effect Sizes in Developer-Centered Usable Security	61
4.1	Motivation	61
4.2	Literature Collection	62
4.3	Systematizing Study and Statistical Test Data	63
4.3.1	Systematization Process	63
4.3.2	Data Structure	64
4.3.3	Data Input and Checking	67
4.4	Meta-Analysis	67
4.4.1	Variable Topics	68
4.4.2	Use of Hypothesis Tests	68
4.4.3	Completeness of Statistical Reporting	69
4.5	Systematic A Priori Power Analysis	71
4.5.1	Effectsize Meta-Analysis	73
4.5.2	Publication Bias Correction	74
4.6	Limitations	75
4.7	Conclusion	76
4.8	Future Work	76
5	Power and Effect Size at CHI	79
5.1	Motivation	79
5.2	Extracting Statistics	81
5.2.1	Sample of CHI Publications	82
5.2.2	Statistics of Interest	82
5.2.3	Extraction Tooling	83
5.2.4	Prompt Engineering	83
5.2.5	Evaluation	84
5.3	Constructing Effect Size Guidelines for CHI Research Areas	85
5.3.1	Categorization of Papers	85
5.3.2	Converting Effect Sizes	86
5.3.3	Data Cleaning	86
5.3.4	Data Analysis	87

5.3.5	Limitations	88
5.3.6	Results	90
5.4	Application of Findings	94
5.5	Conclusion	97
6	Interpretation of Effect Sizes	99
6.1	Motivation	100
6.2	Effect Size Interpretation in Published Work	101
6.2.1	Sample Selection	102
6.2.2	Data Analysis	102
6.2.3	Limitations	103
6.2.4	Types of Papers	103
6.2.5	Overview of Effect Size Reporting	104
6.2.6	Interpretation Approaches	104
6.3	Studies to Investigate How Researchers Interpret and Understand Effect Sizes	107
6.3.1	Positionality	108
6.3.2	Ethics	109
6.3.3	Measurement	109
6.3.4	Vignettes	110
6.3.5	Conventional Effect Size	111
6.3.6	Surveys and Interviews	111
6.3.7	Recruitment	112
6.3.8	Participants	112
6.3.9	Data Analysis	112
6.3.10	Changes in S-SOUPS	113
6.3.11	Recruitment and Participants in S-SOUPS	114
6.3.12	Expert Interview	114
6.3.13	Results	115
6.3.14	Limitations	123
6.4	Discussion	123
6.5	Conclusion	125

7	A Re-replication Study About Experts' and Non-experts' Security Practices and Advice	127
7.1	Motivation	128
7.2	Replicated Studies	128
7.3	Method	129
7.3.1	Positionality	129
7.3.2	Ethics	130
7.3.3	Study Design	130
7.3.4	Recruitment	132
7.3.5	Participants	132
7.3.6	Limitations	134
7.3.7	Data Analysis	136
7.4	Content-level Results	137
7.4.1	Development of Security Behavior over Time (RQ-Rep-1)	137
7.4.2	Differences between Experts and Non-Experts (RQ-Rep-2)	138
7.4.3	Experts' View on the Current State of End User Security Advice Practices (RQ-Rep-3)	144
7.4.4	Discussion	146
7.5	Meta-scientific Results	148
7.5.1	Differences in Study Design	148
7.5.2	Development of Participant Recruitment	149
7.5.3	Differences in Statistical Analysis	153
7.6	Conclusion	158
8	Recommendations	159
8.1	Recommendations for Qualitative Research Practice	160
8.2	Recommendations for Quantitative Research Practice	161
9	Conclusion	165
	List of References	167

A Background	221
A.1 USP Papers Evaluated for Effect Size Reporting at CHI'23	221
A.2 USP Papers Evaluated for Effect Size Reporting at SOUPS'23	221
B Influence Factors on Results from Qualitative Analysis	223
B.1 Data Collection Instruments	223
B.1.1 Interview Analysis Task Description	232
B.1.2 Survey Analysis Task Description	236
B.2 Meta-level analysis outcomes	237
B.2.1 Interview Content Codebook	237
B.2.2 Survey Content Codebook	239
B.3 PC Member Survey	242
B.4 PC Member Survey Codebook	244
C Power Analysis and Effect Sizes in Developer-Centered Usable Security	247
C.1 Included Papers	247
D Power and Effect Size at CHI	249
D.1 Prompt	249
D.2 Effect Size Guidelines	251
E Interpretation of Effect Sizes	257
E.1 CHI'23 Publications Used for Investigation of Effect Size Interpretation Strategies	257
E.2 Vignettes	258
E.2.1 Vignette from Participants' Perspective	258
E.2.2 Research Question and Study Design	258
E.2.3 Descriptive Statistics and Inferential Statistics	259
E.2.4 Effect Size Explanation	261
E.3 Interview guidelines	261
E.3.1 S-CHI-i	261
E.3.2 S-SOUPS-i	263
E.4 Surveys	264
E.4.1 S-CHI-s	264
E.4.2 S-SOUPS-s	266

F	A Re-replication Study About Experts' and Non-experts' Security Practices and Advice	269
F.1	Survey Material	269

Chapter 1

Introduction

The number of scientific publications is continually rising [247, 371]. Besides developments on a content-level in various research areas from biology and computer science through psychology and sociology to arts and humanities, research methods and practices also evolve. Investigating these methods is called meta-science. Meta-science can include work on research methods, reporting practice, reproducibility practices, evaluation methods and incentivizing good science [235]. Aspects of meta-science gained more traction in the course of the so-called replication crisis, where large-scale attempts to replicate important psychological findings, failed to replicate to the degree expected [102, 362]. This led to a call for more transparency in description of methods and reporting of results [20, 262, 492, 280]. While there has been no explicit investigation of replicability in Human Computer Interaction (HCI) or Usable Security and Privacy (USP), practices which have been identified as problematic are also present in HCI [541, 442] and USP [96, 189].

Research in USP specifically and HCI more generally has traditionally adopted research methods from different domains and research areas. Meta-science provides an avenue to investigate how these adopted methods are applied in practice and an introspection into how we as a field want to continue to do research. In USP, some types of meta-scientific research have been explicitly welcomed at venues like the Symposium on Usable Privacy and Security (SOUPS), the IEEE Symposium on Security and Privacy (S&P) or USENIX Security for some time, e.g. as Systematization of Knowledge (SoK) papers or replications [138, 528, 100, 529] and an explicit call for meta-scientific research has recently been introduced at USENIX Security in 2025 [530]. The flagship conference in HCI, the ACM Conference on Human Factors in Computing Systems (CHI) includes replications as a possible contribution [52] but meta-science is not explicitly mentioned and choosing two required subcommittees is hard for meta-scientific contributions, as the available subcommittees very focused on specific content-level topics, rather than methodology. Thus, additional meta-scientific work in USP and HCI would be valuable to help these fields assess and improve their research practices.

This dissertation includes contributions to three areas of meta-science: research on research methods, reporting practice and reproducibility practice, applied in USP. I explore a range of methods used in both qualitative and quantitative USP research and derive recommendations to improve the workflow, reporting and validity of the results we produce.

First, for **qualitative research**, we explored issues around multiple coders during the qualitative data analysis process. Recommendations on whether to conduct qualitative data analysis with one

or multiple coders are not unified [73, 327]. Prior work has also explored the influence of research characteristics on qualitative analysis outcomes [163, 273, 88]. When exploring qualitative research practice in USP, we focused on the impact of data complexity and researcher experience on the similarity of coding output. We found that the coding results for more simple data were more similar both within a group of coders and comparing different groups of coders and that more research experience led to a more abstract representation of the findings from the coding process. To ground these findings and compare them to reviewers' expectations, we investigated the quality criteria reviewers at SOUPS apply to qualitative research using a survey and found that reviewers differed in the specific quality criteria they apply, but generally expect justification of the specific methods used.

For **quantitative research**, we investigated the use of power analysis and reporting of effect sizes in Developer-Centered Usable Security (DCUS) and HCI. A priori power analysis is useful to determine the sample size necessary to detect relevant and interesting effects but this strategy is often not used [134]. A potential reason is that effect sizes are not comprehensively reported [442, 78, 189], so researchers have a hard time judging which effects are typical in their field of research and choosing which to use in power analysis. We first conducted a manual analysis of publications in DCUS, where recruiting of expert participants is particularly effortful and expensive to give an overview over typical effect sizes in this field, for different types of variables. This work was extended to a larger sample of publications in the broader research area of HCI. We used a LLM-assisted process to gain an overview over effect sizes in CHI publications between 2019 and 2023 and derived research area specific effect size guidelines. Beyond the reporting of effect size numbers we were also interested in how researchers communicate about and interpret effect sizes. We conducted a qualitative analysis of effect size reporting in quantitative CHI papers and identified five categories of approaches researchers used when interpreting effect sizes in writing. We additionally conducted an interview and survey study to explore researchers' understanding and interpretation of effect sizes further. We found that additional factors, such as context or point of view taken by the researchers influence their interpretation of effect sizes beyond the numerical size and that they feel uncertainty when interpreting effect sizes and have misconceptions about the measures we investigated, Cohen's d and odds ratio (OR).

Finally, we conducted a **re-replication** study on experts' and non-experts' security practices and advice. Prior work had shown differences between expert recommendations and behavior and non-expert security practice [74, 236] but also emphasized that there is an information overload of security advice [417]. This makes it hard for users to decide which advice to implement. We found that differences in security practice between experts and non-experts remained in 2024 but for some behavior, like use of two-factor authentication (2FA), experts and non-experts were becoming more similar. Consequently, we can say that in the ten years since the original study was published, security behavior has noticeably changed and this is a time frame to consider for future replications. In conducting the replication study we also encountered methodological issues, concerning recruiting both of experts and non-experts in a time when AI is accessible and available, and regarding statistical reporting in the original study and the first replication.

This dissertation makes the following contributions: Concerning qualitative research,

- I derive recommendations on coding practices regarding the number of involved coders in qualitative research from an analysis of coding outcomes of 65 student coders and seven researcher coders.

- I present recommendations in reporting qualitative analysis methodology, based on a survey of 37 SOUPS PC members and said study of coding outcomes.

Concerning quantitative research,

- I make recommendations on reporting of statistical results for quantitative null hypothesis significance testing (NHST) work, grounded in empirical evaluation of reporting and interpretation of effect sizes
- I provide manually extracted methodological and results-related information for 54 top tier publications in DCUS and LLM-assisted extracted methodological and results-related information for 599 CHI papers published between 2019 and 2023, including 122 papers in the area of security and privacy.
- I outline interpretation strategies for statistical effect sizes, based on a qualitative study with researchers and a literature analysis of 67 CHI publications.
- I provide a step-by-step tutorial on power analysis for hypothesis tests with one categorical independent variable with two groups.

Concerning replication,

- I highlight that ten years is a sufficient time frame in research on security advice to expect evolution of advice and behavior.
- I discuss evolution of study design, recruitment and statistical analysis methods to facilitate future replications.

Concerning security advice,

- I provide updated information on the status quo regarding security behavior and advice of experts and non-experts in 2024.

The structure of this thesis is as following:

Chapter 1 motivates the investigation of research methods in USP and describes the contributions of this work, as well as the structure of the thesis.

Chapter 2 first provides background information on USP, with specific focus on DCUS and security advice. It then introduces meta science in general and specifically in USP and HCI, and the research practices I investigate in the studies presented in this thesis, i.e. qualitative analysis and NHST-related quantitative methods, including power analysis, effect sizes and their use in USP and HCI.

Chapter 3 focuses on research methods and reporting in qualitative research. I describe an empirical evaluation of the qualitative coding process using two content-level studies where researchers with different levels of experience analyzed data of different levels of complexity and two meta-level studies analyzing the analysis results from the content-level studies. I also discuss a survey on SOUPS PC members' quality criteria as applied to qualitative analysis.

Chapters 4, 5 and 6 focus on research methods and reporting in quantitative research, specifically power analysis and effect sizes. Chapter 4 describes the methods and results of a manual literature analysis of 54 DCUS papers and includes information on effect size distribution for different variables used in DCUS. In Chapter 5, I present an LLM-assisted extraction process for statistical information and resulting effect size distributions in different research areas, based on information extracted from 599 CHI papers. Chapter 6 focuses on effect size judgment and interpretation. It describes an analysis of written effect size interpretations from 67 quantitative CHI papers and I identify five approach categories that researchers use when interpreting effect sizes. It further describes an interview and survey study conducted with researchers at two conferences and examines factors influencing researchers' interpretation of effect sizes.

Chapter 7 describes the second replication of Ion et al.'s study on experts' and non-experts' security behavior and opinion on security advice [236]. Besides detailing the methods and results of this iteration of the study, which include an updated status quo of experts' and non-expert's security behavior, I also present observations on methodological challenges in replications and replicability.

Chapter 8 derives recommendations for research practice from the work discussed in Chapters 3, 4, 5, 6 and 7.

Chapter 9 summarizes the findings of this thesis.

This thesis is based on work which has been published in five peer-reviewed full paper conference publications. An additional workshop paper provided the basis for a part of the background section. In acknowledgment of my co-authors and contributors to this research, I will use the plural personal pronoun "we" in place of singular "I" when describing this research. I highly appreciate their ideas, support and our collaboration. My collaborators' and my distinct contributions to individual publications will be summarized at the beginning of the respective chapters.

Chapter 2

Background and Related Work

Disclaimer:

Parts of this chapter were previously published as part of the following six publications:

1. “Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of Interview and Survey Study on Security Advice”, published at CHI 2023. This paper is the result of a collaboration with my co-authors Matthias Fassel, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz and Matthew Smith. This chapter uses text from this paper’s related work section. I conceived the initial structure of this section, conducted literature research on meta-studies of qualitative analysis and prepared a compilation of different coding practices used in USP research and wrote an initial draft of these subsections. Matthias conducted a literature research on quality criteria for qualitative analysis. The section was iteratively revised by Matthew, Katharina, Matthias and me.
2. “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”, published at SOUPS 2023, together with my co-authors Christian Tiefenau and Matthew Smith. This chapter contains text from the paper’s related work section. The first version was written by me, and it was revised by Christian, Matthew and me.
3. “Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation”, published at CHI 2025, together with my co-authors Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas and Matthew Smith. This chapter contains text from the paper’s related work section. I wrote the initial draft of the section, and it was iteratively revised by Florin, Lisa, Matthew and me.
4. “A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen’s d Effect Sizes”, published at CHI 2025 with my co-authors, Julia Angelika Grohs, Simon Lenau, and Matthew Smith. This chapter contains text from the paper’s related work section. I wrote the first version of that section, Simon contributed input on statistical terminology in a revised version, and Julia, Matthew and I conducted further revisions.

5. “Meta-Science in Usable Security and Privacy and HCI”, presented at Meta-HCI: First Workshop on Meta-Research in HCI at CHI’25, on which I collaborated with Florin Martius. Florin and I came up with the idea and conception for this paper. I conducted and wrote the literature research on meta-science in USP and the overview of calls for papers at USP venues and CHI. Florin and I summarized the recommendations derived from our previous meta-scientific work. We both participated in streamlining and revisions.
6. “Replication: "No one can hack my mind" - 10 years later: An update and outlook on experts’ and non-experts’ security practices and advice”, published at SOUPS 2025, with my co-authors Jenny Tang, Arthi Arumugam, Daniel Huschina, Lisa Geierhaas, Florin Martius, Luisa Jansen, Kolja von der Twer, Lilly Jungbluth, and Matthew Smith. This chapter contains text from the paper’s related work section. I conceived the initial structure of the section and wrote the first draft of the description of the original studies ([236, 74]). Daniel, Kolja, and Lilly conducted literature searches about specific areas of security behavior and security advice. Florin, Luisa and I iteratively revised the section.

This section introduces background on the overarching topics of this thesis, USP (Section 2.1) and meta-science (Section 2.2), as well as the three types of research which this thesis meta-scientifically explores, qualitative research (Section 2.3), quantitative research (Section 2.4) and replication (Section 2.5).

2.1 Usable Security and Privacy

The field of USP is situated at the intersection of HCI and the Security/Privacy community, and arose out of the attitude that users should not have to adapt to security mechanisms, but rather that these should be designed with users in mind for them to be successful [9, 566]. Garfinkel and Lipford identified eight challenges specific to USP research in their summary of the field from 2014 [170], which I summarize as following:

1. USP is **interdisciplinary**, using methods and aiming for goals from multiple research areas, both in the more user-focused area of HCI and the originally more systems-focused area of security and privacy.
2. Experts may already be **familiar** with the problem space in USP due the occurrence in every day usage. Elegant solutions to USP problems may seem simple after they have been introduced, leaving the work to develop them underappreciated.
3. The necessary balancing in the **interrelation** between security/privacy and usability makes designing solutions more complex than when considering each problem space individually.
4. Many studies rely on self-reported data but **user evaluations** of security are not reliable, complicating study design. Nevertheless, perceptions of security matter in addition to actual security when considering technology adoptions.
5. Security is often a secondary task to users making it challenging to design **ecologically valid** studies that do not prime participants to pay attention to security or privacy issues.

6. Studies in the context of security require **modeling an adversary**, who opposes the user. This adds complexity to the study design, and also to evaluations as users' perceptions of the adversary may influence the results.
7. Studies in USP evaluate technology which is subject to rapid changes and developments. This **velocity** can make replications difficult.
8. There is no clear single entity that benefits from or funds USP, i.e. could be considered a **customer** for USP research. This makes it challenging to decide who to address in USP research.

Conducting USP research includes challenges specific to this field of science and this makes it especially interesting and valuable to conduct meta-scientific research in this area. Two subdomains of USP research are of particular relevance to this dissertation: DCUS as the research area at the focus of the first publication on quantitative meta-science [369], and security advice, which was the topic of the content level studies in our evaluation of the qualitative analysis process [365] and of our replication study [368]. I give more context on these in the following.

2.1.1 Developer-Centered Usable Security

DCUS is a subfield of USP, but with a focus on the challenges and needs of expert users, such as software developers or administrators, instead of end users, who are at the center of typical USP-studies [185, 6]. DCUS extends USP's notion that security mechanisms should be designed with users in mind, to developers, which are themselves users, e.g., of cryptography APIs [185, 340, 4], programming languages [401] and other security tools [e.g. 395, 246, 28]. Tahaei & Vaniea provide an overview of topics addressed and methods used in DCUS [495]. While developers have been the main focus, this thesis also includes other expert users such as administrators in this field (e.g. [510]).

2.1.2 Security Advice

IT security advice has been investigated by numerous studies [236, 74, 414, 140, 171, 417, 215, 416, 553, 554, 437]. Prior work found a gap between expert security advice and end-user behavior [236, 74]. One reason for this gap is that end users are overwhelmed by the volume of security recommendations available, making it difficult to discern which advice is essential and actionable [417]. Users may reject security advice for several reasons: they may perceive it as marketing-driven, as a threat to their privacy, or as coming from untrusted sources [416]. One challenge in security education is that users must trust that their security-related actions will yield positive outcomes, despite these benefits being difficult to see or verify [76, 140], leading to rejection of advice due to a poor cost-benefit trade-off [215]. Also, users may believe that such practices are unnecessary due to differing mental models of security [553].

Compounding this issue, even security experts do not agree on best practices, making it challenging to establish a clear and consistent set of recommendations [236, 74, 417]. Security beliefs vary across different demographic groups [554], highlighting that a one-size-fits-all approach to security education is ineffective. In fact, security recommendations often need to be tailored to specific user

groups. For example, older adults have distinct security concerns and may require communication strategies that differ from those for younger users [335]. Similarly, software developers require security guidance that integrates seamlessly into their workflows [8]. Activists, who often operate under high-risk conditions, need advice that is contextually relevant to their security threats [63] and LGBTQ+ individuals may face unique privacy and security challenges that necessitate specialized guidance [171].

2.2 Meta Science

Meta research is the study of research, which according to Ioannidis et al., encompasses five major themes: investigating research methods, reporting practice, reproducibility of research, evaluation methods, and how to incentivize good science [235]. Methods meta-research focuses on aspects of performing research, i.e. includes research on study design, collaboration between researchers, data analysis, including statistics but also ethics and classical meta analyses which synthesize findings across multiple prior publications [235]. Reporting meta-research investigates the communication of research, through-out the research process, from study registration, informed consent procedures to reporting of results and communication of findings after publication [235]. Classically replication studies fall under the category of reproducibility meta-research, but it also includes other aspects of verifying research, e.g work on transparency and sharing of data and procedures [235]. Evaluation meta-research studies how research assessed, e.g. through peer review, and other means of evaluating research quality [235]. Finally, incentives meta-research studies how good research is rewarded in the evaluation processes at the focus of evaluation meta-research [235]. These five meta-research areas partially overlap, and at the same time are not exhaustive but they can provide an overview and broad map of meta-research.

Meta science received more attention in the wake of the replication crisis, which was a critical event, whereby many scientific results were shown to be difficult or impossible to reproduce [102], most famously in psychology [362] and medicine [233], but also in other fields, such as economics [45]. In the following, I provide an overview of meta-research in USP through an analysis of calls for meta-scientific work across venues publishing USP papers as well as an overview of meta-scientific work in USP and HCI.

2.2.1 Support for Meta-Scientific Research at Venues Publishing USP Papers

One way to evaluate the support for meta-scientific research at a venue is to analyze the published calls for papers for the respective venue. I conducted an analysis of the calls for papers for five venues where USP research is frequently published: SOUPS, USENIX Security, S&P, the ACM Conference on Computer and Communications Security (ACM CCS) and CHI for the years from 2005, the first year that SOUPS took place, until 2025 in the present. I extracted the calls for papers from the available conference websites, in some cases using the internet archive to access websites which were no longer available. The results of this analysis are presented in Figure 2.1. I identified four meta-scientific topics solicited in calls for papers: Replications, SoK papers, meta-scientific contributions named as such explicitly, and other meta-scientific contributions, which are

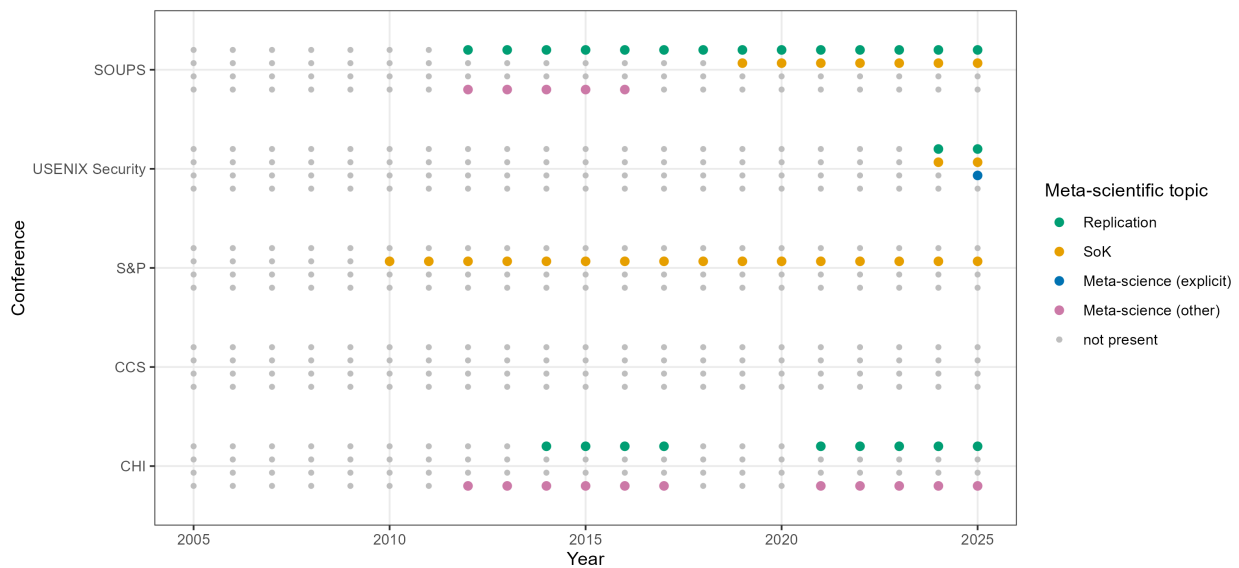


Figure 2.1 – Meta-scientific topics in calls for papers at venues publishing USP research

not clearly categorizable. The degree of detail and the formatting of information present in the calls for papers varied. The calls for papers for ACM CCS were extremely brief [305], with the main topical statement being merely:

The [...] ACM Conference on Computer and Communications Security (CCS) seeks submissions presenting novel contributions related to all real-world aspects of computer security and privacy. Theoretical papers must make a convincing case for the relevance of their results to practice.

This is likely contributing to the absence of any mention of meta-scientific contributions in these calls for papers.

S&P was the first of the analyzed venues to mention meta-scientific contributions in their call for papers in 2010, introducing SoK papers as a contribution:

In addition to the standard research papers, we are also soliciting papers this year focused on systematization of knowledge. The goal of this call is to encourage work that evaluates, systematizes, and contextualizes existing knowledge. These papers will provide a high value to our community but would otherwise not be accepted because they lack novel research contributions. Suitable papers include survey papers that provide useful perspectives on major research areas, papers that support or challenge long-held beliefs with compelling evidence, or papers that provide an extensive and realistic evaluation of competing approaches to solving specific problems. Submissions will be distinguished by a checkbox on the submission form. They will be reviewed by the full PC and held to the same standards as traditional research papers, except instead of emphasizing novel research contributions the emphasis will be on value to the community. Accepted papers will be presented at the symposium and included in the proceedings.

This explicitly welcomes papers contributing to the areas of methods and reproducibility. This type of contribution was later adopted by other security-focused venues, SOUPS in 2019 [528] and USENIX Security in 2024 [529]. In the excerpt from the 2010 S&P call for papers above, replications are solicited as “papers that support or challenge long-held beliefs with compelling evidence”, but not explicitly called replications. SOUPS was among the first of the analyzed conferences to explicitly list replications in their list of possible contributions in the call for papers in 2012 [100] as “reports of replicating previously published studies and experiments”. The conference put more emphasis on this type of contribution in 2017 [527], adding a paragraph giving details on types of replication papers and giving pointers on how to report on replications below the brief list of contributions. In 2019, the phrasing of this description was adjusted to what it still reads in 2025 [531, 528]:

In addition to original work, we are soliciting well-executed replication studies that meaningfully confirm, question, or clarify the result under consideration. Please prefix the title of these papers with the word “Replication:” for the review process.

Replication papers should aim to replicate important/influential findings from the literature. They may not necessarily offer new or unexpected findings; papers confirming previous findings are also considered contributions. Replication of a result that has already been replicated many times is less valuable. Replication of an obscure study that originally had only minimal influence on the community is less valuable. Authors should clearly state why they conducted a replication study, describe the methodological differences precisely, and compare their findings with the results from the original study.

Replication papers will be held to the same scientific standards as other technical papers. They should use currently accepted methodologies and technologies. Authors should not reuse outdated methods/technologies simply because they were used in the original paper. Replications may follow the same protocol as the original study, or may vary one or more key variables to see whether the result is extensible (e.g., re-running a study with a sample from a different population).

From 2012 to 2017, SOUPS also listed “reports of failed usable security studies or experiments, with the focus on the lessons learned from such experience” as a separate possible contribution. I categorized this as other meta-scientific contributions, as it can be seen as a methodological meta-scientific contribution but meta-science is not explicitly named.

USENIX Security adopted meta-scientific contributions relatively late compared to SOUPS and S&P but at the time of analysis it was the only venue explicitly soliciting meta-scientific contributions, starting in 2025:

Meta-science in Security and Privacy Meta-science, or the study of scientific research itself, aims to enhance the efficiency, quality, and outcomes of research activities in our community. Submissions in this broad topic should focus on evaluations of research practices, replicability/reproducibility, ethics, research methodologies, data transparency, and peer-review processes. Contributions should extend beyond analysis, aiming to influence future research practices.

Under the same heading, this call for papers also called for replications, explicitly naming “novel investigations into why certain studies fail to replicate” as a further contribution, similar to the earlier SOUPS calls for papers.

At CHI the calls for papers were often not published as a single page or document. Instead the information on possible contributions is spread across different sites. Structurally, the call for papers links to the “Guide to a Successful Submission at CHI”, which links to “Contributions to CHI” but also to a description of “Selecting a Subcommittee”, which describes contributions relevant to different subcommittees. Since these pages are linked at different points in the overall CHI website, it is somewhat hard to parse whether all content related to the call for papers has been analyzed. Consequently I may have missed some meta-scientific topics. While the structure described above applied to most of the CHI calls for papers I analyzed, there were some deviations, e.g. the “Contributions to CHI” page was not available in the form described above in 2018, 2019 and 2020. In addition to soliciting replications in the form of “Validation and refutation studies”, the “Contributions to CHI” also include methodology as a contribution, starting 2012, with interruptions until 2025:

Methodology contributes a novel method or method variation that helps us design an HCI artifact or evaluate an HCI artifact or setting. That is, it improves how we do what we do. The method may be intended for use in research or development settings (or both), but the paper should be clear about the intended audience. Your paper will likely include answers to the following questions: Is the methodology novel? Do you show how the methodology builds on or contrasts with related work? Does the methodology have relevance and importance to HCI practice and/or research? Is the methodology described in sufficient detail so that others can use it? Has the value of the methodology been demonstrated, e.g., by showing how it helps product design, development, and use, or by demonstrating its use in a particular setting, or by contrasting the results obtained with a competing methodology, or by showing how its use improved the process it was intended to address?

While the “method [...] intended for use [...] in development settings” is similar to what new implementations or more efficient algorithms might be in other fields of computer science, methods “intended for use in research” can be considered meta-scientific contributions. However, despite this type of contribution being listed, submitting meta-scientific work to CHI is not straightforward. At this venue, authors need to select two subcommittees to facilitate reviewer assignment to their submission. These subcommittees are mostly very topically focused, e.g. as examples from the subcommittees available when submitting to CHI 2025, *Games and Play* or indeed *Privacy and Security*. The subcommittees on *Understanding People*, using *Qualitative*, *Quantitative*, or *Mixed Methods*, represent somewhat of an exception as their categorization focuses on the methods used, rather than the topics investigated, but their description still focuses on user studies, and does not explicitly welcome, for example, meta-analyses. In summary, this makes it hard to choose an appropriate subcommittee for researchers aiming to submit meta-scientific work to CHI.

2.2.2 Meta-Scientific Research in USP and HCI

In the following, we present an overview of meta-scientific work in the USP community, along the five areas of meta-research posited by Ioannidis et al. [235]. We supplement this with meta-scientific

research published in broader HCI. While not all of the challenges of USP research apply to HCI, the two research domains are both interdisciplinary in methods, HCI research problems may also be familiar to experts and non-experts alike, and HCI is likewise affected by rapid technology developments. As such, we believe that many of the findings from meta-scientific research in HCI are transferable to USP and vice versa.

This is not intended to be a formal review, but a sample of meta-scientific work in USP and HCI. Le Pochat and Joosen recently reviewed meta-scientific work in security and privacy, including USP research next to more systems focused research [299]. We extend their work, on the one hand narrowing our focus to USP rather than security and privacy more generally, and on the other hand broadening the focus to HCI.

Most meta-scientific USP research focuses on methods, and several studies investigated aspects of study design. Danilova et al. [108] and Fulton et al. [165] compared security study tasks for developers – asking them to write, review, or fix code. Schechter et al. examined the effect of role-playing in studies [450] and found that role-playing participants behaved less securely than those using their own credentials. In multiple studies, Naiakshina and colleagues explored priming effects, finding that priming significantly impacted secure password storage implementation [345, 344, 343, 342]. Deception is used in USP [345, 121], but also HCI [259]. Danilova et al. found that removing deception in a password storage study did not affect outcomes [106]. USP research also explored differences due to the study environment, either within a single publication [142] or a replication in a different publication [479] and identified differences between online and laboratory studies. Like Fahl et al., Mazurek et al. also analyzed password characteristics and they compared real passwords from a university setting to passwords collected in studies on Amazon MTurk and passwords found in leaks [325]. They found that passwords created during MTurk studies were reasonably similar to the real passwords, while leaked passwords were weaker [325]. Redmiles et al. and Wash et al. found systematical differences between survey responses and field data, with respondents self-reporting faster software update speeds in surveys than observed in practice [418], and self-reported intentions regarding password use only correlating weakly with behavior [555].

Some USP research requires access to hard-to-reach participants, such as software developers. Meta-scientific work has focused on remote recruitment, e.g. for crowd-sourced data collection [107, 105, 461], and identifying suitable recruitment channels [260, 496]. Serafini et al. [459, 460] focused on recruiting software developers working in companies and work by Naiakshina et al. compared different types of participants with programming skill: students, freelance developers and company developers [342], finding effect sizes differed but trends aligned. More broadly, studies have examined the representativeness of online recruitment for security studies, noting shifts over time [415, 501, 255]. A meta studies found that even more participants in USP studies belong to Western, Educated, Industrialized, Rich, and Democratic (WEIRD) demographic groups than at CHI [207, 306] and that the way identity influences security behavior differs across demographic groups [560].

Meta-research in HCI has investigated aspects of research methods from various stages of the research process, from literature search [372], recruiting [440, 375], types of participants [306] and study methods [413, 440, 123], through measurement [223] and biases, like participant response bias [114] and social desirability bias [18] to data analysis of quantitative [445, 315, 410, 337, 547, 264], and qualitative data [404, 319, 167], and meta-analysis [585, 251, 249].

USP meta-research on reporting is not as frequent as methodological meta-research. Groß [189] and Coopamootoo and Groß [96] found that statistical reporting in USP is often incomplete. Tang

et al. found that most of the papers in their sample of SOUPS publications contain at least minor issues or deviations from best practice of statistical reporting and tests in 23% of the papers contain errors that lead to potentially invalid claims [500]. A poster publication on SOUPS future work statements suggested limited impact of these statements on future research [489, 490].

In HCI, various aspects of reporting have been explored, including reporting of participant compensation [383], openness and transparency [38, 442], positionality [473] and race and ethnicity [82]. Oppenländer describes the increase of citation counts in CHI papers and analyzes influencing factors [364]. Robinson et al. discuss conventions in writing CHI papers and the development of their AI-tool KITSUNE which aims to support the writing process [426]. Liu et al. discuss visualizations for communicating such decision-making during study design [315]. Research investigating visualization of uncertainty [220] and presentation of effect size [272] suggests that currently common ways of presenting statistics, such as confidence intervals [220] or means and mean differences [272] may lead to research readers overestimating effects. There have also been efforts to foster transparent reporting of statistical information [262, 263, 546, 125]. In an analysis of quantitative reporting at CHI Play, Vornhagen et al. find inconsistencies in statistical reporting, and misuse of hypothesis testing in studies intended to be exploratory [541]. Rogers et al. propose guidelines for how to report systematic reviews in HCI [430] and analyze reporting in currently published systematic reviews [429].

Regarding replicability, as we discuss in Section 2.2.1, several USP venues explicitly support replications, leading to published replications across various areas [51, 33, 501, 288, 390]. Meta-science of issues of replicability beyond replications themselves is less frequent in USP. Hamm et al. [197] found study descriptions and materials are increasingly available at S&P, ACM CCS, and USENIX Security, but data remain scarce. Guo et al. interviewed researchers and data curators to investigate challenges with de-identifying data, which include perceived unrealism of threats, lack of incentives and support and lack of knowledge and support of tools for de-identification [192]. Jallow and Bugiel found that the replicability of studies using data from the software development Q&A platform Stack Overflow is limited and recommend to treat this data as time-series data [240].

While the calls for papers at CHI have included replication as a possible contribution, the venue also places much value on novelty, which is sometimes seen as conflicting with replications [430, 370]. Nevertheless the RepliCHI movement has advocated for incentivizing replications at CHI between 2011 and 2014 [572, 573, 570, 571]. In their 2014 analysis of replications, Hornbaek et al. found that only 3% of the papers in their analyzed sample explicitly apply replications and that small changes in reporting results could greatly facilitate comparison to prior work [224]. Replications published at CHI are not explicitly marked as such, in contrast to e.g. SOUPS, where a prefix is applied to the title. This makes it more difficult to recognize replication papers. Cockburn et al. conducted an analysis of issues contributing to the replication crisis, e.g. dichotomous reporting of results and publication bias in the literature on empirical computer science, which includes HCI. Echtler and Häußler investigated the occurrence of published open source software at CHI and find that this practice, which contributes to open science, was not common in the proceedings for CHI in 2016 and 2017 [131].

Meta-research on evaluation is rare. In USP, Soneji et al. interviewed program committee (PC) members on peer review in security conferences [476]. While venues adapt submission and review processes, these changes rarely undergo scientific evaluation. Recent shifts at venues publishing USP work include the introduction of an artifact availability badge and review process, e.g. at USENIX

Security in 2020¹ and a move toward journal-style submission cycles at security conferences, e.g. S&P, starting in 2018 with 12 submission dates per year, then adjusting this to quarterly in 2021 and then thrice yearly in 2022. In addition, some venues publishing USP work, e.g. USENIX Security in 2014 and 2015 and S&P from 2016 to 2021 have implemented so called shadow PCs, whereby junior researchers participate in a separate peer review process independent of the main review process, in order to gain reviewing experience. This would represent a good opportunity for evaluation but to the best of my knowledge, any possible findings have not been published. In HCI evaluation meta-research, some publications analyze what factors lead to high citation counts, using citation count as a substitute for quality [252, 398]. The interdisciplinarity of the field is identified as an issue in the review process, when qualitative papers are reviewed by primarily quantitative reviewers [130]. Jansen et al. survey authors on the usefulness of the reviews they received for their CHI submissions [242].

We did not find USP meta-research which mainly focused on incentives for good science, even if some of the meta-research on reporting includes calls to action to improve statistical reporting [500, 96, 189]. In HCI, there are publications evaluating the extent of use of practices for good science [89] and publications discussing possible solutions for identified problems [90, 405, 120, 152, 382, 150]. Cockburn et al. identify many of the same problems which contributed to the replication crisis in other fields in publications in empirical computer science [89]. Other publications discuss incentives for replication [150], designing for pre-registration [405, 90] and open science [120, 152, 382].

In summary, there is a wide range of meta-scientific work published both in USP and HCI, albeit with slightly different focuses. Much of the meta-scientific work published at CHI is not published as a full paper, but rather in alt.chi, which “is a forum for controversial, risk-taking, and boundary-pushing research at CHI” [281] or in the form of an extended abstract, a format which is used to publish late-breaking work, workshop proposals or case studies. A possible reason for this may be that meta-scientific research is hard to categorize into the currently available subcommittees at CHI, which are used when submitting full papers. In our opinion, many USP meta-research findings, especially about methods, are at least partly transferable to HCI, and vice versa. For example, research on priming or recruiting software developers, though focused on security studies, can inform HCI research targeting the same demographic. It is necessary to reinvestigate issues outlined in descriptive meta-research [207, 197, 189], although for some topics, like participant demographics, the USP study [207] followed the methods of work in HCI [306].

2.3 Qualitative Analysis

Qualitative analysis was the focus of the meta-scientific work in our first paper “Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of Interview and Survey Study on Security Advice” [365]. Qualitative analysis is an approach for understanding how texts, e.g., interview transcripts, open-ended survey responses, or other documents, answer specific qualitative research questions, although it can also be applied to videos or images. Usually, researchers start with a *close reading* of the underlying material. Afterward, they assign labels, so-called *codes*, to relevant parts of the text to structure

¹<https://www.usenix.org/conference/usenixsecurity20/artifact-evaluation-information>

the content and get an overview of the entire material. There are two general approaches to assigning codes: *inductive coding*, where researchers create new codes based on the analyzed content, and *deductive coding*, where researchers apply a pre-defined set of codes. The set of codes is also called a *codebook*. Some researchers extract *themes* from the data for the analysis [64]. When the analysis involves multiple researchers, they usually discuss disagreements about code assignments. Sometimes they also calculate and report an *inter-rater reliability* value.

2.3.1 Coding Practices in USP

Many different coding practices are used in the field of USP. Recommendations like those of McDonald et al. [327] suggest that only one coder codes when a researcher has special expertise, which is common in ethnographic research [e.g. 487, 373, 520], or where data is simple, e.g., when coding survey responses [e.g. 164]. However, most published work in USP involves multiple coders coding the same data. These may code everything, either independently [e.g. 535, 286] or jointly [e.g. 494]. Multiple coders may also code a subset of the data, before splitting up the remaining data [e.g. 33, 164, 34, 497, 596]. The subset of double coded data ranges from 8.7% of responses [596], through 20% of data [497], 7 of 16 interviews (43.8%) [164] and one third (33%) of transcripts [34] to 16 of 26 (61.5%) transcripts [33]. Usually, splitting the data and continuing the coding process separately is justified by high agreement between multiple coders. Qualitative analysis encompasses more than just coding data and the coding process may serve different purposes. In all scenarios presented above, more researchers may be involved in different phases of the analysis process, e.g., when establishing a codebook or discussing findings and developing a theory.

Reporting and agreement measures differ as well. Agreement may be based on a codebook [e.g. 33] or the assignment of concrete codes to data [e.g. 497]. Different measures of agreement may be reported for specific topics in the codebook [e.g. 436], or as an overall agreement [e.g. 497]. Agreement procedures are often iterative and include phases of coding, discussions, and adjusting the codebook [e.g. 33]. They can also involve calculating inter-rater reliability (IRR) or similar statistical measures [e.g. 164]. Another aspect is the status of agreement after the coding process, i.e., whether all conflicts have been resolved and full agreement has been reached [e.g. in 176, 23], or whether disagreements remain [e.g. in 509, 497], and if and how these disagreements are discussed.

Researchers also refer to different methodological approaches, such as Grounded Theory [e.g. 344, 316, 109], the General Inductive Approach [e.g. 373, 520], or Thematic Analysis [e.g. 384, 535, 208]. However, the analysis process is then not always described in detail, making it unclear whether there is a shared understanding of the analysis process among researchers stating to use the same method.

In summary, this variation can make planning a qualitative analysis process daunting, especially for novice researchers. Due to the wide range of different analysis approaches used in the USP literature, researchers seeking to justify their methodological approach can always refer to a similar approach, or may even mix and match different approaches.

2.3.2 Quality Criteria for Qualitative Research

The review process ensures quality control of research before further dissemination. For quantitative research the quality criteria are well established, usually, reviewers focus on reliability, validity, and

generalizability [65]. If and how these criteria apply to qualitative research is part of an ongoing discussion [65, 284, 191, 301].

For example, IRR aims to measure the reliability of the coding process, e.g. how reproducible this part of the analysis is [127]. However, this is criticized as applying quantitative quality criteria to qualitative work and assuming that coding should be objective [65]. Where reliability is used to judge the quality of qualitative research, the term is used not necessarily in its form as exact reproducibility, but rather as consistency [301]. Guba [191] proposed a model of trustworthiness that instead focuses on credibility, transferability, dependability, and confirmability of research.

According to Guba’s model, reliability corresponds to dependability [191], which can be ensured by replication, peer examination, triangulation, dense description of research methods, code re-code procedure [284]. Two well-established methods of implementing criteria for trustworthiness are member checking, i.e., asking participants or members of the participants’ community to review results, and triangulation with other data sources [65, 327, 191]. Other methods include re-interviewing participants [327], maintaining transparency of the research process in general [327, 356] and the (theoretical) position of the researcher in particular [327], and using so called “thick description” with samples of raw data material [356]. Checklists for evaluating qualitative research, such as the COREQ [55] or Braun and Clarke’s list [64], have been proposed but not universally adopted into practice.

Despite the criticism of IRR, calculating IRR is a common practice in qualitative research. Reliability in qualitative research can be assessed within the same coder, between different coders and between coding schemes [327]. A related term, which refers to the type of reliability named second above, is IRR. It is a statistical way of measuring agreement between coders [327, 198]. They can be applied to ensure consistency between multiple coders, so that more data can be analyzed [327, 356]. If coding requires much interpretation of concepts and codes, to facilitate replication, or if researchers plan to report results in a quantitative way, e.g. using the results from qualitative analysis in statistical hypothesis testing, calculating IRR can be used to try to reduce confirmation bias of individual researchers [327]. During the coding process, IRR can be useful to discover where disagreements are [327] and to enable a more formal discussion about them [356]. Finally, in communicating results, IRR can be used to show that structures in the data exist beyond what a single individual sees, or, pragmatically, to use a recognized (quantitative) representation of reliability when publishing in venues not familiar with qualitative research [356]. By contrast, there are also situations when using IRR is not necessary, for example, when codes are not the product of the research, but just part of the research process, when coding is straightforward, for example marking the presence or absence of simple phenomena, in a social context when participants are also researchers or when there is a single expert researcher with unique experience and expertise, as is often the case in ethnographic studies [327]. Using statistical agreement can even be harmful in certain research domains, for example research on power structures or empowerment [327] or it can be abused as validation that reliability is sufficient, when its statistical measure meets a certain threshold [198]. In such cases disagreements should be openly discussed and not dismissed as negligible, because of a certain value of IRR. Finally, using IRR may contradict the agenda of qualitative research, which usually takes a more interpretivist stance towards the world [356].

2.3.3 Meta-Analyses on Qualitative Analysis

Prior work comparing different methods of qualitative analysis focused on specific methods or approaches. For example, Blair subjectively compared open coding and template coding [53], concluding that the coding technique should fit a researcher's mindset and research paradigm. Dufour and Richard found that using Grounded Theory and the Generalized Inductive Approach led to comparable results regarding the insight into the phenomenon, but there were differences regarding the depth of analysis reached [128]. Since each method was performed alone, with each researcher choosing the one with which they had more experience, there was no discussion of inter-rater reliability [128]. Thematic Analysis and Rapid Analysis were found to produce largely similar outcomes with much overlap but also some distinct findings [504]. However, these may be attributed to different levels of immersion into the topic for the different groups of researchers [504]. Wertz et al. analyzed the same data using five different approaches: Phenomenological Psychology, Grounded Theory, Discourse Analysis, Narrative Research, and Intuitive Inquiry [564]. While they discuss and compare these approaches and find that all incorporate some similar methods in their analyses, such as beginning with an open reading of the data, taking on a reflective stance, and letting patterns emerge from the data, there are also methodological differences [564]. Their analysis is not focused on assessing the similarity of results; instead, they acknowledge each researcher's approach to analysis as having a unique impact on the results [564]. Work about analytical pluralism, i.e., applying multiple qualitative analysis methods to the same data within the same study, also combines and compares the use of different qualitative approaches [88], e.g. variants of phenomenological analysis [273], variants of narrative analysis [162], or more different methods, like grounded theory, interpretative phenomenological analysis, Foucauldian discourse analysis and narrative analysis [163]. A single researcher can use multiple methods in their analysis [162, 567], or the different methods are applied by different researchers [163, 273].

Sanders and Cuneo investigate reliability in the coding process for a relatively simple, somewhat ordinal coding scheme applied to judge student submissions, and they focus on the social dynamics between coders [443]. However, their coding was not intended to aid sense-making [443].

Other publications investigate reliability but not in the context of qualitative analysis. Expert and novice users of a website accessibility evaluation tool were compared concerning their agreement in accessibility judgments [35]. Reliability of heuristics in the common usability method heuristic evaluation was also evaluated in various contexts [216], e.g., basic user interface elements [271], websites [491], and gaming [565].

2.4 Quantitative Analysis

The papers "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" [369], "Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation" [367] and "A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen's d Effect Sizes" [366] focused on topics related to methods and reporting in quantitative analysis.

In contrast to qualitative analysis, in quantitative analysis, researchers investigate the relationship between one or more independent, or predictor variables, and one or sometimes more dependent or

outcome variables with the goal of achieving evidence to support claims about the world, which are phrased as hypotheses. There are different frameworks used to conduct quantitative analysis, frequentist NHST and Bayesian analysis. NHST is the most commonly used framework in HCI [257] and based on anecdotal evidence, most quantitative research in USP also uses NHST. There has been critique of this state of the art, e.g. positing that NHST focuses on a binary decision of accepting or rejecting hypotheses and that interpretation of frequentist results is counter-intuitive, leading to common misconceptions around e.g. p-values [257, 264, 194, 499], confidence intervals [101, 218] and other concepts in NHST [188]. Because Bayesian analyses include distributions for prior and posterior estimates, rather than point estimates of distribution parameters, such as means or effect sizes, results from these analyses include more information for further analyses to build on [534]. Bayesian analysis also generally makes it easier for researchers to use this information to build upon prior work when analyzing their own studies, through the use of more informative priors [264]. Explicitly defining priors also makes assumptions more clearly evident [534].

However, given that NHST is still widely in use in HCI and USP [47, 189, 500], we focused our meta-research on this framework and will go into more detail on various relevant related concepts: Effect size (Section 2.4.2), power analysis (Section 2.4.3) and reporting of inferential statistics (Section 2.4.4). So, while these concepts may have a Bayesian equivalent, e.g. Bayes Factors or regions of practical equivalence are used similar to effect size indices [268, 317], when we describe these concepts in the following, we do this in relationship to NHST.

2.4.1 Null-Hypothesis Statistical Testing

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 [136]. Effect sizes measure the strength of the relationship of independent variables with dependent variables [136]. 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, assuming the null hypothesis is correct [136]. Commonly the threshold of $\alpha=0.05$ (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 [558, 434]. 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 [136]. 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 [274], 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 Chapter 6, 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.

2.4.2 Effect Sizes

There are many different types of effect size [275]. Simple effect sizes report the size of the effect in the units of the dependent variable [32], 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 [32], e.g. Cohen's d , or risk estimates such as odds ratios where variables are categorical. When comparing two means, e.g. with a t -test, the difference in mean completion time between two different interface variants represents a simple effect size, measured in units of time, e.g. minutes, while a standardized effect size for this scenario, such as Cohen's d , takes into account the standard deviation in the two groups. 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 [432], or risk estimates comparing different groups, like the odds ratio [151]. Especially for risk estimates, the base rate of the risk is important to evaluate the meaningfulness of an effect [151]. Some effect sizes additionally incorporate corrections, e.g. for sampling variability (adjusted R^2) or for shared variance (partial r) [151].

There is also meta-scientific work on effect sizes. For example, 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 U_3 , 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 [200]. Probably due to habituation, significance reporting without an effect size was also rated very informative [200]. 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 [272, 250]. 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 [272, 250, 220]. 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 [220]. On the other hand, Kale et al. found that showing densities, without explicitly specifying the means was most effective [250], while in Kim et al.'s study, the probability of superiority performed best [272]. 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 [272]. 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 [250]. This suggests that effect size interpretation differs depending on the task [250].

2.4.3 Power Analysis

A priori, or prospective power analysis is used to calculate the sample size necessary to detect an effect of a certain size at a specific significance level, with a specific statistical power. It is a procedure used in study planning.

2.4.3.1 Theoretical Background of Power Analysis

Four parameters are relevant to power analysis: Power, i.e., the probability of the test correctly rejecting the null hypothesis, the significance criterion α , the reliability of the sample results, and the effect size [91, 136]. The largest and invariably present influencing factor on reliability is sample size [91] - larger samples produce more consistent and reliable estimates than smaller ones. Consequently, the sample size is often used as a stand-in for reliability in power analysis.

The power of a statistical test is the probability of the test correctly rejecting the null hypothesis, i.e. that a statistical test yields a significant result, when the alternative hypothesis is true [136]. Power can also be represented as $1 - \beta$, wherein β is the Type II error, i.e. wrongly rejecting the null hypothesis. This means that if a test has a statistical power of 0.8, as is an often used acceptable value [91, 118], an actual effect will be detected 80% of the time. The significance criterion α represents the threshold of maximum accepted probability of making a Type I error, i.e. wrongly assuming the alternative hypothesis, detecting an effect, when there actually is none [91]. Using the widely accepted threshold of 0.05 for statistical significance means that only in 5% of cases, an effect is detected in the sample, even though in the population, it does not exist. Reliability refers to how well a sample estimate represents the corresponding population parameter [91]. Reliability is influenced by different factors, depending on the type of estimated parameter, such as the quality of the measurement instrument, and controlling sources of variance in the data, which might distract from the effect you are trying to measure [136]. The largest and invariably present influencing factor however, is sample size [91], such that larger samples produce more consistent and reliable estimates than smaller ones. Finally, the effect size measures the amount of impact of an independent variable on dependent variables, rather than only judging the presence or absence of an effect [136].

These four parameters are interdependent, such that when three of them are available, it is possible to calculate the fourth. These calculations are referred to as power analysis. In general, there are four different kinds of power analysis, each used to determine one of the parameters from the other three [91]. The focus of work in this thesis is on so-called a priori, or prospective power analysis, which is used to calculate the necessary sample size to detect an effect of a desired size with a chosen power and significance criterion, before actually conducting the study.

To be able to conduct an a priori power analysis, researchers need to know or guess either a standardized effect size they expect or want to detect or related statistics that can be used to calculate such an effect size. These effect sizes [136] and the resulting power analysis procedures [149] vary depending on the test used. The work in this thesis focuses on hypothesis tests with one categorical independent variable with two groups, and one dependent variable, i.e. the Fisher's exact,

chi-squared, McNemar's, Wilcoxon rank-sum, Wilcoxon signed-rank and independent as well as dependent-samples t-tests. The standardized effect sizes needed for these tests are ϕ for the chi-squared test, the odds ratio for McNemar's test, Cohen's d for independent samples, and Cohen's d_z for paired samples mean-comparison tests. While effect sizes can be converted between each other, what is generally seen as small, medium, or large differs between the effect size types [91], and this makes the process of guessing more difficult since not every researcher is familiar with the same types of effect sizes. Some procedures require different information, such the Fisher's exact test, where success probabilities for both groups are needed instead of a standardized effect size, or McNemar's test, which requires the proportion of cases, where changes occur in subjects' responses (proportion of discordant pairs) in addition to the odds ratio effect size. Alternatively, researchers can also guess unstandardized effect sizes, such as group means for both groups, i.e., the difference between these means. This can be more intuitive, especially when taking the approach of aiming to detect the smallest practically relevant effect sizes [136]. However, additional statistics are needed to calculate the necessary standardized effect size from these values. For the tests we focus on in this work, these are the standard deviations for the two groups for independent and paired t-tests, Wilcoxon rank-sum tests, and Wilcoxon signed-rank tests. Additionally, the correlation between the two groups is necessary for within-subjects tests with continuous dependent variables.

Several strategies can be employed to determine which size of effect to use in a priori power analyses. Estimates can either be based on similar studies from the literature [136], researchers' judgments of the smallest effect size of interest [292, 21]. or guidelines for judging the size of effects [136]. The most frequently used of these are Cohen's guidelines, whereby a Cohen's d standardized effect size of 0.2, 0.5 and 0.8 constitute small, medium and large effects respectively, and for Pearson's r , 0.1 is judged to be a small, 0.3 a medium, and 0.5 a large effect [91]. Guidelines for risk based effect sizes are not as commonly cited, likely because their interpretation is highly dependent on the base rate of risk in the population, i.e. when the base rate of risk is higher, then interventions lowering this risk are more relevant. Other guidelines exist, but are not as frequently referenced, e.g. Ferguson describes minimum thresholds for practically relevant effect sizes of different types in Table 1 [151]. These guidelines have been criticized for not taking into account context and methodological issues [151, 91, 465, 448]. Even Cohen himself cautioned against using the guidelines indiscriminately when proposing them [91](p.532). This warning has largely gone unheeded. To remedy this, various domain specific guidelines, have been proposed, based on actual effect sizes common in specific research areas [355, 448] but not yet in USP.

A priori power analysis is important, as both under- and overpowered studies are detrimental to the furthering of knowledge. Conducting an underpowered study means that failing to reject the null hypothesis is likely [136]. Since non-significant results are less likely to be published [26, 452], the effort in planning and conducting the underpowered study may be wasted. On the other hand, overpowered studies are wasteful, too [136]. Highly powered tests can detect very small effects so that in extreme cases, it is possible to find a highly statistically significant, albeit very small actual difference, which may be irrelevant in practice. Less power and fewer resources would have been sufficient to detect a practically relevant effect [136]. In addition to waste of resources, only collecting data from as many participants as necessary minimizes the amount of data collected, with positive effects on participants' privacy. Additionally, both underpowered and overpowered studies may be interpreted incorrectly when focusing on p-values. Underpowered non-significant results may be dismissed as irrelevant, even in the case of a large effect, while overpowered significant results representing trivial effects can be posited as important due to the statistical significance [136]. In

DCUS, it is especially important to be mindful when recruiting since developers as specialists are usually time-constrained, and payment is often much higher than in end-user studies.

2.4.3.2 Application of Power Analysis

The tools to conduct power analysis have evolved from power and sample size tables [91] to online calculators², designated computer programs, like G*Power [149] and multiple implementations in programming languages commonly used for statistical analysis, like the `pwr` package [79], among others in R or the `statsmodels` library [457] in Python.

Nevertheless, historically, power analysis has often not been applied [91, 48]. Unfortunately, in many fields, this is still the case according to more recent reviews, such as in psychology in 2014, where only 5% of 183 reviewed publications mentioned power analysis [537]. In other fields, often in the medical domain, power analysis is more prevalent, e.g., 43% of studies in a 2018 review of obesity interventions in schools [213] and over 60% of reviewed publications in NEJM and Lancet reported prospective power analyses in a review published in 2013 [518]. A possible reason is adherence to submission guidelines [517], which we recommend updating for the field of USP. If there is enough prior information, power analysis is recommendable for grant applications to ensure sufficient funds are planned for recruitment.

In HCI, power analysis is also frequently not applied. E.g., in 2018, only five of 519 experimental papers at CHI used prospective power analysis [134]. In interviews evaluating a prototype of a program to facilitate power analyses, some researchers were explicitly skeptical of power analysis as a research tool [134]. In a more cursory evaluation of terminology used in the CHI proceedings of 2017 - 2019, rather than manual inspection of publications, only between 1.5% and 2.7% of the papers containing the term “experiment” also contained the term “power analysis” [133]. In our own analysis of USP publications at SOUPS and CHI from 2020 to 2022, we found that only 7.8% of quantitative SOUPS papers used power analysis in some form and 15% of quantitative USP CHI papers did so. That means, over these three years, only 13 of 111 (11.7%) quantitative USP papers at CHI and SOUPS used power analysis in some way. One of those papers conducted post hoc power analysis, which is controversial since power and p-values are directly related, and if a result is not significant, i.e., the p-value is high, then power was too low to detect an effect of the size present in the sample. So little is gained by the post hoc power calculation [219, 182]. For another paper, it was unclear if the power calculation happened post-hoc or a priori. For most of the papers in our sample, description of power analysis, if present, was very brief.

Consequently, power to detect small effects in published studies is frequently low [136]. Literature shows this to be the case for such diverse research areas as management information systems, including ACM transactions [40], psychology [435, 458], and health professions education [95]. While not a formal review of power, Cockburn et al. examine empirical computer science literature and find evidence of the same practices that contributed to the replication crisis in other domains [89]. At the largest HCI conference, CHI, quantitative studies may even be underpowered to detect large effects [77].

When power analysis is not used to determine appropriate sample size, alternative approaches include recruiting the maximal number of participants possible, based on population, time, and

²e.g., powerandsamplesize.com or jakewestfall.org/power

monetary constraints [136], following prior practice and experience in the domain of interest [136, 77] or using rules of thumb, such as 10 or 15 cases of data per predictor [153], $50 + 8 \times k$, where k is the number of predictors in regression for testing the overall model [186], or two subjects per variable to estimate the coefficients in linear regression [27]. However, none of these methods guarantee that studies will have sufficient power [136].

To increase power, especially in fields like DCUS, where recruiting is challenging, researchers should also consider adapting their research design to increase the reliability of measurements and reduce random errors, e.g., by conducting within-subjects research [136]. Another possibility is conducting sequential analyses, whereby a study can be stopped at planned intervals if a large enough effect can be detected at this time, which on average reduces the necessary number of participants [292]. There is some discussion about whether stopping studies early like this introduces an additional bias towards larger effects [42, 334] or not [455, 159].

2.4.4 Reporting of Statistical Methods and Results

Reporting statistical results accessibly, understandably and completely is important to ensure other researchers can build upon prior work, and form their own interpretation of findings. The SOUPS calls for papers between 2010 and 2018 included a guideline on “Common Pitfalls in Writing about Security and Privacy Human Subjects Experiments, and How to Avoid Them” by Stuart Schechter [449], which also included some information on reporting statistical results: “For each test, remind the reader of the hypothesis to be tested, present the data used to test the hypothesis, explain your choice of statistical test, and then present the result of that test”. The document discusses e.g. reporting on checking of assumptions for tests, corrections for multiple testing and discussing results but does not give details on what information should be reported for various statistical tests and how it should be reported. The CHI guide to a successful submission stresses that “[...] statistical analyses should be described with significant detail” and that such papers “should include enough detail for an independent researcher or practitioner to (1) independently evaluate the correctness validity, and reliability of your [...] analyses and (2) reproduce and replicate [...] experimental methods”, in its Transparency Section [512]. However, it neither provides more detailed guidance on how to achieve this, nor references guidelines that do so. A special interest group on transparent statistics took place at CHI’16 [262] and the group has since organized further discussions at CHI and IEEE VIS [515, 546, 263]. Work is ongoing on a guideline for transparent statistics reporting in HCI, with the section on effect size currently in alpha, and other sections in the drafting stage, however, changes to the final document have not been made since 2019 [514]. The topic is nevertheless current and being pursued in the USP community. The Workshop on Developing Research Transparency Guidelines took place at SOUPS 2025, and explicitly included drafting reporting guidelines, with the aim to publish them by the end of 2025 [144].

Other disciplines have guidelines on reporting statistics which are endorsed by some publication venues, of which we discuss examples. In medicine, extensive guidelines exist for many specialized variants of study designs [348, 349]. One of the most frequently referenced is the CONSORT checklist [331]. As the term checklist implies, this does not give detailed instructions on how to report information, but rather specifies what to report. Evaluations comparing quality of reporting of endorsing venues and non-endorsing venues suggest that it has a positive effect on reporting, although problems with reporting remain [394, 522]. Other medical reporting guidelines are more

detailed, e.g. even the abbreviated website version for the AHA/ASA journals specifies the order in which to present statistical values [248]. A more comprehensive version gives some examples about commonly used effect sizes in the application domain and consolidates the CONSORT guidelines within their section on randomized controlled trials [15]. The publication manual of American Psychological Association (APA) gives detailed instructions regarding various aspects of the publication process, including reporting [25]. The main text remains non-specific due to the variety of possible tests, but the sample tables include examples of the information which should be reported for several commonly used tests in psychology, including t-tests, χ^2 -tests, ANOVA and regression [25].

Prior work suggests that reporting statistical results in a complete and transparent way is a challenge that applies to many disciplines, e.g. usable security [96, 189, 500] and HCI [78, 352, 442] but also other disciplines related to aspects of HCI, such as psychology [354], social science [486] and software engineering [253]. Coopamootoo and Gross found in their analysis of studies in USP, published between 2006 and 2016, that methodological information was mostly reported to a sufficient degree to enable replications but effect sizes were largely not reported, with only 6% of the considered studies providing sufficient information for meta-analyses [96]. Tang et al. found that 97% of the papers in their sample of SOUPS publications from ten years between 2005 and 2024 contain at least minor issues or deviations from best practice of statistical validity, reporting or interpretation [500]. Issues included using the wrong hypothesis test for the type of data in the analysis (e.g. independent test for dependent data), lacking information to judge scientific or practical significance, i.e. not reporting on effect sizes or interpreting the results on the wrong scale, i.e. reporting changes in log odds as changes in odds [500]. Salehzadeh Niksirat et al. investigated reporting practices at CHI, including of statistics, and found that reporting of descriptive statistics, clearly stating the test procedure, and reporting of test statistics and p-values was largely sufficient, but reporting of effect sizes and even more so, confidence intervals was lacking [442]. Besançon and Dragicevic focused on the communication of results as dichotomous, i.e. using significant vs. not significant as a proxy for the results being interesting or relevant and found that phrasings related to dichotomy are common at CHI [47]. Dragicevic et al. propose reporting multiple possible analyses, i.e. multiverse analyses, to promote transparency of statistical reporting [125]. Reporting of power analyses in publications may not be sufficient to reproduce the analysis. In an analysis of 105 preregistrations and 155 IRB proposals in psychology, Bakker et al. found that while between 45% and 72% used power analysis to decide on sample size, reporting on these power analyses was largely incomplete [37]. Only 20% of the descriptions of power analysis were sufficiently complete for replication, and not all of them matched the main hypothesis under investigation [37].

2.4.4.1 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 identified the CHI USP papers by searching the conference proceedings for sessions and papers 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. [254] 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 [91] 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. [56] 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 [442] in 2023 and suggests an improvement from 2020 [272]. However, in our sample, deeper discussion of effect sizes and size judgment justification was sadly not common.

2.5 Replications

Replications are an aspect of meta-science which is quite often solicited in calls for papers, as discussed in Section 2.2.1. Conducting and publishing replications enables validation of published results, and assessing the development of results concerning a particular research area over time. Our paper “Replication: “No one can hack my mind” - 10 years later: An update and outlook on experts’ and non-experts’ security practices and advice” [368] represents an example of a replication study.

To be able to replicate prior work, information about its methodology, relating to both data collection and data analysis need to be available to the researcher who wants to conduct a replication. When discussing replicability, transparency and open science are also often mentioned. While there is no universally accepted definition, Vicente-Saez et al. defined Open Science as follows, based on a systematic literature review: “Open Science is transparent and accessible knowledge that is shared and developed through collaborative networks” [539]. Different recommendations and practices fall under open science [468], from providing material on request, to publishing study material, analysis scripts, the data itself or even pre-registering studies [62]. Material provided can alternatively be broadly categorized as research design related, research infrastructure and research data [168]. Nevertheless, making this available to share and re-use has challenges, such as maintaining continuity, availability and comprehensibility of the shared knowledge [168]. In their interview study with USP researchers, Klemmer et al. found that benefits of transparency often apply to the community as a whole, while drawbacks concern individual researchers or publications [277]. The drawbacks discussed by their participants include making themselves vulnerable to criticism in the review process when publishing full study materials and the time and effort necessary to

publish study materials and data [277]. However, sharing the full data analysis and results not only makes science more transparent, but also helps researchers estimate effect sizes to be used for power analyses in their future studies or for meta analyses.

Klemmer et al. conducted a literature analysis of the current status of transparency in USP research considering 52 criteria [276]. They found that some practices, like reporting on the study protocol, sample size or recruitment strategies are widely practiced, while others, like providing the codebook for qualitative analyses or data analysis code or software used for quantitative analyses, less so [276]. While full transparency, such as beginning a scientific inquiry with pre-registration, is not yet common in HCI and USP in comparison to disciplines like psychology [471, 38], there are nevertheless some examples of open science practices and re-use in our community. There are many examples of re-using methodology. Validated questionnaires, such as the System Usability Scale [68, 279] in HCI or the Internet Users' Information Privacy Concerns questionnaire [318] and the secure software-development self-efficacy scale [542] in USP are commonly used. Re-using questionnaires is not limited to validated ones, non-validated questionnaires are also re-used, such as Naiakshina et al. [342] reusing some questions about software developers' background from Assal & Chiasson's work [24]. While a collection of experiment tasks, such as the Repository of Assigned Search Tasks for interactive information retrieval [160] does not exist for USP, tasks are nevertheless reused across studies. An example is Geierhaas et al. [173] basing their password storage task on a task in Acar et al.'s study [7]. There is also re-use of infrastructure in USP, e.g. the Developer Observatory [483] was used in studies investigating the effect of security re-ranking on code security [156], the security development practices of Github users [7], and a usability evaluation of a website presenting security related code snippets [173]. Finally, as discussed in Section 2.2, replication studies, i.e. re-use of results, also regularly appear at SOUPS, e.g. [288, 501, 390], and are explicitly solicited in the call for papers (see Section 2.2.1).

2.6 Positionality statement

As a researcher conducting research about research, I want to give some background on my own research experience, practice, and background to provide context on how to interpret the work done in this dissertation. I believe that the conclusions which are drawn from data depend to some extent on the researchers that conduct the analysis, and their personal background and experience. This aligns with a relativistic ontological viewpoint and a social constructionist epistemological stance.

Courses exclusively on statistics were not part of my program of studies during my Bachelors and Masters. In my course work as an undergraduate student, more time was spent on quantitative methods of data analysis than qualitative methods, specifically NHST, although data collection methods were also covered. This focused mostly on methods that HCI researchers were likely to use in their studies and at this point did not involve practical exercises. I was a student assistant for the empirical methods course later, when it was undergoing a re-design to make it more hands-on. While designing practical exercises for the students, I learned the material myself, although of course the lecturers for this course were available to me for questions and feedback. Hypothesis tests covered in the course were hypothesis tests with one independent variable with two groups, and one dependent variable, and in addition, correlations, simple and multiple linear regression and analysis of variance (ANOVA). The course was application focused, so that students could use

the tests and interpret the results with respect to statistical significance but it did not focus much on statistical theory. Topics of this thesis, such as power analysis, effect sizes and replications were mentioned in this course, as something you should do or consider, but since the course covered a lot of material, there was not enough time to go into details. Throughout my undergraduate studies, when working on student projects, which were all constrained to fit into a single semester, there was likewise not enough time to research and run, e.g. power analyses for the hypothesis tests which we used in our analyses. During these projects, when making decisions, e.g. about study design, data collection or analysis and asking for feedback, I often heard “It’s hard to say which is better, both are okay as long as you can justify the decision”. While at that point, this was sometimes not very satisfying to hear, it stuck with me. Starting my PhD, I had the time to consider these topics in more depth and try to back justifications up with empirical data.

Despite my focus on methodological topics in my research, I see myself as a researcher using these tools, like power analysis, hypothesis testing or qualitative coding to get research results, not as a statistics or methodological expert, and I believe that many researchers in USP feel the same. I am trying to apply the recommendations derived from my research to current projects, e.g. on reporting the hypothesis tests conducted for our re-replication of the “No One Can Hack My Mind” study, discussed in Chapter 7. As such, I recognize that external constraints like limited expert counseling, page limits and limits on appendices in publications, and reviewers’ feedback influences the form of publications and that not all recommendations can be applied at all times.

Chapter 3

Influence Factors on Results from Qualitative Analysis

Disclaimer:

The contents of this chapter were published at CHI 2023 as part of the publication “Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of Interview and Survey Study on Security Advice”. This paper is the result of a collaboration with my co-authors Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz and Matthew Smith, as well as the student participants of the Usable Security and Privacy Exercises at the University of Bonn in summer semester of 2021 and 2022, and Anastasia Danilova, Alena Naiakshina and Lea Gröber. Matthew Smith and I conceived the research questions and idea. Later, Matthias, Lea, Katharina, Alexander, Matthew, and I discussed and decided on an extension of the research questions, including the survey of PC members. Anne, Matthew, and I designed the content- and meta-level studies for the summer semester of 2021. Florin and I adjusted the study design from 2021 to plan the content- and meta-level studies for the summer semester of 2022. Students of the Usable Security and Privacy Exercise conducted the content-level study in 2021. Anne and I conducted the meta-level study in 2021. Florin and I conducted the meta-level and content-level studies in 2022. Alena, Alexander, Anastasia, and Matthias, as well as the student participants from 2021 each analyzed a subset of the content-level data from 2021. The student participants from 2022 analyzed the content-level data from 2022. Anne and I analyzed the the full set of content-level data from 2021 and conducted the meta-analysis for this data. Florin and I analyzed a subsample of content-level data from 2022 and conducted the meta-analysis for this data. Florin, Matthew, and I planned the survey of SOUPS PC members. Matthew carried out the recruitment, Florin and I coded the free text answers of the survey, and I calculated descriptive statistics. Finally, I prepared the paper for publication together with Matthias, Katharina, Florin, and Matthew. In particular, Matthias was responsible for related work on qualitative analysis, Florin was responsible for figures, I was responsible for writing the methods and results section, and all of us took part in revisions and proof-reading.

3.1 Motivation

Qualitative content analysis is a common research method in the field of HCI. Different forms of qualitative analysis are used to generate theory [e.g. 487], investigate novel topics [e.g. 584] and inform follow-up quantitative analyses [e.g. 164, 509] or add depth to existing ones [e.g. 108, 340]. One commonality between many types of qualitative analysis is the process of coding textual, image, video, or audio data, where one or more researchers assign so-called codes to segments of data [564].

In this chapter, we will examine aspects of this coding process, including the effects of multiple coders for two different kinds of data, inter-rater reliability as well as the effect of experience and background on coding outcomes. This research was done in the context of the HCI sub-field of USP. The sub-field was chosen since the authors have been involved in the paper reviewing and acceptance process as reviewers, PC members, area chairs, and program chairs in the area of usable security and privacy. In these roles, the authors have participated in discussions about the quality criteria of coding practices that papers need to fulfill to be acceptable. For instance, we have seen lively debates and papers being turned down, with the argument that only one researcher coded all the data weighing heavily in the decision. In these debates, arguments were based on informal common practices within the community. This is similar to reviewers referencing community norms on conducting usability evaluations [187] in HCI and for adhering to perceived statistical standards [30], e.g. regarding sample size [31] in medicine.

Consequently, when advising our students, we usually recommend the use of at least two coders to fulfill the informal quality criteria even when other guidelines suggest that these quality criteria are not applicable in all cases, e.g., that when data is straightforward and easy, multiple coders may not be necessary [327]. To offer empirical insights into these kinds of decisions, we present the results of an experiment evaluating the coding process involving 65 students and 7 researchers. We limit our claims to the USP sub-field since we are less familiar with the informal quality criteria in other sub-fields. While our motivation and analysis are rooted in this sub-field, we hope that our results will be helpful in the broader HCI community as well.

We conduct a study to analyze qualitative coding outcomes of coders working in groups of two or three, to answer the following research questions:

RQ-Qual-1.1: How does simple vs. complex data affect the similarity of results of two coders within a group of coders?

RQ-Qual-1.2: How does simple vs. complex data affect the similarity of results between groups of two coders?

RQ-Qual-2: How does coders' previous experience with qualitative data analysis affect the results?

RQ-Qual-1 was motivated by our experience of discussions on this topic during the paper review process and the divergence between recommendations of always using two coders [73] and more relaxed recommendations [327]. RQ-Qual-2 arose when we found differences between student and researcher coders. Prior work has examined the influence of researcher characteristics, such as epistemological stance [163, 273], experience [163] and background on qualitative analysis outcomes

when using different qualitative analysis methods [88]. To look at researcher influence, we compare outcomes when different researchers use the same method. To compare our results with the reviewers' perspective, we additionally explored the following research question:

RQ-Qual-3: Which quality criteria do reviewers currently apply to qualitative research in the field of USP?

Based on McDonald's definition of simple and complex data [327], we replicate and extend parts of the "No one can hack my mind" study [74, 236] to collect what we think are good representatives for simple and complex data in the field of USP. For the simple data, we asked participants to state security advice using a survey instrument. We consider this data to have only a small interpretative range and thus could potentially only require one coder. For the complex data, we interviewed participants about their views on the practicality and effectiveness of that security advice, leading to data with a larger interpretative range.

We had a total of 65 students who were taking a course in the field of USP, and seven researchers analyze different parts of this data. We analyzed their codebook creation process, similarity of outcomes, inter-rater reliability and compared the student to the researcher outcomes. We also surveyed SOUPS PC members from the last five years about their views on coding. We evaluate and discuss our findings in the context of current reviewer preferences and make recommendations for authors and reviewers for whom this data is applicable.

3.2 Empirical Evaluation of the Coding Process in Qualitative Analysis

In the following, we present two content-level studies on security advice, one generating complex data from interviews and the other simple data from survey answers. Two meta-level studies analyze researchers' and students' analysis processes for these content-level study data. Table 3.1 shows an overview of the involved participants. Table 3.2 provides an overview of the conducted studies. While the R abbreviations in the tables represent a single researcher, the S abbreviations represent a group of students (usually two) since we always analyze them as a group.

3.2.1 Choosing a Suitable Security Topic as a Basis for our Meta-Study

In their review of qualitative analysis practices, McDonald et al. describe different scenarios when or not to use IRR [327], with one aspect centering on the ease of coding. Concrete statements can be coded with relative ease when trying to code, e.g., for the presence or absence of clearly defined phenomena or when the coding task itself is clearly specified. On the other hand, less concrete statements are harder to code, especially when the exact area of interest is not known before the coding process and researchers are identifying areas of interest through the coding process. To investigate this empirically, we chose a topic from the domain of USP where we could gather these two different types of data: Security advice.

Abbreviation	Experience	Involvement	Institution
S1c - S15c	Students 2021	interviewed participants of content-study-complex coded data of content-study-complex	A
S1s - S19s	Students 2022	coded data of content-study-simple	A
R1	Researcher	coded data of content-study-complex	A
R2	Researcher	coded data of content-study-complex	A
R3	Researcher	coded data of content-study-complex co-author	B
R4	Researcher	coded data of content-study-complex co-author	B
R5	Researcher	course instructor 2021 coded data of content-study-complex coded data of meta-study-complex co-author	A
R6	Researcher	course instructor 2022 coded data of content-study-simple coded data of meta-study-simple co-author	A
R7	Researcher	course instructor 2021 and 2022 coded data of content-study-complex coded data of content-study-simple coded data of meta-study-complex coded data of meta-study-simple co-author	A

Table 3.1 – Overview of all parties involved in the empirical evaluation. S abbreviations represent groups of students. R abbreviations represent single researchers.

Study	Type	Researchers	Participants	Year
content-study-complex	interview	S1c - S15c, R1 - R5, R7	recruited by S1c - S15c	2021
content-study-simple	survey	S1s - S19s, R6 - R7	recruited on MTurk	2022
meta-study-complex	meta-study	R5, R7	S1c - S15c, R1 - R5, R7	2021-2022
meta-study-simple	meta-study	R6, R7	S1s - S19s, R6 - R7	2022

Table 3.2 – Overview of the studies conducted for the empirical evaluation of the coding process.

We conducted two studies, see Table 3.2: (1) content-study-simple, a survey on security advice and related behavior, replicating Ion et al.'s and Busse et al.'s work [236, 74], and (2) content-study-complex, a qualitative interview study about reasons for trusting security advice and perceptions of realism and effectiveness.

On the one hand, security advice itself is a concrete statement or recommendation and can be coded with relative ease to identify advice, especially since previous work already identified categories of advice. In our study, this type of data, which we collected in a survey, is represented as simple to code (**s** in participant quotes). On the other hand, reasons and opinions regarding the judgment of security advice are not as straightforward. For example, experts disagree on the effectiveness and realism of security advice [236, 74]. We consider this type of data, which we collected through interviews, complex to code (**c** in participant quotes). We recognize that complexity does not arise purely through the method of data collection, however, within our meta-analysis, the survey data represents simple data, and the interview data embodies complex data. Other USP examples for simple coding could be identifying types of vulnerabilities [543], assigning security scores [344, 184] or other cases, where the scope of coding is well-defined. These examples have in common that segmentation is not an issue, and that coding, once the codebook is established, pertains to judging whether a concept is present or absent in the data. Further examples of coding complex data are coding sketches of mental models of various concepts [256, 285] or exploratory data analysis of interviews about implementing cryptography [201]. In general, when it is not clear what aspects of the data should be coded, as in open coding [484] or more general, in inductive coding, this can be considered complex.

In general, we believe these two types of data are typical for commonly analyzed data in the USP domain regarding the degree of necessary interpretation. These two content-studies serve as the basis for our meta-study about the coding process.

3.2.2 Meta-Studies

To understand how different data types and researcher experience affect the outcomes of qualitative analysis we conducted a meta-study based on the two content-level studies described above. At the meta-level, we compared how coders at different experience levels (students, researchers) perform the qualitative analysis of the content level following a general inductive approach [507].

To investigate the usefulness of multiple coders, we wanted to compare the outcomes of the coding process within (RQ-Qual-1.1) and across multiple different groups of coders (RQ-Qual-1.2). We started by recruiting students from an introductory course on scientific methods and USP, who participated in designing and conducting content-study-complex. To analyze reliability, multiple students coded the same sets of interviews and submitted their progress at predefined points in the analysis process. The two researchers teaching the course (R5 & R7) coded the interviews following the same procedure to have an overview of the data and compare the students' assessment to their own. During the data analysis process, we noticed differences between the students and the instructors, which we hypothesized, were due to the difference in expertise. To test this (RQ-Qual-2), we recruited two additional researchers from the same research group (R1 & R2) and then two researchers from a different institution (R3 & R4).

To investigate the effect of simple and complex data (RQ-Qual-1), and empirically evaluate the guidelines put forward by McDonald et al. [327], we collected security advice in a survey like the

one used by Busse et al. [74] in the next iteration of the aforementioned course, and repeated the analysis process with a different sample of students, and R6 and R7 as the instructors of that course iteration.

3.2.3 Interview Study (Complex data)

In the following, we describe the data collection and analysis process with regards to the complex interview data.

3.2.3.1 Interview Procedure

We conducted the study in the summer semester 2021. During the course and exercises, students learned about interviewing and participated in developing an interview guideline for a semi-structured interview.

The interviews extended Busse et al.'s work on IT security advice [74] by focusing on participants' reasons for selecting and trusting specific pieces of advice. We especially asked participants about the advice's effectiveness at keeping users secure and its realism, i.e., the likelihood that users follow advice, since Busse et al. highlighted this contentious topic. The interview script was designed to answer the following content-level research questions:

- Why is some advice rated effective, but not realistic?
- What factors influence how much advice is trusted?

3.2.3.2 Content-Study Analysis

Participating students each recruited an interview participant, conducted the interview via Zoom and recorded the audio for transcription. A professional transcription service transcribed the recordings using the simple rules of Dresing and Pehl [126]. As shown in Figure 3.1, we selected a random sample of 18 of 28 of the interviews and split this sample into three subsamples of six interviews each. We balanced the length of the interviews in each subsample, so that students had similar workloads. Students analyzed one of the three interview subsamples. We followed this approach of dividing the data and having multiple subsamples for analysis to ensure that characteristics of the analysis outcomes were not due to the exact analyzed content.

R5 and R7 went through the transcriptions and checked them for accuracy. For each of the three subsamples, we selected two initial interviews, for which we checked they were sufficiently complex to yield multi-faceted codebooks. For the content-level analysis process described in the following, everyone who coded the data of content-study-complex (see Table 3.1) was advised to base their coding on the research questions named in Section 3.2.3.1. Each student chose a research partner to work with on the coding task and followed a general approach to inductive coding [507]. We randomly assigned partners to those who had not chosen a partner. Students submitted their progress at predefined steps in the coding process to allow us to monitor their work, see Figure 3.2.

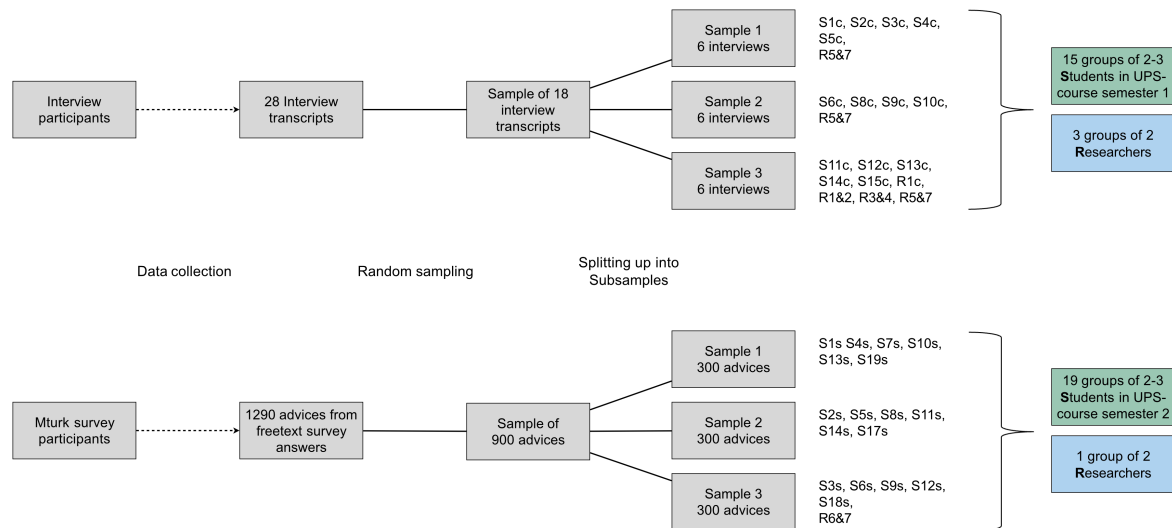


Figure 3.1 – Process of Data Preparation and Visualization for Empirical Evaluation

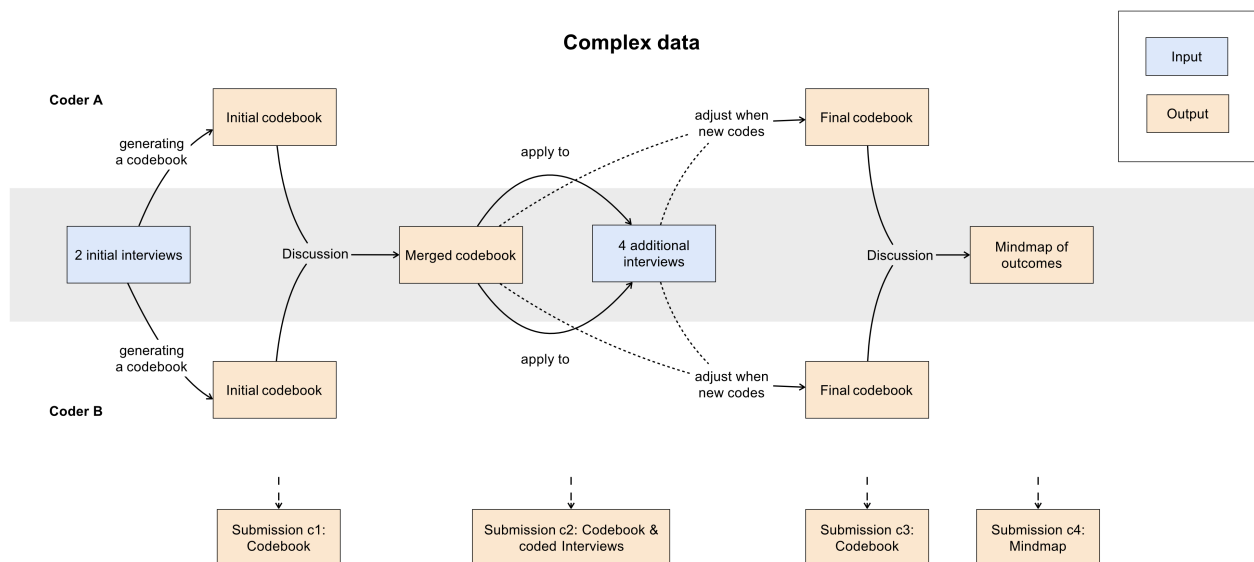


Figure 3.2 – Interview coding process and submitted outcomes

First, students individually coded the two initial interviews and submitted their initial codebook with their coded transcripts (submission c1). They refined and merged their codebook with their chosen research partner, independently recoded the initial interviews with the merged codebook, and submitted the merged codebook and the recoded interviews (submission c2). Students then coded the remaining four assigned interviews using the merged codebook. If they found that new codes were necessary, they could add those codes and discuss them with their research partner. They then submitted the remaining coded interviews and the final codebook (submission c3). Finally, the students developed themes from their codes and summarized the most important ones in a mindmap, which they also submitted (submission c4).

R5 and R7 followed the same process as the students, but for the whole set of 18 interviews. They had already seen all of the 18 interviews during the anonymization process, and thus could not unbiasedly work on a subset anymore. R1 - R4 followed the same process as the students, for a subsample of 6 interviews.

3.2.3.3 Meta-Level Analysis

We used a meta-level analysis to understand how different researchers analyzed the interview content. The meta-analysis of the interviews was conducted by R5 and R7, jointly, using the printed documents, whiteboards, sticky notes, and the software MaxQDA [323]. Non-digital phases of the analysis were documented through photographs and partially digitized to make the results more accessible for further analysis. We traced the appearance of concepts throughout the analysis process, to see which of those present in the final result (mindmaps) had already been present in the initial codebooks. To answer RQ-Qual-1.2 and compare the analysis results with respect to similarity between groups of data analysts, we analyzed mindmap content and structure. We inductively and iteratively developed a codebook of structural aspects of the mindmaps. When noting a new, repeatedly occurring structural element, we recoded previously analyzed mindmaps. Mindmap content was coded as related to the content-level research questions, i.e., either to effectiveness, realism, or trust (or a combination of two or more of these) and only presence or absence of concepts was coded. We were liberal in assigning codes, coding a concept even when it was adopted from the interview guideline, so the results can be considered an upper limit of similarity. We only coded concepts when we recognized which research questions they belonged to, or else if the relationship to a different abstract topic was clear.

To investigate the influence of different data analysts within a group on the coding process (RQ-Qual-1.1), we coded the initial codebooks with respect to their content, using the same codebook as for the mindmaps to be able to see which concepts came from which of the researchers. In our analysis we again consider only presence or absence of concepts, even though we coded multiple instances of the same concept per codebook, to prevent overlooking concepts.

We also used MaxQDA to calculate the achieved IRR for all coded interviews in the final submission, as well as separately for the initial two and the remaining four interviews as coded in the final submission. Since we had not pre-specified segments prior to coding, we chose to calculate IRR for different levels of minimum overlap between segments and chose 85% overlap to use in further analysis, as this was the point where agreement levels stabilized. This accounted for small differences, such as one coder deciding to exclude punctuation marks or additionally coding filler words.

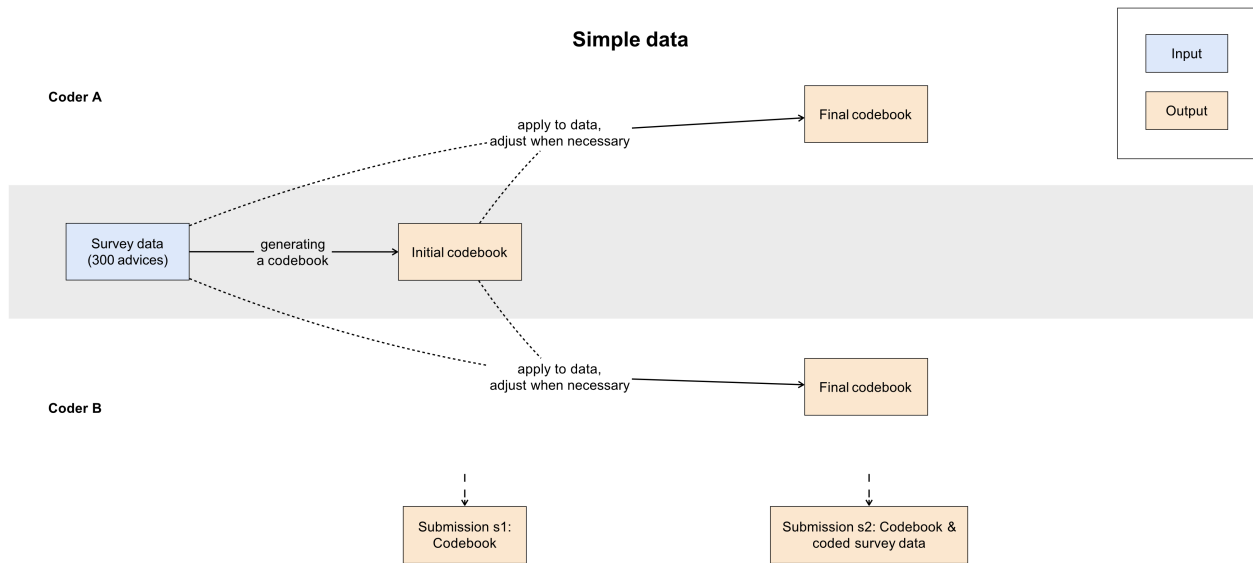


Figure 3.3 – Survey coding process and submitted outcomes

To summarize, to answer part of RQ-Qual-1, we analyzed artifacts (mindmaps, codebooks, and coding) of the analysis of complex data for similarity, both between groups of data analysts, e.g. when comparing final results (RQ-Qual-1.2) and within a group of data analysts, e.g. by calculating IRR or conducting change tracking from the initial codebooks to the mindmaps (RQ-Qual-1.1). To answer RQ-Qual-2, in our comparisons between groups of data analysts, we specifically focused on differences between groups of researchers and groups of students.

3.2.4 Survey Study (Simple Data)

3.2.4.1 Survey Procedure

We re-ran two surveys from Busse et al. on Amazon Mechanical Turk [74]. The first survey asked: “What are the top 3 pieces of advice you would give to a non-tech-savvy user to protect their security online?” The second survey asked: “What are the 3 most important things you do to protect your security online?” The difference between these two questions is not relevant to this paper but is stated for completeness.

3.2.4.2 Content-Study Analysis

Figure 3.1 shows the selection process for the data analyzed in content-study-simple.

We selected a random subset of submitted security advice. This included submissions we considered invalid, i.e., participants submitting random words or numbers, to give the students a realistic scenario. We divided this sample into three subsamples of 300 pieces of advice each.

Figure 3.3 shows an overview of the survey analysis process. Students again worked with self-

selected partners. Their first task was establishing a codebook with their partner, which was submitted (Submission s1). Students then independently coded their assigned advice with the shared codebook and were allowed to amend the codebook during their further analysis. In the end, they submitted both the coded advice as well as their final version of the codebook (Submission s2).

3.2.4.3 Meta-Level Analysis

R6 and R7 mixed deductive and inductive coding to analyze the final submitted codebooks, starting with the codes reported as the most frequent types of advice in prior work [74, 236] and adding additional codes, as they appeared in the students' codebooks. The four broad areas from previous work were *Account security*, *Mindfulness*, *Security software*, and *Updates*. Other advice from prior work was not listed under a specific category, so we subsumed it as *Technical*. All differences were resolved through discussion.

The results were compared for the final codebooks, which were the final outcome of the survey study. We compared appearance of concepts in codebooks using outputs of the code matrix browser supplied by MaxQDA. Similarly to the interviews, we also calculated IRR between each pair of researchers for their final coded submission.

Advice given by the current sample of MTurk participants was similar to the advice reported most frequently in prior work [74, 236]. Like in the replication by Busse et al. [74], some new advice appeared in our sample, while other advice which had been popular in the past, did not appear as often anymore. Since we focus on the meta-study, we do not report on details of the specific security advice.

3.2.5 Participants

There are two groups of participants in our empirical evaluation of the coding process. At the content level, participants were interviewed or answered the survey. At the meta-level, the students and researchers who analyzed the data, were also research participants, as we further analyzed their analysis process. As our focus is on the meta-studies, we only provide further information on participants of these studies (see Table 3.2) here.

3.2.5.1 Students

We recruited the data analyzing students from two iterations of a Bachelor's course on topics and research methods in usable security at a Central European university in the summer of 2021 and 2022. All students studied either IT security or computer science. They participated in the development of a data collection instrument and in analyzing the collected data. Our analysis will focus on the analysis process since we had the most control over this aspect of the collaboration. However, they were able to earn bonus points for their exam through their voluntary participation in each of these steps.

In 2021, 31 students signed up to analyze interviews, and in 2022, 40 students analyzed surveys. Six did not submit their work (2 analyzing interviews, 4 analyzing surveys).

We asked the 2022 cohort of students if they had prior experience in qualitative coding before taking the course. Three of 36 (8.3%) stated they did. Unfortunately, we did not ask this question in our 2021 cohort of students at the time. We contacted the students in 2022 but only got a response from 30% of them. These all stated to have no previous experience. We followed up with those students stating to have experience and asked them about the types of data they had analyzed, and the methodological frameworks they had used and to roughly estimate the amount of data/time spent in analysis. One had analyzed interview data but did not specify further details about their experience, one had spent about 2 hours using the General Inductive Approach [507] to code originally written material, like survey answers, and one had spent about 160 hours applying critical discourse analysis [145] to data from interviews and focus groups.

With few exceptions, the students did not have previous experience with qualitative analysis and we will refer to them as **S**tudents in the following (*Sc* for those analyzing the more complex data from the interviews and *Ss* for those analyzing more simple survey-data). The students worked in groups of two to three, and when referring to specific groups of students, we use abbreviations as follows: S[number]c/s. Numbering includes those groups which were excluded from the analysis. We append alphabetic characters to denote individual students within the groups, so the two students in group S2c would be referred to as S2c-A and S2c-B. Numbering starts afresh for the different data types.

3.2.5.2 Researchers

A total of 7 researchers participated in data analysis for this project, as shown in Table 3.1. We will refer to groups by adding an ampersand between the researchers' identifiers, e.g., R5&7 for the 2021 course instructors.

R5 and R7 coded the whole set of 18 interviews. Both had prepared course materials on qualitative analysis, both had been employed at a university in positions related to research for about four years and R7 had conducted and published a study using thematic analysis on roughly 15 hours of interview data. R5 was a student research assistant at the time of the study, while R7 was a PhD student.

We recruited R1 and R2 from our research group and R3 and R4 from a group at a different institution but working on similar topics. R1 - R4 analyzed a subset of the 18 interviews, like the students. We had intended for them to code the sample 3 interviews, but due to an error in the assignment process, they received the initial interviews of sample 3, but the additional four interviews of sample 2.

However, we did not notice differences in outcomes between the different subsamples of data, so this mishap did not further influence our analysis. These four researchers had participated in the analysis process of multiple qualitative and mixed-methods studies prior to this study, but all of them had done research for at least four studies using qualitative methods and using both interview and survey data, and R4 had additionally analyzed websites, videos and social media posts. R1 was a post-doc at the time of the study, while R2-R4 were PhD students in intermediate to advanced stages of their PhD.

Their epistemological outlook differs: R1 and R2 both have a positivistic viewpoint, and identify most with internal realism, while R3 and R4 both see themselves as social constructionists, whereby R3 leans towards relativism and R4 towards nominalism.

R6 and R7 analyzed a subset of the survey data. R6 was a student research assistant at the time of the study, had been employed in research-related positions for 2 years, and had participated in the analysis of roughly 6 hours of interview data, which resulted in a publication. Due to our findings from the Student-simple groups' analysis of survey data, and the low amount of differences both within and between the groups, no additional groups of researchers coded the data.

R5 and R6 were students at the time of the analysis, but more advanced than our student meta-level participants, and both had prior experience with research and had been employed at the university in research-related positions. In our meta-study-complex, we found the results of R5&7 to be closer to the researchers than the students, in terms of identified concepts and level of abstractness.

3.2.6 Ethics

We conducted this study in Europe and thus collected and stored data in accordance with the strict privacy regulations of the General Data Protection Regulation (GDPR). One of our institutions' Ethical Review Boards (ERBs) approved the meta-analysis.

3.3 Results

To investigate the similarity of outcomes within and between groups (RQ-Qual-1), we first compare the coding results from complex data achieved by different groups, as well as the coding results from the simple data. Then we trace the origin of concepts from the mindmaps, which are the results of the interview analysis back to the initial codebooks. Based on this we describe structural differences in the mindmaps and final codebooks. Finally, we share insights into the codebook merging process and compare the IRR achieved in different groups and analyzing different types of data. We specifically compare outcomes of student and researcher groups to cover RQ-Qual-2.

3.3.1 The Influence of Different Data Types and Researcher Experience on the Analysis Outcomes

To understand how the type of qualitative data and researcher experience influence the resulting codes, we compared the content and structure of codebooks from different pairs of researchers with different levels of experience (RQ-Qual-2) and between the interview and the survey study (RQ-Qual-1.2).

3.3.1.1 Comparing Outcomes for Complex Data

The final submitted outcome in the interview study was the mindmap depicting the most important concepts the students and researchers identified. We compared the concepts appearing in the 16 different groups' mindmaps to examine mindmap similarity. In total, we found 4 main concepts, 3 relationships between the main concepts, and 27 lower-level concepts, some of which appeared in association with multiple main concepts. First, we examined main concepts and noted if a group's mindmap either directly contained information about this concept or concepts related to it.

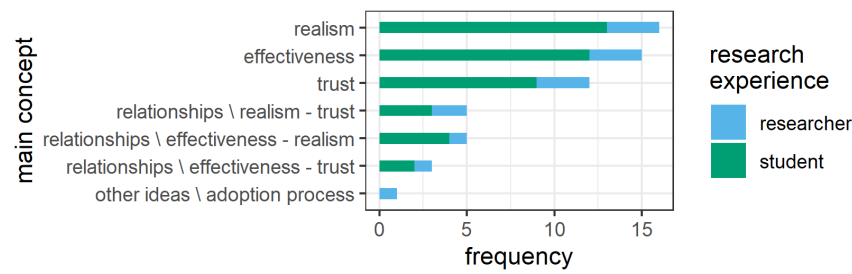


Figure 3.4 – Frequency of main concepts in mindmaps based on complex data (N=16)

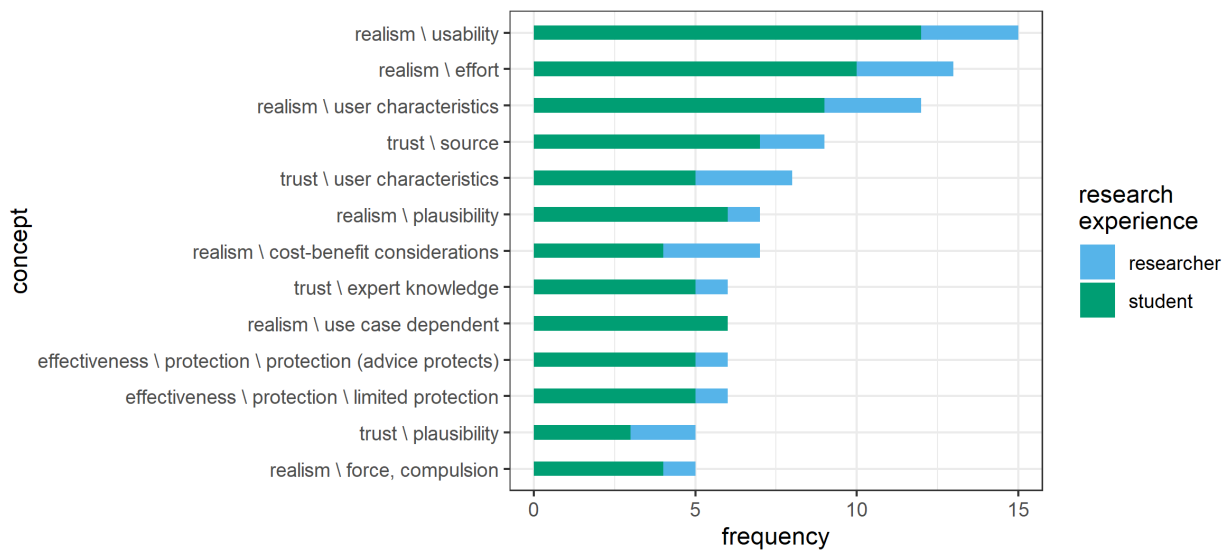


Figure 3.5 – Frequency of concepts in interview-based mindmaps (only showing concepts that were present in the mindmap of at least 5 of the 16 groups)

Figure 3.4 shows that some relationship to the main concepts from the research question, i.e. realism, effectiveness, and trust was present in all the researchers' mindmaps, and in most of the students' mindmaps. The concept of *trust* appears slightly less often than realism and effectiveness, perhaps since it was the focus of the second, rather than the first of the two content-study research questions. For an example of a mindmap with some mention of concepts related to effectiveness and realism, but not trust, see Figure 3.6a. R3&4 introduced a new main concept that had not been part of the content-study research questions, examining the adoption process of security advice, see Figure 3.6b. Relationships between the three main concepts as depicted in the mindmaps were present less common than the main concepts themselves in both the researchers' and students' mindmaps. In our annotated mindmaps in Figure 3.6, relationships are represented by an overlap of codes associated with different main concepts, or by explicit connections between such codes.

We also checked for lower-level of concepts, as displayed in Figure 3.5. Since there was a large amount of concepts, including 27 different unique concepts, and 38 different combinations of lower-level and main concepts, we only visualized concepts and combinations which were present in at least 5 of 16 mindmaps containing this concept, to get an overview. Many of those lower-level concepts

represented often are relatively concrete, such as *usability*, *effort*, or *user characteristics*. These were probably easier to identify and are thus more common across mindmaps of different groups. Concepts related to effectiveness are less common than those related to realism and trust. This mirrors observations when coding, that the interviewees had trouble understanding effectiveness and naming factors that would influence their judgment of effectiveness and the interviewing students had difficulties clarifying the question for their participant. So lack of clear understanding on both the interviewers' and the interviewees' side may have led to less common ground regarding effectiveness and consequently less representation of effectiveness in the mindmaps. As noted above, some concepts, like *plausibility*, or *user characteristics* appeared in relationship to different main concepts, so they are represented multiple times.

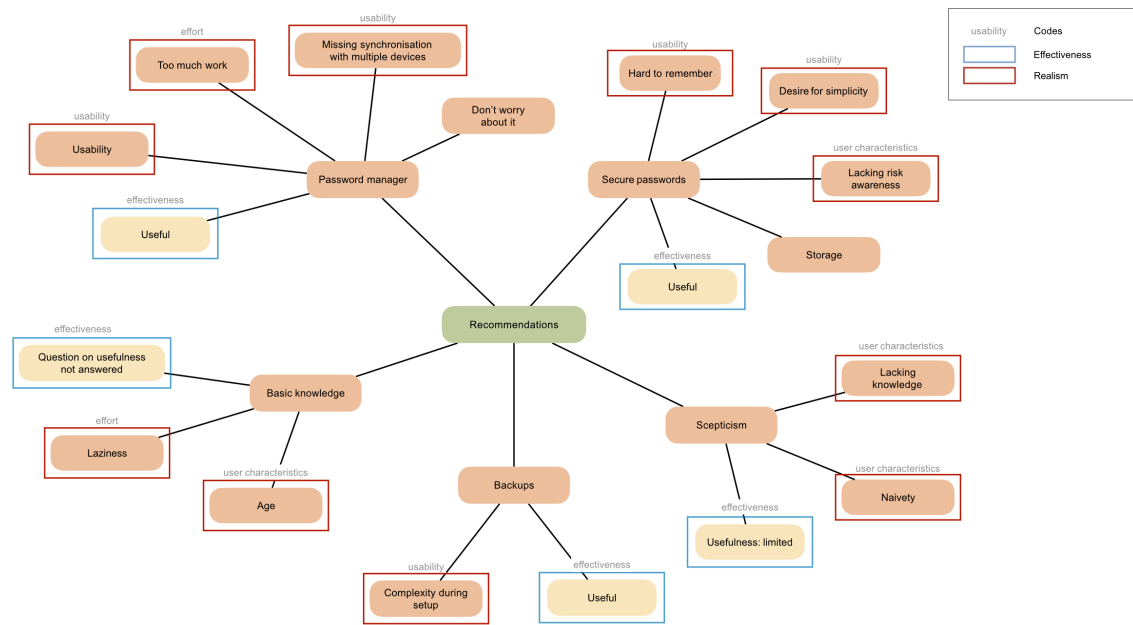
Comparing appearance of concepts for different levels of research expertise, Figure 3.5 suggests that while the top concepts are present in all or most of the mindmaps of the researchers, the proportion of students' mindmaps where this is the case declines more rapidly.

3.3.1.2 Comparing Outcomes for Simple Data

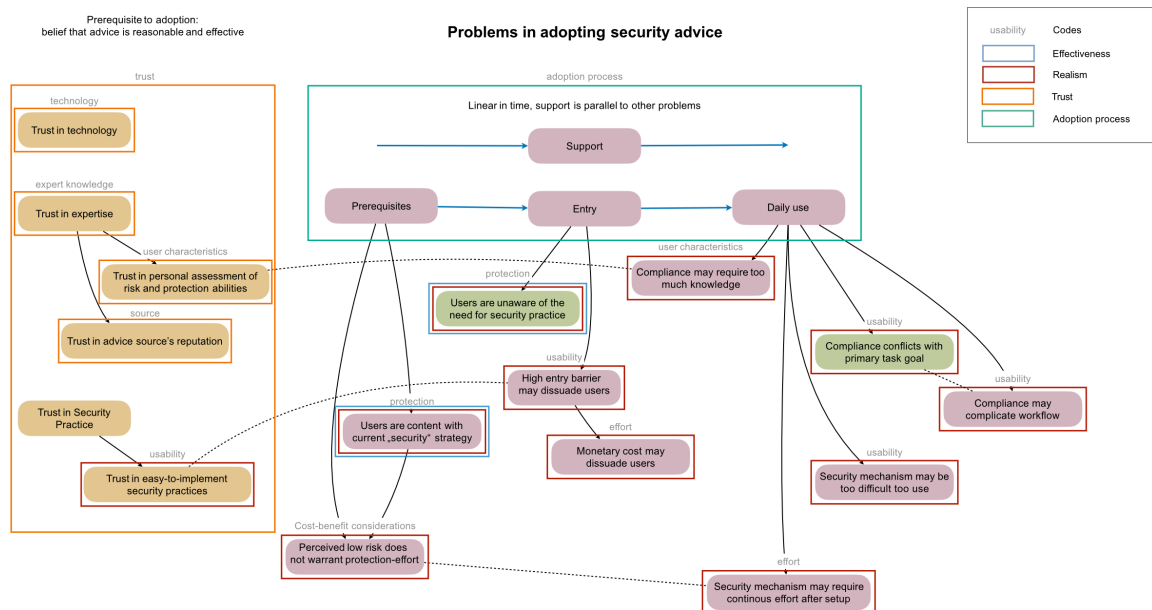
Similar to the interview study, we compared the final outcomes of each group's analysis process. For the survey, this was the final submitted codebook. Based on prior work [74, 236], we identified five different categories of advice and 71 individual pieces of advice throughout the 18 different groups. We identified 28 pieces of advice relating to *Mindfulness*, 18 for *Account security*, 9 each relating to *Security software* and *Technical*, 5 for *Updates* and 2 which did not fit any of these categories. An overview of the total amount of concepts identified in each group's codebook can be found in Figure 3.7.

All five categories of advice appeared in all final codebooks except for S13s and S2s, as shown in Figure 3.8. S2s's codebook was very brief (2 categories and 4 subcodes) and they misunderstood the assignment somewhat. They judged the effectiveness of the advice from their point of view, rather than identifying types of advice. Out of the five broad areas described above, their codebook only contained one identifiable code about *updates*. S13s's codebook was relatively brief as well and contained two very abstract codes which could not be clearly coded at all, as well as two types of advice which did not fit into any of the broad areas described above. Nevertheless, their codebook contained one code about *mindfulness*, one about *security software*, and one about *updates*. However, due to the lack of detail, we exclude these two groups from the rest of the analysis described in this section.

Figure 3.9 shows the frequency of occurrence of individual pieces of advice in the outcomes for the simple data. Like for the more complex data, there was a long tail of advice identified by only a few groups. Due to the larger amount of codes for this type of data, we adjusted our cut-off point to those concepts present in at least 10 groups, which were 15 of 71 total identifiable concepts. It is noticeable that even though *Mindfulness* boasts the largest number of advice, only a proportionally small amount (3 out of 28) of this advice appears in more than 10 outcomes. This reflects that there are many different ways mindfulness can manifest and that there were multiple similar, but not equal pieces of advice such as *Don't click links from unknown people* and *Be suspicious of links*, or *Be suspicious of e-mail*, *Don't enter passwords on links in Email* and *Don't open email attachments*, and the concept of being suspicious applied to different, but related instances, e.g. links, e-mails, downloads, or files. On the other hand, 5 of the 9 pieces of advice relating to *security software* are



(a) S14c's mindmap



(b) R3&4's mindmap

Figure 3.6 – Examples of submitted mindmaps, recreated and translated by the authors where applicable. The mindmaps themselves were submitted by our participants, where each group used their own color scheme. Colored rectangles show our coding. The name of our code is in gray above the rectangle, the color depicts the associated main concept.

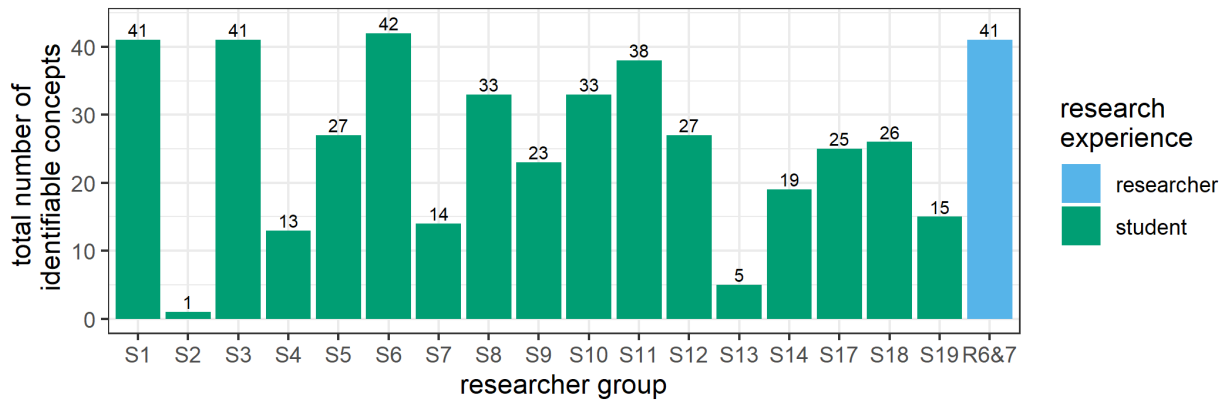


Figure 3.7 – Number of identifiable concepts for each of the submitted final codebooks for simple data

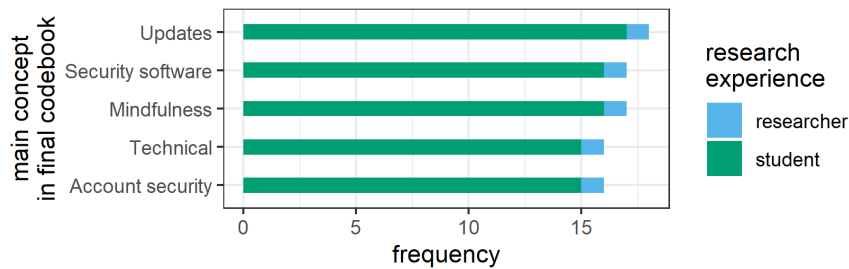


Figure 3.8 – Frequency of main concepts in 18 final codebooks based on simple data

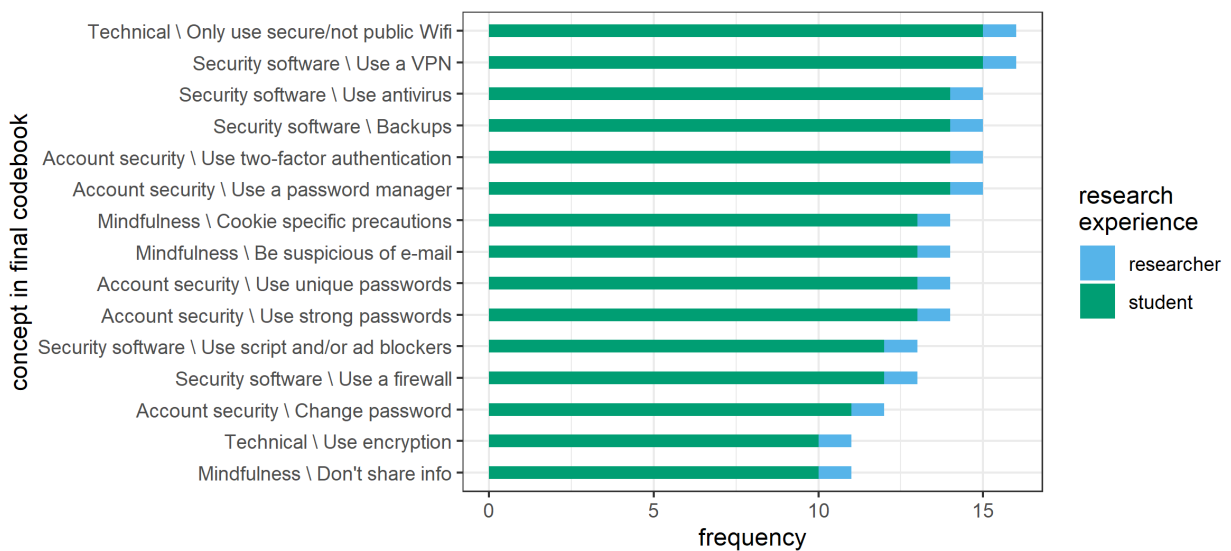


Figure 3.9 – Frequency of lower-level concepts in 16 final codebooks based on simple data

present in at least 10 outcomes. Security software is a more concrete topic than mindfulness, which may be the reason for the higher similarity in the results.

When comparing across the different subsamples of the data, some concepts do not appear in all, or only in one of the three different subsamples of survey data. For example, *Use biometrics* is not present in subsample 1, and *Use cloud storage* is not in subsample 2. On the other hand, *hide data* and *check logs / search for anomalies* are only present in subsample 3. These are instances where a concept simply did not appear in a subsample of data.

Likewise, some concepts which featured prominently in prior work and were thus included in the codebook from the beginning, did not appear at all. In some cases, e.g., *Use Linux*, this may have been caused by our participant sample, which only consisted of MTurk workers, and likely did not include experts like in prior work [74, 236]. Other reasons could include different coding practices, e.g. *Don't enter passwords on links in e-mails* was likely too specific and replaced by *Be suspicious of links*. Finally, with regard to e.g. *Turn on automatic updates*, a practice may have become so widely adopted that participants did not deem it worth mentioning.

3.3.1.3 Summary

In meta-study-complex, all three main concepts appeared across all the researchers' outcomes, but especially trust did not in many of the students' outcomes. There were only a few lower-level concepts, which a high number of groups had identified. The percentage of student groups identifying a concept declined faster than for the researchers, suggesting higher overall similarity among the researchers, than among the students (RQ-Qual-2).

In meta-study-simple, the five categories of advice were recognized by almost all different groups. Regarding specific pieces of advice, the length of the codebooks varies, and so does what each group considered relevant enough to be a code, leading to slightly different results for the groups. The necessary level of detail for the advice codes was not pre-specified and may have additionally depended on the amount of effort the students were willing to expend. However, given that the average amount of identifiable concepts in the analyzed groups' codebooks was 28.6 (SD=10.2), 15 specific pieces of advice identified by more than 10 of 16 groups suggests higher similarity for the coding of simple data than for complex data (RQ-Qual-1.2).

3.3.2 The Influence of Researcher Experience on Mindmap Structure in Complex Data Analysis

We examined several structural features of mindmaps: The level of abstraction of the mindmaps, presence of concrete security advice and how it was structured, if and how the research questions were present, the number of total concepts, the number of cross-relationships outside of a tree structure, and the degree of detail in including advice sources. We also noted some aspects related to both content and structure. These were the presence of trust as a concept in the mindmap, distinguishing between private and professional actions, between expert and non-expert opinions or reasons, and between positive or negative framing in relation to the content-level research questions (effectiveness, realism of and trust in security advice). We will only address these concepts briefly, and focus instead purely structural aspects in the mindmaps. There were no structural differences

immediately noticeable between the students analyzing different subsets of the data, except for the aspects related to distinguishing between private/professional, and expert/non-expert, both of which only appeared for subsample 1, as this differentiation was present specifically in the interviews in this subsample. Other variation in mindmap structure was not caused by a group analyzing a specific subsample.

In the following, we describe the mindmap structure of researchers and students. For two examples of mindmaps, from a researcher and a student group, see Figure 3.6. We use the terms small, medium, or large based on the observed spread of the specific feature. All researchers had a medium number of nodes/concepts in their mindmaps, although the number of relationships outside of a tree structure differed. R1&2 did not have any such relationships in their mindmap, while R5&7 had a medium amount and R3&4 had many such cross-relationships, see Figure 3.6b. Their mindmaps also reached a sufficient degree of abstraction, where it was possible to interpret and understand concepts in the mindmaps without having access to the interviews, and at the same time, the concepts were not too close to the actual interview content. Concrete instances of security advice were not present in the mindmaps. The main concepts from the research questions were at least to some extent present in the mindmaps, R1&2 named the research questions explicitly in their mindmap, R5&7 had the main concepts from the two content-study research questions as root categories, and for R3&4, not all research questions were thus represented, as trust showed up in codes, but not explicitly. Advice sources were not examined in detail in any of the mindmaps of the researchers. So, in summary, the structure of the researchers' mindmaps was relatively similar amongst each other.

Of the student groups, S7c did not submit a final codebook or a mindmap, and S11c did not hand in a mindmap, so we omit these two groups from this analysis. There was more variation in structure between the 13 remaining student groups. Most of the students, like the researchers, had a medium amount of nodes/concepts in their mindmap, but two groups also had a high amount, and three had a low amount of nodes. Most student groups did not diverge from a tree structure in their mindmaps, while one group had a few cross-relationships (S10c), one had a medium amount (S6c) and two had many (S2c and S3c). The degree of abstraction varied as well, nine of the groups reached a level of abstraction that could be interpreted similarly as with the researchers. Three had mindmaps, where the concepts were not abstract at all and thus very close to the interviews, and one mindmap was so abstract, that it was not interpretable at all without having knowledge of the interviews (S12c). A big difference to the researchers was the presence of concrete security advice in the mindmaps. Only two groups did not have security advice at all in their mindmaps, while eight had a node for security advice, which was linked to different examples of advice, in seven cases. Two, including S14c, as shown in Figure 3.6a, had security advice as categories with subnodes, and one group categorized the security advice with respect to aspects of the research questions. This behavior is also an aspect of abstraction since even though the interviews were about concrete instances of security advice, the research questions were about effectiveness and realism instead of the advice itself. The main concepts of these research questions were not at all present in the mindmaps of five student groups, including S14c, two groups had the research question concepts as main nodes, while four more groups represented some, but not all of the research questions this way. Two groups explicitly named the research questions in their mindmaps. Advice sources did not show up in much detail for most of the student groups, but one had details related to trust, and two only had concrete sources in their mindmap but did not draw more abstract conclusions. In summary, there was more variation between the different student groups.

In partial answer to RQ-Qual-2, comparing different levels of experience, the amount of abstraction was generally lower in student researchers' mindmaps. The most obvious sign of this is the presence of concrete advice in the mindmaps, as described above. Additionally, students' mindmaps frequently featured utterances from interviews or references to the interview guideline, such as *question on usefulness not answered* in S14c's mindmap in Figure 3.6a. While content-wise, their mindmaps contained concepts related to the three main concepts, such as S14c's mindmap to *realism* and *effectiveness*, they often did not name these concepts or explicitly connect them to the research questions, thus staying on a less abstract level of sense-making.

This also becomes visible, when examining the more content-related aspects of structure. All three researcher groups did not distinguish between any of the three dichotomies private/professional actions, between expert/non-expert opinions or reasons, or between positive/negative framing of advice. However, five of the student groups distinguished between positive and negative framing of the research questions, e.g. between *not effective* and *effective*. Additionally, within the analysis of subsample A, two groups distinguished between private/professional, and one group between expert/end-user.

3.3.3 The Influence of Multiple Coders on the Qualitative Analysis Process for Complex Data

To further understand how multiple coders within a group influence the qualitative analysis process (RQ-Qual-1.1) for complex data, we compared the codebooks between different stages of the analysis process.

3.3.3.1 Development of Codebooks

We examined content differences between the initial codebooks, where meta-level participants worked alone, and the final outcomes, the mindmaps, in addition, we specifically describe the merge process between different codebooks. In the mindmaps of the different groups, we found between 2 and 19 different concepts. We coded concepts in relation to the three main concepts from the content-level research questions: effectiveness of, realism of, and trust in security advice. Due to this process, relationships between the three main concepts are also coded as concepts. Through coding the initial codebooks with the same codebook of concepts, we investigated the origin of different concepts visible in the final mindmaps.

As can be seen in Figure 3.10, for all of the groups, except one, some concepts in the mindmap had not been present in either of the initial codebooks from the group. S12c's mindmap was very sparse and only contained two identifiable concepts. Likewise, the three students in this group also had short initial codebooks. The abstract concepts present in the mindmap had been present in all their initial codebooks. For the other groups, the proportion of new codes, which had not been present in any of the mindmaps, varies between 5% and 67% ($M=35\%$, $SD=21\%$). The amount of concepts which had previously been present in multiple (in most cases: both) of the initial codebooks varies between 6% and 100% and was on average lower ($M=28\%$, $SD=26\%$) than the proportion of new concepts. Some concepts had only been present in one of the initial codebooks, so these concepts were introduced into that group's outcome by that coder. It is noticeable that for many student

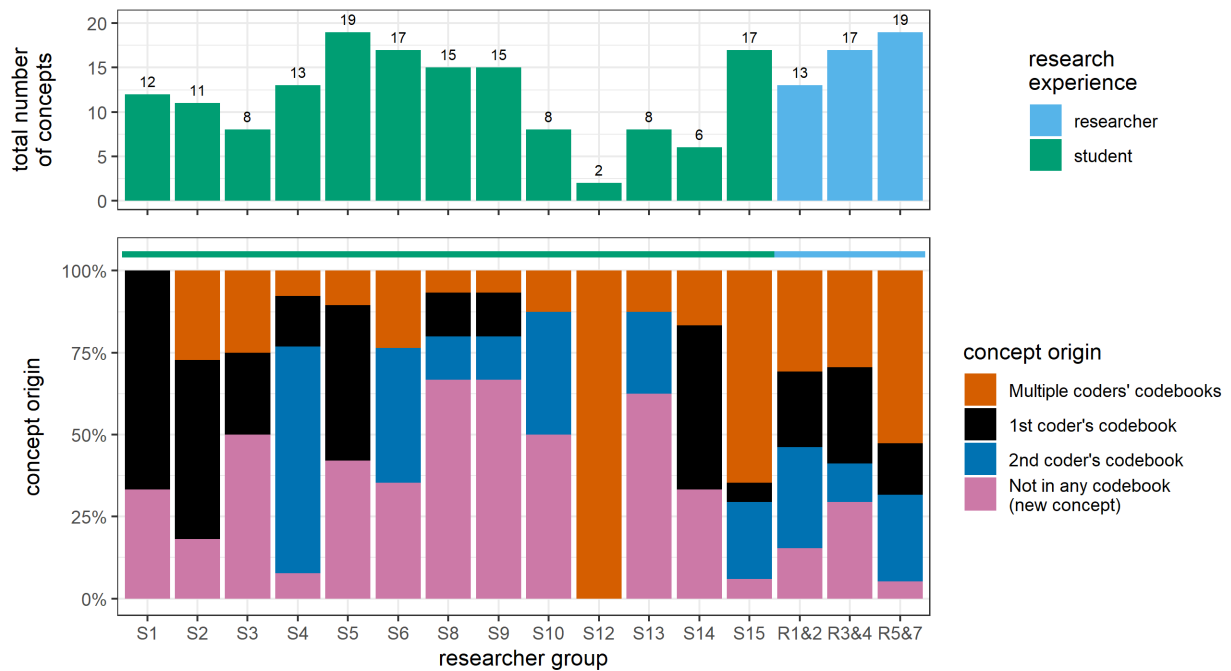


Figure 3.10 – Overview of concept origin: Upper part of the graph shows the total number of concepts present in the mindmaps, while the lower part of the graph shows the proportion of different origins for the concepts in the mindmap.

groups, the amount of concepts introduced is quite different for the different partners, e.g., in S1c, S3c, and S6c, one coder did not introduce any concepts of their own that ended up in the mindmap, and for S4c most of the total amount of concepts came from the second coder. For S1c, it is even the case that all concepts in the mindmap either stem from only one of the coders or were introduced after the coders had started working together. In contrast, for other groups, such as S8c, S9c, and R1&2, the proportions of concepts originating from one or the other group member were relatively equal. In other cases, both group members contributed, but not quite in equal measures, such as in the remaining two researcher groups, R3&4 and R5&7.

In summary, this shows that multiple coders working together had an impact on the content of the final outcomes, i.e., the mindmaps. In partial answer to RQ-Qual-1.1, multiple coders were beneficial in analyzing complex data.

3.3.3.2 Codebook Merge Process

When delving deeper into the codebook merge process in investigating the influence of multiple coders on results from qualitative analysis (RQ-Qual-1), we found four main strategies when merging individual (initial) codebooks into a shared (merged) codebook. They can roughly be summarized as adding concepts, removing concepts, rearranging concepts, and changing concepts.

In some cases the merge process led to completely new codes or concepts appearing in the shared codebook, which had not been in any of the individual codebooks prior to the merge. For the most part, only a few new codes appeared in shared codebooks, e.g. S3c added a new upper-level

code *motives* for codes referring to reasons to follow security advice, and S2c added a lower-level code *criteria for use*. However, R1&2 added multiple upper-level concepts / groupings in their codebook, such as *technical hurdle* and *Necessity* (both as subcodes to *effective, but not realistic*). In addition, they added new lower level codes, for example *personal experience* as a subcode to *characteristics* [of users] and *configuration necessary* as a subcode to *technical hurdles*.

On the other hand, codes and concepts were also dropped in the merge process. In S5c, while S5c-A's codebook spanned over 7 pages, and S5c-B's codebook had 2 pages, the shared codebook had 3, so quite a number of codes were dropped during the merge. In many cases, subcodes were dropped, e.g. multiple subcodes for *usability* from S5c-A's codebook, including *personal failures*, *cumbersome menus and services*, and *insufficient help*.

When rearranging concepts, the content of concepts appearing in the shared codebook stayed the same, but order or level in the hierarchy changed, such that what had been a low-level code moved into a higher level of the codebook hierarchy or vice versa. S5c-A had first used individual categories for the two interviews, leading to a lot of repetition of subcodes, such as *usability*, *obligation* or *personal experience*. In the shared codebook (S5c), these concepts then became upper-level codes. Lower-level codes were also moved around staying on their own level of the hierarchy, but categorized under a different category, for example *users are lazy* and *assessment usability vs. security* (S13c), were moved from *effectiveness of measures* in S13c-A's codebook to *practicability* in the shared codebook.

Finally, some concepts were also changed in the merge process. An example of this is in S8c, where two codes (*for companies, less increase in security* and *for end users, huge increase in security*) from S8c-A's individual codebook were merged into one code (*Protection depends on prior situation*) in the shared codebook. A different type is unification, where codes from multiple of the individual codebooks result in one new code in the shared codebook, which differs in phrasing, but not content from all the previous codes. This strategy seems to be used to unify phrasing, e.g. the code *factors for a successful security measure* was derived from S12c-A's *aspects for successful protection* and S12c-C's *factors*. One code from an individual codebook can also be split up into multiple codes in the shared codebook, e.g. *Mobile*, *Tablet*, *Laptop*, a subcode of *technical affinity* in S14c-A's codebook, was split up into single codes for each of the devices, on a higher level in the code hierarchy of S14c's shared codebook.

3.3.4 The Influence of Different Data Types and Researcher Experience on Calculated Inter-rater Reliability

When investigating agreement for complex data, we use MaxQDA's measure of 85% overlap constituting agreement. We investigated agreement in different subsets of the coded interviews in the final submission, after all interviews had been coded independently with the shared codebook.

As can be seen in Figure 3.11, for most of the groups, agreement was higher for the initial interviews and lower for those coded after establishing a shared codebook, resulting in a medium level of IRR for all interviews taken together (see Figure 3.12). This was less pronounced for some of the meta-level participants, such as two of the researcher groups. A notable exception to this pattern was S6c, where agreement was lower for the initial interviews and higher in the second round of coding. High agreement in the initial interviews could be caused by recoding of the first two interviews already

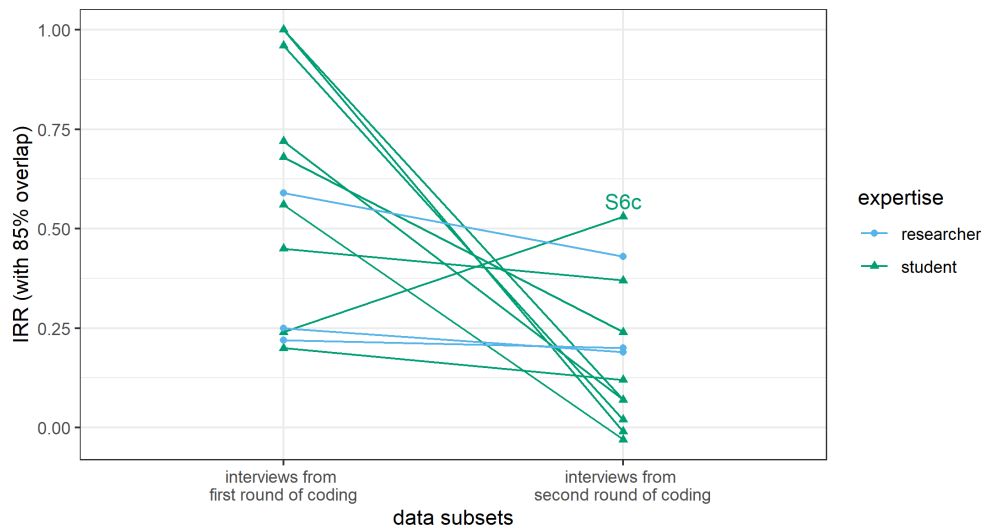


Figure 3.11 – IRR for different subsets of interviews

taking part jointly during the codebook merging process rather than individually afterward. For those groups with very high agreement, up to Cohen's Kappa κ of 1, coded segments were adopted as is for the first two interviews. For example, S4c generated a shared codebook by merging the two initial codebooks, and both group members took over those coded segments from each other that they had not included before. In S2c, adoption of coded segments took place one way: S2c-B adopted the exact coded segments from S2c-A's merged codebook for their final submission, although S2c-A updated their coding in the merged codebook to incorporate some structure from S2c-B's initial codebook.

Those groups that did recode individually might have profited from the shared codebook being based on those two interviews, which were then recoded, and for which higher levels of IRR were then reached. The coders in S6c stuck closer to their original coding for the recoding of the initial interviews with the shared codebook. Examples of this behavior included renamed codes used to recode a specific segment but not adding new ones, and S6c-A generally coding smaller segments, compared to S6c-B coding larger segments of text. For the remaining four interviews, they did not have prior coding to be attached to, and both coded a combination of smaller, more focused segments and very large segments of text for broad concepts, such as text referring to a specific advice, and thus reached a higher IRR.

When calculating IRR for the simple data, there were no segmentation issues since each given advice was coded as a whole. When comparing the achieved IRR for the simple data to the complex data, IRR was notably higher for the survey data (range:0.30 to 0.97; median=0.70) than for the interview data as a whole (range:0.16 to 0.66; median=0.41), see Figure 3.12. This difference becomes even more pronounced when comparing to the interviews coded in the second round of coding (see Figure 3.11). Regarding RQ-Qual-1.1, this shows that with complex data, there is more room for coders to see things differently. In contrast, the codes based on simple data show less variation between the coders.

On RQ-Qual-2, for the complex data, two researcher groups were relatively similar to each other regarding the achieved IRR, but one of them reached a higher level of agreement. The researchers'

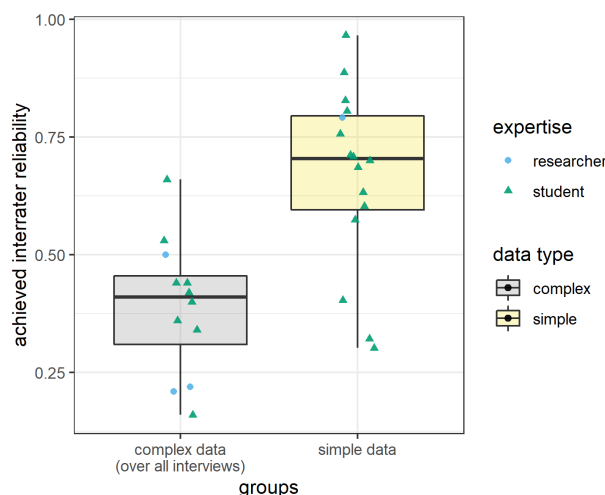


Figure 3.12 – IRR compared for different data types and experience levels

achieved IRR was within the range of the different student groups' IRR. For the second round of coding, where the influence of those groups who had probably worked together to recode the first two interviews was removed, R1&2 and R5&7 reached an IRR close to the median of all groups, while R3&4 had the second highest IRR. In total, the IRR of the researchers was in a similar range to the IRR of the students. Since the contents of their analysis outcomes differed, it may be sensible to include students in the coding process either when the codebook is already established or when working together with more experienced researchers.

3.3.5 Summary: RQ-Qual-1 and RQ-Qual-2

We found differences in outcome based on the type of analyzed data and the level of experience of the data analysts. For the complex data, outcomes on main concepts were similarly present for most groups. However, for lower-level concepts, there was a lot of variation. For the simple data, the main topics appeared in all relevant codebooks, and the clearer the definition of a concept or specific security advice was, the more similar the codebooks were on this topic. Overall outcomes between groups of data analysts were more similar when simple data was analyzed, than for complex data (RQ-Qual-1.2). Within the groups analyzing complex data, members contributed different concepts to the final outcomes, and some concepts appeared after interaction between coders. In general IRR was higher for simple data than for complex data (RQ-Qual-1.1).

Regarding RQ-Qual-2, there were differences in the outcomes of researchers (higher experience) and students (lower experience). Researchers reached a higher level of abstractness in their outcomes than students. While the students' outcomes were not as developed as those of the researchers, their direction of analysis was more similar to those of the researchers from their own institution (R1&2 and R5&7). Content-wise, differences regarding the complex data were somewhat more pronounced between R3&4 and the other two researcher groups, R1&2 and R5&7. While there was variation in the outcomes of content-study-simple, there was no clear difference between R6&7 and the student groups as a whole.

3.4 Survey of SOUPS program committee members

Due to our experiences of reviewer discussions at SOUPS and CHI, we wanted to get a more complete overview of current reviewing practices regarding qualitative analysis. To achieve this, we surveyed researchers who had been members of the SOUPS PC in the last five years (2018 - 2022) about their research background, experience, and the criteria they apply when judging qualitative research for this venue. We chose SOUPS since qualitative USP research is regularly published at this venue, and it is at the intersection of a highly quantitative research domain (security) and the more user-centered HCI community. Additionally, SOUPS has a close-knit community, with which some of the authors are familiar, which made it more likely for PC members to respond to the survey. The survey focused primarily on criteria as applied to the coding process.

We asked participants to judge criteria as mentioned by McDonald et al. [327] and additionally asked questions relating to our empirical evaluation. Since PC members have busy schedules, we kept the survey short and mostly relied on closed questions while still offering the possibility to specify detailed answers if desired.

We conducted two pilot tests using the think-aloud method [43, p. 169f]: One with a researcher who had previously reviewed for SOUPS but had not been a PC member, and the second with a PC member of the Conference on Human Information Interaction and Retrieval (CHIIR), a different conference of a similar size, which is also situated at the intersection of two larger research domains, HCI and information retrieval. We clarified some question phrasings and shortened the survey considerably after the pilot. Two further pilot testers, who were part of our intended sample, completed the survey to test the required time. They took 7.27 and 7.6 minutes to complete the survey. Since this time was within the time frame of our intended duration, we did not change the content of the survey but only changed the formatting of a question at the request of one of the pilot testers. We retained the data of these final two pilot testers. We include the final version of the survey in the supplemental material.

3.4.1 Recruitment and Participants

We contacted all 85 SOUPS PC members of the years 2018 to 2022 via e-mail through the account of a senior researcher who had been active in the community for a decade. 6 of our e-mails bounced, and 3 prompted an *out of office* notice as a response. We received 37 responses (response rate: 47%). They took a median time of 8.9 minutes (min: 2.4 minutes, max: 72.8 hours) to complete the survey.

3 of the participants identified as mostly qualitative researchers, 5 as mostly quantitative researchers, and the rest (29) used both qualitative and quantitative methods. Overall, experience levels as PC members were evenly distributed, with 12 having been on the SOUPS PC one or two times, 11 three or four times, and 14 five or more times. We raffled three 100 € gift certificates of the participants' choosing among participants who wished to join the raffle.

3.4.2 Ethics

Like for our empirical evaluation of the coding process, we collected and stored data in accordance with the strict privacy regulations of the GDPR. We did not collect any personal data in the main

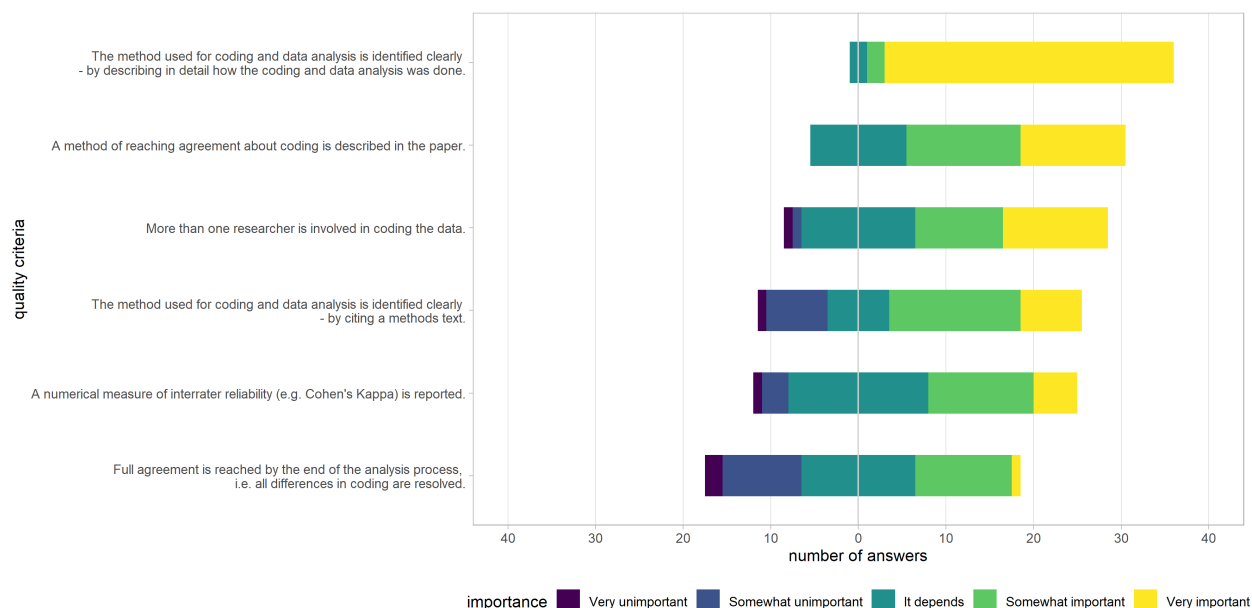


Figure 3.13 – PC members’ importance judgments for given quality criteria, There was one response of “I don’t know” each for “agreement described” and “full agreement reached”, which were excluded from this visualization

survey and gathered e-mail addresses for inclusion in the raffle through a separate survey to avoid participant identification. One of our institutions’ ERBs approved the PC member survey.

3.4.3 Data Analysis

Two authors used the General Inductive Approach [507] to jointly analyze responses to open-ended questions, i.e., additional input about “It depends” answers, to find influencing factors on the quality criteria. Participant quotes from PC members are identified with **PC**<number>.

3.4.4 Important Quality Criteria

We asked PC members to judge the importance of different criteria in their reviewing process. Figure 3.13 shows that there is agreement among our participants that some method of reaching agreement should be described in the paper and that a detailed description of the method used for coding and data analysis is necessary. For other criteria, such as whether full agreement should be reached by the end of the analysis process or whether the data analysis method should be identified by citing a methods text, reviewers’ opinions are split. In general, all of these criteria are also applied by many PC members based on other contextual factors, except perhaps the method description, where this was only the case for two participants. They may consider the type of study, e.g., the research questions, the type of data analyzed, and the type of analysis conducted. They refer to method-specific recommendations for specific research paradigms, e.g., for the reporting of numerical IRR PC29 states “For content analysis it is important, for other

types of coding, not so much”, and regarding the number of coders, PC26 recommends: “For thematic analysis only one coder is adequate” and PC36 states “[...] someone doing true Grounded Theory where data collection and analysis are intertwined probably wouldn’t have a second coder”. Another consideration is how the coding results are used, e.g., whether for “performing quantitative analysis on the codes” (PC26) or whether “the data is analyzed for *theory*” (PC13). Even if PC members have a preference for a particular practice, e.g., reaching total agreement, and if otherwise would expect justification for the deviation from their personal standards, i.e. “would expect the paper to make an argument about why it’s not needed here” (PC9), they realize that there are exceptions to their applied criteria: “I would say that by default/in most cases I expect more than one coder, but there are sometimes good and appropriate reasons for just one” (PC9). Regarding the criterion of citing a method-text for the data analysis method used, novelty was frequently a factor, although in different ways, e.g., “If it’s entirely new to the community, that kind of citation is important” (PC15), or that “sometimes researchers come up with a refined / new method” (PC34) where a citation is not possible. There was also some criticism when citing a methods text is used as a substitute for a thorough description of the analysis method, as methods texts may be “only loosely aligned with how the research appears to have been conducted” (PC31).

PC members named additional quality criteria which they apply. Many stressed the importance of providing a detailed description of “the process in enough detail for the reviewer to understand and evaluate what was done” (PC27). This can include choice of sample size and participant recruitment process, “how the codebook was developed” (PC28, PC30), and the inclusion of “good quotes that support[...] the themes and assumptions draw[n] from the process” (PC15). The description can be enhanced by providing supplementary material, like interview scripts or survey question phrasing. The methodology should not only be described but justified: “authors [should] argue for what they did and why they think it’s appropriate” (PC7). The strength of the claims made and their appropriateness given the outcomes are also used to judge qualitative submissions, although this criterion may also be applicable for quantitative data. Finally, PC3 mentioned appropriateness of research goals: “Qualitative analysis is best when it tries to find interesting narratives and patterns of thoughts and behaviors and not to report on a phenomenon that can be generalized to the general population”.

3.4.5 Acceptable Coding Practices

We specifically investigated acceptable coding practices for simple and complex data, which corresponded to the data analyzed in our evaluation of the coding process. One scenario was similar to our survey study and presented as “study where short, textual survey responses are coded”. We explicitly mentioned Ion et al.’s work [236] as an example. The other scenario was similar to our interview study and presented as “a study that analyzes complex answers such as from interviews”. We did not want to bias the participants by pre-labeling these scenarios as simple and complex. Participants could select all coding practices that they considered acceptable separately for complex and simple data.

Contrary to our expectations based on McDonald et al.’s quality criteria [327], we don’t see any large differences between complex and simple data, e.g., Figure 3.14 shows that using only one coder is largely seen as unacceptable even for simple data. Interestingly, despite the two coder variants being mostly seen as acceptable, all three variants have a fair number of PC members who

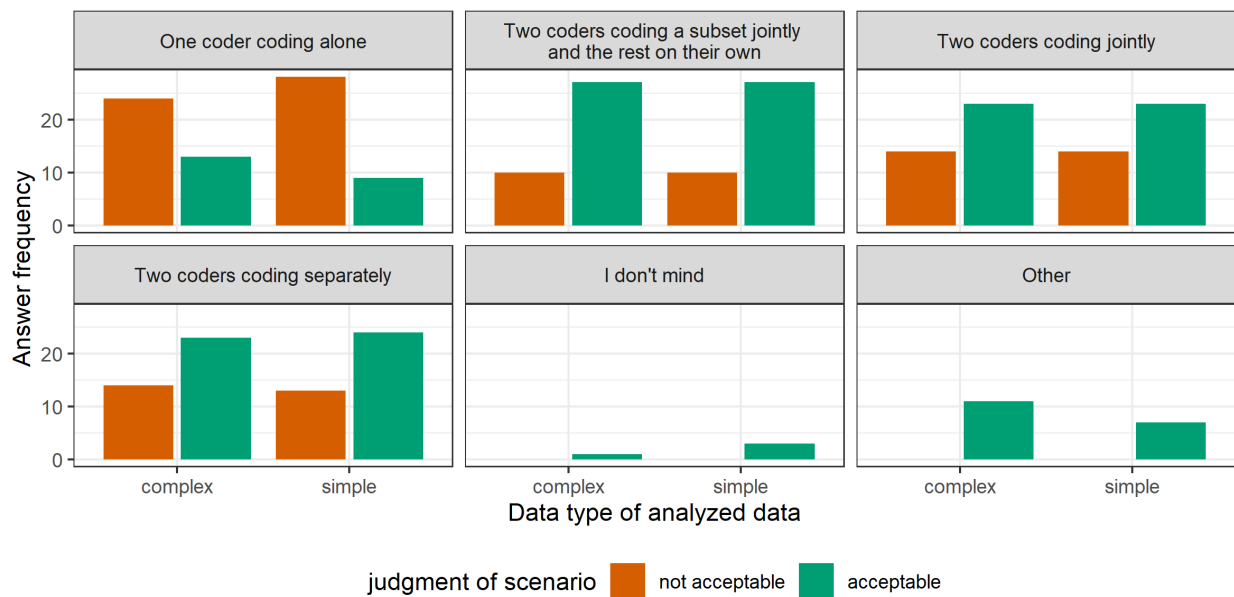


Figure 3.14 – PC members’ acceptance of different scenarios regarding the number of involved coders and the type of analyzed data. Non-answers for “left-over” categories of *I don’t mind* and *other* are not graphed since they don’t carry information.

do not approve of that variant. Note that the scenario “two coders coding a subset jointly, and the rest on their own”, which is acceptable to the largest amount of PC members in our survey, is also closest to the coding process in our evaluation. When analyzing the data from content-study-simple, coders jointly developed a codebook and then independently applied it, while for the complex data from content-study-complex, coders started working independently but consolidated a codebook jointly before applying this shared codebook independently. A few participants indicated that they don’t care which process is used. Participants were also able to specify other coding procedures that they considered acceptable and named using more than two coders and a grounded theory-specific process involving “iteratively cod[ing] and consolidat[ing] to a set of themes”(PC8). Other participants gave similar reasons for accepting a coding process, as the ones we described in Subsection 3.4.4, such as sufficient justification for the process, how the coding results are used, and the type of analysis conducted.

3.4.6 Acceptable Levels of Calculated Inter-Coder Agreement

We asked for the PC members’ minimum acceptable level of agreement and gave both word judgments and the corresponding values for Cohen’s Kappa according to Landis and Koch [295]. Figure 3.15 shows that of those who expect reporting of a numerical agreement value, most require substantial agreement. Lower than moderate values of numerical agreement were not accepted by any of our participants, although some stated that they do not care about the exact number or don’t require any numerical agreement to be reported. However, a large number of respondents stated that the level of agreement they require depends on different factors, such as the type of study, type of analyzed data (e.g., more structured data needs higher agreement (PC37)), type of

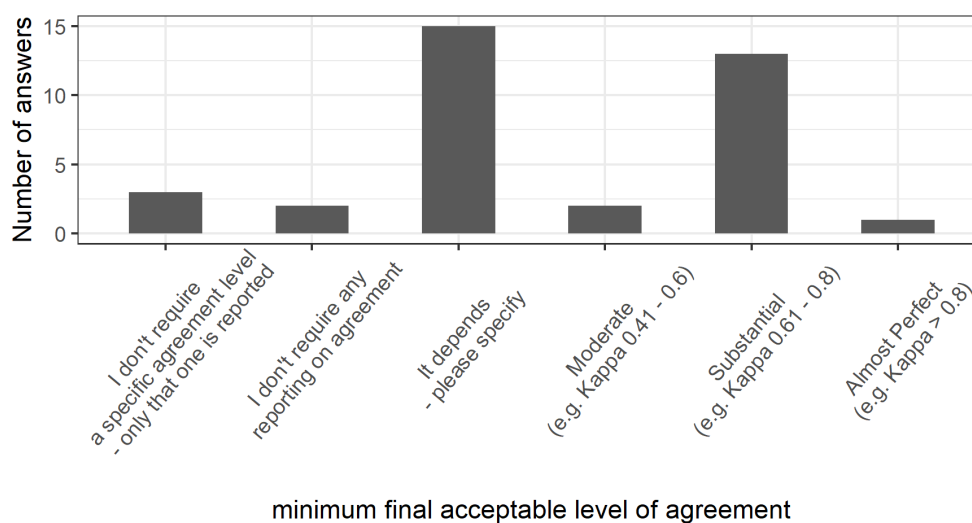


Figure 3.15 – Minimum acceptable levels of numerical agreement

analysis conducted, or the claims which the analysis tries to corroborate.

About half (18) of our participants support reporting this numerical IRR after the coding is complete, 9 think it is appropriate to report IRR after the codebook was established, 7 think IRR should be reported after each step in the analysis which involves independent coding, and 7 state that there is no need to report IRR. Six participants gave criteria on which the point of reporting IRR could depend, such as the method used (PC10, PC22), claims made in the paper (PC13, PC31), the weight of phases of the analysis for the outcome (PC14), or the purpose of reporting IRR, e.g. to “test coding reliability” after coding or “checking differences between judges” (PC24). PC25 stated that “agreement can change, it is important to discuss at each stage and recode and recalculate if necessary”. Three PC members gave the caveat that reporting IRR at all is not always necessary.

3.5 Limitations

In our empirical evaluation of qualitative analysis, we only investigated the coding stage of qualitative research, although other stages may also be influenced by the involvement of different researchers. Reproducibility, especially with less standardized interviews, may also be influenced in the interviewing stage of research and prior work has found interviewer characteristics to differ, even when following the same guidelines and being trained in interviewing the same way [388]. However, interviewing is also influenced by interviewee characteristics, and interviewing the same interviewee multiple times by different interviewers is not feasible due to fatigue and learning carrying over to subsequent interviews. Since coding is a central part of many different approaches to qualitative analysis [564], used to generate insight from raw or transcribed qualitative data, we focus on this part.

There are some limitations with respect to the different participants in our empirical evaluation: On the content level of data analysis, our interview participants were a convenience sample, but since

we were largely interested in the coding process applied to data generated from these interviews, this was not as grave. While other work has shown that student participants are similar to more professional participants concerning the direction of effects in secure programming tasks [342], this had not been investigated for qualitative analysis. For this reason, we compared our student meta-level participants with researchers, and while some aspects, especially the degree of abstractness in the outcomes were different, other aspects were similar. Certainly, a first-year PhD student may not have experience in qualitative analysis either, so it is important to examine the characteristics of less experienced researchers' coding process to be able to support them better. However, we were not able to collect data on prior experience with qualitative analysis for all the participants in meta-study-complex, and as such our findings regarding the influence of experience are tentative and should be explored further.

Regarding our researcher participants, R5-R7 were researchers, as well as participants of the meta-level studies which we report on in this work, meaning that they analyzed not only other groups' coding outcomes but also their own. To be able to interpret analysis outcomes, it was necessary for them to have sufficient in-depth knowledge of the analyzed data and research questions, which they gained by participating in the analysis of content-level data. However, their own work may have subconsciously presented the reference frame to which other outcomes were compared in the meta-level studies. We tried to mitigate this by taking breaks of two months (for meta-study-simple) and six months (for meta-study-complex), between the content-level analysis and the meta-level analysis. We also sought an outside reference frame by incorporating the advice and advice categories from prior work for meta-study-simple, and by specifying concepts taken from the research questions as main concepts for meta-study-complex. Furthermore, we did not pre-specify expectations, e.g. of the number of recognized concepts, or regarding structural elements of the mindmaps, but rather assessed them by comparing to the variance in observation among all groups' outcomes.

Using the full set of interviews and survey responses was not feasible due to the large workload it would have imposed both for the students, who also partook in other tasks during the course, as well as the more experienced researchers, whose time is generally scarce. We accounted for this possible bias from the concrete instances of data used in the analysis, by sampling groups of interviews and survey responses, while balancing the interviews for length to ensure a similar workload across all three groups of data. We did not notice any important differences between the groups assigned to different samples, neither for the complex interview data, nor for the simple survey data.

For our meta-analysis, coding the mind maps and codebooks of other researchers, in most cases without memos present, meant interpreting what the researchers aimed to express through a, often short, code. It is unclear whether the connections we drew were always intended by our meta-level participants, and we may have over-interpreted connections, due to being more familiar with the data. We assume that the researchers and students meant a connection to the research questions, when we are able to see it. This makes our reported levels of content presence in the initial codebooks and the mind maps an upper bound, as some of the meta-level participants might not have intended a connection to the concepts of trust, realism, or effectiveness where we saw it.

The most important limitation of the PC member survey was its brevity and the lack of more open-ended questions, even though expert reviewers' perceptions of quality criteria for qualitative work in the domain of USP have previously not been investigated much. However, it was our utmost priority to keep the survey as short as possible to motivate busy PC members to participate. For

this reason, we limited our questions to those most relevant to the topic of this paper: qualitative coding, and the measurement of agreement, and defer further topics to be investigated in future work.

3.6 Discussion

We empirically investigated the coding process of coders with two different levels of experience (students and researchers) and two types of data (simple to code survey data and more complex interview data), and found differences in outcomes. For the complex data, while the main concepts, which were predetermined through the given research questions were present in nearly all outcomes, for lower-level concepts there was a lot of variation. In our analysis of the origin of concepts, we also noted concepts being introduced into the codebook from both coders. For the simple data, the main topics appeared in all relevant codebooks, and the clearer the definition of a concept or security advice was, the more similar the codebooks were on this topic.

Concerning the experience of the coders, researchers had more abstract codebooks and mind maps than students. However, we note that content-wise, differences regarding the complex data were somewhat more pronounced between R3&4 and the other two researcher groups R1&2 and R5&7, even though R5&7 can be considered junior researchers, as they do not have as much experience as the other two researcher groups. R3&4 introduced an entirely new main concept in their outcome, which was not present in any of the other student and researcher groups, which were more similar in that regard. We hypothesize that this may be due to R3&4 working at a different institution than the other meta-level participants, and they may thus have a different background of research experience. Their differing epistemological outlook to other researchers also supports this interpretation, as prior work has shown epistemology to influence results in qualitative analysis [163]. The influence of researcher characteristics on analysis results is not only present in qualitative analysis but can also apply to quantitative methods. In a re-analysis of two studies by seven researchers each, already the pre-processing steps differed so much that no re-analysis used the same sample size [229]. The social sciences have used positionality statements to make prior knowledge, social and personal background, and other possible factors that may influence researchers' data analysis process clear to readers [111]. In some instances HCI [505] and USP [561] have adopted this practice, although, in a literature review focused on privacy and marginalized communities, positionality statements were not reported in any of the studies published in privacy-focused venues, which included SOUPS [444]. We checked USP papers from SOUPS and CHI 2022 and found 49 papers using qualitative analysis methods, but only 3 of those papers contained positionality statements. Our data suggest it would be a good idea to further adopt this practice in USP research where appropriate, describing researchers' familiarity with the methods they are using.

Also based on our data, we suggest that for simple data and clear research questions, it is sufficient if a single coder codes data with an established codebook. This means that if a codebook is established together, the coding following it does not need to include multiple coders for simple data. Researcher time and effort should instead be invested in coding complex data, where multiple researchers can lead to different perspectives in the outcomes, and collaborations across different institutions and research focus can be a valuable contribution. Section 3.2.1 defines in more detail what we mean by simple and complex data in the scope of our study and recommendations.

For the complex data, both students and researchers differed in their interpretation of which lower-level concepts were the most important to represent, and R3&4 introduced an entirely new main concept, the adoption process of security advice, as taking part over time, in their outcome, which did not appear in any of the other groups' outcomes. This shows that different groups of researchers, especially from different research groups, can find different layers of meaning in complex data using the same method, like previously shown for using multiple qualitative analysis methods [163, 88]. Consequently, we suggest that replicating qualitative work even if it only consists of another set of researchers analyzing existing qualitative data sets could be worthwhile. We realize that such re-analyses carry the risk that not enough new insights or differences can be found to convince reviewers to accept a paper in highly competitive venues. Thus, there are clear disincentives to take such a risk in a "publish or perish" world. But what we would definitively encourage is that qualitative researchers make anonymous data available to open up the option. At the very least, we think this would greatly benefit teaching because students could re-analyze the data. Being able to analyze real data can motivate students, while the re-analysis of important data sets benefits the assessment of reproducibility for these findings, and thus this serves a double purpose if new results emerge from this form of crowd-analysis.

3.7 Conclusion and Future Work

We investigated quality criteria for qualitative research, specifically the qualitative coding process, both through an empirical lens and from the reviewer's perspective. We found more similarity for simple data than for complex data with more possibilities for interpretation. More research experience led to a more abstract representation of the findings from the coding process. Future work could examine further variations of when and how much interaction is beneficial and other factors influencing qualitative research outcomes in more detail, such as institutional, and thus educational background, and epistemological or ontological positionality, e.g., by extending this study to include students from R3&4's institution. We hope that our work may help qualitative researchers to make decisions regarding the planning of their coding processes and spur discussion in the USP community regarding the quality criteria we want to apply to qualitative research.

Chapter 4

Power Analysis and Effect Sizes in Developer-Centered Usable Security

Disclaimer:

The contents of this chapter were published as part of the publication “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” at SOUPS 2023. This paper is the result of a collaboration with my co-authors Christian Tiefenau and Matthew Smith, and various students and colleagues involved in the data collection. The research idea was conceptualized initially with Matthew and the idea evolved through project groups, labs and seminars with the students Ahmad Alaya, Somar Aljabr, Ahmad Assaf, Marwan Jaradat, Tansen Khan, Heng-yi Lin, Florin Martius, and Atacan Süder. I conducted an initial formal search for DCUS literature, which was expanded for recent literature from 2021 and 2022 by my colleagues at the time: Lisa Geierhaas, Eva Tiefenau, Maximilian Häring, Mischa Meier, Stephan Plöger, and Klaus Tulbure. I designed the data base for methodological information from DCUS papers. Christian designed and implemented a Django application for the data base. I adjusted the application where necessary. I trained and supervised Theo Raimbault and Simon Bong to extract data from DCUS papers and input it into the data base and I conducted a check of all extracted data. Christian and I extracted summary statistics from the data base. I conducted a meta-analysis of power of statistical tests. Matthew and I discussed key insights and strategies of the work. Finally, I prepared the paper for publication together with Christian and Matthew. Ahmad Assaf implemented a companion website for this publication under my supervision, and Radmehr Nouri Inanlou, Felix Jeske, Maximilian Mundt and Michelle Stivala adjusted the tutorial for power analysis to be more usable, also under my supervision.

4.1 Motivation

A priori power analysis can be used to calculate the necessary sample size for a study to detect an effect with a given probability, a given significance criterion, and a defined effect size [91]. The probability is often set to 80% and the significance criterion to 5% by convention and the effect size needs to be chosen by the researcher based on their research goals [91, 118]. Using a priori power

analysis, researchers can avoid running underpowered studies and missing effects that are actually present in the population and thus wasting resources or, worse yet, potentially publishing results that are misinterpreted as stating that there is no effect. It also prevents researchers from using more resources than necessary by running overpowered studies [136]. This is especially problematic for the sub-field of Developer-Centered Usable Security (DCUS) [185] since developers are often both hard and expensive to recruit. Running underpowered studies in this field is especially undesirable due to the large amount of effort coupled with low chances of finding the desired effects even if they are there. Despite this, power analysis is not common in the field of DCUS.

In the 54 DCUS papers we analyzed for this SoK, only 9,3% contained any form of power analysis. When the power of studies has been assessed in other fields in meta-analyses, these have frequently shown low power, such as in a review of ACM transactions [40] or in psychology [458]. The lack of a priori power analysis is one likely reason for this. Our analysis raises similar concerns of underpowered studies in DCUS. A potential reason for so few studies being planned with power analysis is that performing a power analysis is non-trivial and researchers must know, estimate or guess key population values such as standard deviations or proportions to calculate effect sizes or estimate the effect sizes themselves. This is especially tricky in the fields of USP and DCUS, due to both the heterogeneity of the populations studied, as well as the heterogeneity of variables being measured and the many non-standardized measurement instruments. On top of that, many tests are not published with enough statistical details to use them for estimating power for similar future studies [189].

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

4.2 Literature Collection

In the following, we describe the creation of our literature corpus for DCUS. As a catch-all, we will refer to this sort of literature as a developer paper in the remainder of this work. We define a developer paper as literature including a user study in some form, in which the participants

are software developers, software testers, administrators, other people responsible for planning, developing, testing, or managing software, or proxies for such people. An example of proxies would be computer science students, which are commonly used as a stand-in, e.g., for software developers [344]. We focus on the domain of usable security and privacy, which means that the studies should be focused on privacy or security problems and technology. We exclude any papers which do not include a study with actual users.

We started collection of literature in early 2021 and collected developer papers from four major conferences about security and privacy, which were published between 2010 and 2020: SOUPS, USENIX Security, S&P, CCS, and additionally ICSE and the USP tracks of CHI. Abstracts were used to determine whether a paper fits our definition of a developer paper, and in case of uncertainty, the method section of the paper was additionally used to clarify. We updated our literature basis in March 2022. Our final sample consists of 54 papers. Of those, 20 were published at SOUPS, 11 at CCS, 8 at ICSE, 7 at USENIX Security, 5 at S&P and 3 at CHI. The list can be found in Appendix C.1.

4.3 Systematizing Study and Statistical Test Data

Our goal was to collect and systematize information on studies and statistical tests from the domain of DCUS, focusing on what is necessary to conduct power analyses for user studies in this domain. We also wanted to add general information on the data collection process and the types of participants, since this might also be relevant when planning a new study. To aid researchers in planning new studies we created a data structure of our systematization and will offer this to the research community as a database. The database can be queried via a companion website for the relevant information to conduct power analysis¹. An excerpt of two entries can be found in table 4.2. Further entries are on the companion website. The entries are categorized to help researchers query the database and find similar studies, on which they can base their power calculation. For those cases where no directly similar previous study exists, we have created aggregated data based on our systematization that researchers can use as rough guides.

4.3.1 Systematization Process

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

In addition to the papers themselves, two sources further informed our structure: We made sure to represent information necessary to conduct power analysis, based on the G*Power-software [149], since G*Power is a commonly used and very powerful tool for power analysis.

To help guide our work, we created a set of hypothetical research questions for developer studies, for which we would want to run a priori power calculations, e.g., *Do Freelancers recruited from Freelancer.com and Upwork differ in their self-assessment of their reverse engineering skill and*

¹<https://powerdb.info>

# Type of IV	# IV	# IV Levels	# Type of DV	# DV Levels	Study design	Hypothesis test
Cat.	1	2	Cat.	2	B	Fisher's Exact Test, Chi-Squared Test
Cat.	1	2	Cat.	2	W	McNemar's Test
Cat.	1	2	Cont.	-	B	Independent t-test, Wilcoxon Ranksum Test
Cat.	1	2	Cont.	-	W	Paired t-test, Wilcoxon Signed Rank Test
Cat.	1 +	2 +	Cont.	-	B	ANOVA, Kruskal-Wallis Test (1 IV)
Cat.	1 +	2 +	Cont.	-	W	Repeated Measures ANOVA, Friedman's ANOVA (1 IV)
Cont.	1	-	Cont.	-	B	Pearson correlation, Kendall's tau, Spearman's correlation, Polychoric correlation
Any	1 +	Any	Cat.	2	B	Logistic Regression
Any	1 +	Any	Cont.	-	B	Linear Regression, Poisson Regression (DV: counts)
Any	1 +	Any	Any	Any	Any	Generalized Linear Mixed Model
Constant	-	-	Cat.	2	-	Binomial Test
Constant	-	-	Cont.	-	-	Z-Test

Table 4.1 – Hypothesis tests appearing in DCUS papers. In this paper we focus on tests with single categorical IV and single DV (highlighted in pink) for our assessment of reporting and effect sizes. Remaining tests have a green background. All tests in our sample had a single dependent variable. Three tests did not fit into this categorization: Bernoulli trial, factor analysis and k-means cluster analysis. DV = dependent variable, IV = independent variable, Cat. = categorical variable, Cont. = continuous variable, B = Between Subjects Design, W=Within Subjects Design

in their performance while completing a short reverse engineering task?. Our aim was to have a mix of hypothetical studies which were closely related to previous work as well as some that had no relations. This was done to a) ensure that it would be easy to find very specific data from closely related work as well as to b) ensure that our categories were useful to guide researchers in uncharted territory as best possible. Based on the case studies, we added features to categorize variables. Finally, after laying the theoretical foundation, we implemented a database and started to enter information from the collected literature.

4.3.2 Data Structure

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

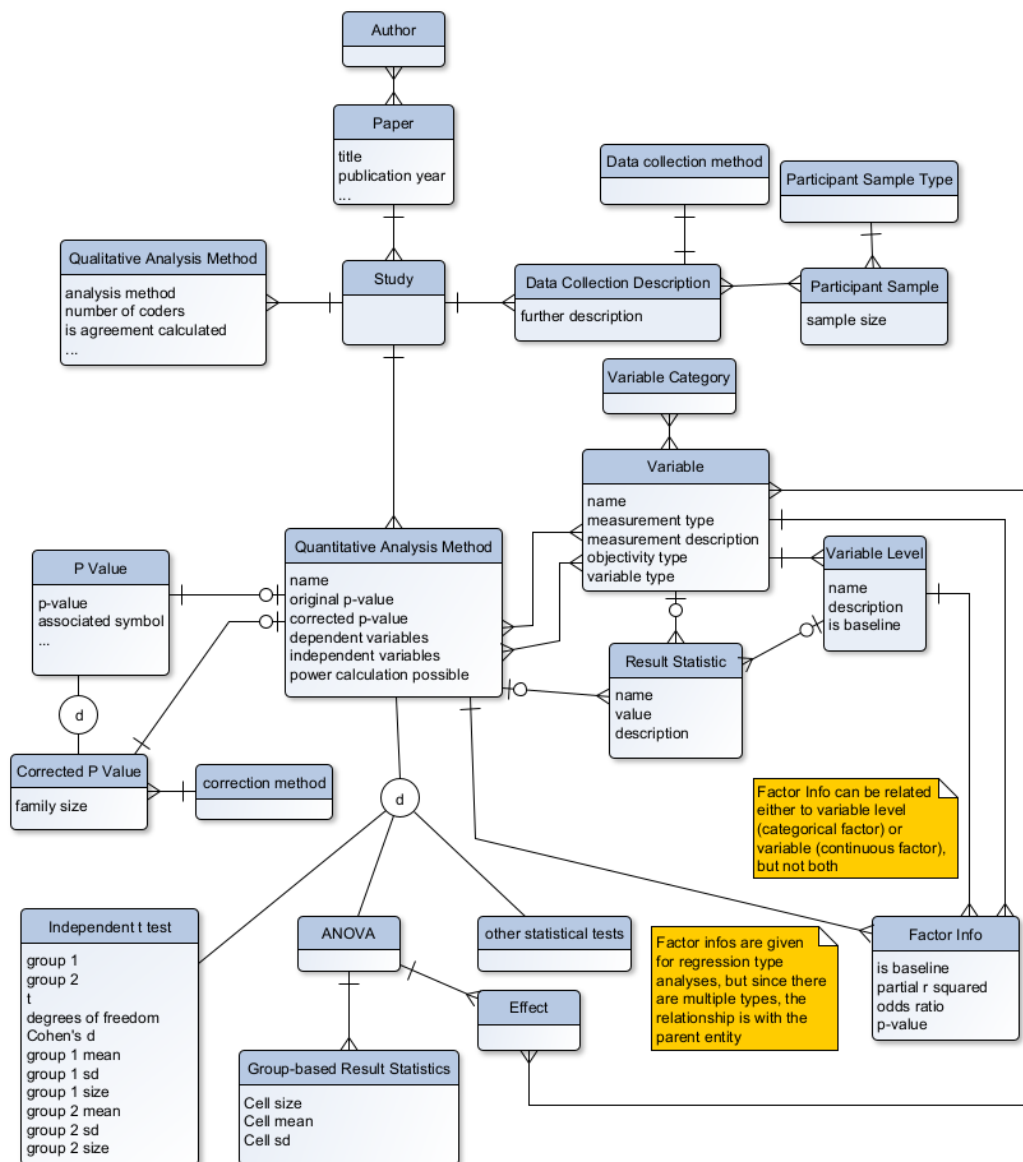


Figure 4.1 – Simplified ERD of the database, not all attributes and entities are depicted.

4.3.2.1 Overview

A general overview of the data structure as a simplified entity relationship diagram (ERD) is in Figure 4.1. For each paper in our set, we first collected meta information about it. Each paper can have multiple studies assigned to it. A *study* is a self-contained unit of a combination of data collection and analysis, which is often presented in a separate section in a paper. We separate *data collection descriptions* from the *participant samples* involved. To support filtering and generalization, *Participant sample types*, instances of which could be “student”, “security expert”, “freelance developer”, and *data collection methods*, where instances are, e.g., “interview”, “survey” or “experiment/task-based evaluation”, are represented as separate entities. These are more easily

reused across multiple studies and multiple papers. The results of qualitative analyses cannot directly be used for power analysis or to support meta-analysis since effect sizes are not calculated. However, because insights from qualitative analysis often help inform further quantitative work, we also collected some information on the *qualitative analysis methods* used. Reporting of NHST-type analysis methods all share certain properties, which we collected for all *quantitative analysis methods*, i.e., the name of the hypothesis test, the p-values, and the dependent and independent variables. For more information on the representation of variables, see Section 4.3.2.2. We categorized the different hypothesis tests used in our sample according to the number and type of dependent and independent variables, the study design, and, in the case of categorical variables, the number of levels in the variable, see Table 4.1. In the following, we focus on those hypothesis tests with one categorical independent variable with two levels. We collected additional test-specific data for these tests, see Section 4.3.2.3.

4.3.2.2 Representation of Variables

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

Finally, a *Variable* can be tagged with one or multiple *Variable categories*. These are broad categories of variables, which frequently occur in studies in DCUS, and which serve to ease the search for a specific variable or test and were used to create generic guides for researchers when no specific prior work exists. We generated these categories by open coding all the variables in eight of the papers from our sample, which we chose to cover a broad range of topics and both descriptive and inferential work. One researcher did all the coding alone, and the generated categories and the corresponding variables were frequently discussed together in an iterative process with a second researcher and modified when necessary. Since we were assigning fixed categories to clear units of data, and a second researcher was involved in the generation of the variable categories, we consider this data simple to code and independent recoding to be unnecessary [327, 365]. The twelve categories which emerged from this process are:

usability Measures relating to overall usability, i.e., incorporating effectiveness, efficiency, and user satisfaction according to ISO 9241-11 [237].

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

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

efficiency Measures relating to time spent, e.g. how fast tasks are completed or worked on.

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

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

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

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

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

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

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

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

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

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

4.3.2.3 Relevant Test-specific Information for Power Analysis

We specifically focused on collecting information that would be needed to conduct a priori power analysis, as well as additional values, which are typically reported with a hypothesis test according to APA style [25]. For some tests, not all data could be contained in a single entity, e.g., ANOVA, linear regression, and logistic regression. For these, we created multiple related entities in our database, but the meta-analysis will be carried out in future work since very few of these tests were reported with enough detail for a robust analysis at this point.

4.3.3 Data Input and Checking

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

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

4.4 Meta-Analysis

In the following, we analyze and further systematize the data we gathered. For our meta-analysis of the current state of research in DCUS, we focus on information related to power analysis, e.g., we

analyze effect sizes in this field and investigate whether reporting is sufficiently detailed to enable power analysis using the data. We analyzed the data from our database using the Python packages *numpy*, *pandas*, and *matplotlib* and the R *tidyverse* [407, 568].

We analyzed a total of 54 developer papers, which encompassed 64 individual studies, of which 24 were quantitative, 24 were qualitative, and 16 used both quantitative and qualitative analysis methods. On average, 127 (12 - 330, median=83, SD=103) participants took part in quantitative and 22.2 (4 - 63, median=18, SD=15.4) in qualitative studies, and for mixed methods studies on average, 131 (36 - 408, median=53, SD=130) participants took part. This is similar to Caine’s analysis of papers at CHI 2014, where the sample size was also smaller for qualitative than quantitative work [77].

4.4.1 Variable Topics

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

Of those categories referring to components of usability, e.g., related to either effectiveness, efficiency, or satisfaction [237], *Participant judgment* was the most used variable category (109 times). This may be because it applied to all variables representing some sort of participant judgment of an evaluated system or a task, e.g., preference [109], confidence in task correctness [5], or criticality of data [467]. This was followed by *security* (44), as a specific form of effectiveness, which is due to our sample focusing on DCUS. *Functionality* (24) as another form of effectiveness, *usability* overall (14) and *efficiency* (12) appeared less frequently.

Of the other variable categories, which were not related to usability, *participant characteristic* (93), *behavior* (69), and *experience* (47) were the most frequent. Examples for *participant characteristic* are type of participant, e.g., [351, 285], although there is a separate category specifically intended for this, demographics, like state of employment [345] or type of organization [342], and other characteristics relating to participants’ opinions or attributes [509, 66]. *Experience*, e.g., with technology like programming languages [e.g., 438, 109], or specific tasks [e.g., 343, 286], often occurs together with *participant characteristic*, as experience can also be considered a defining characteristic of participants. In fact, the most frequent co-occurrence (30) between variable categories was between *experience* and *participant characteristic*. Variables measuring *behavior* included variables tracking participants’ behavior during a study, e.g., the number of times they executed a program [202], number of visited websites [286], or lines of code submitted [543], and variables assigned to participants’ outcomes retrospectively by researchers [e.g., 543]. Behavior could also be self-reported [e.g., 353, 509]. More specific categories, i.e., *participant type* (9) and *system type* (27) and *artifact related variables* (19) were not as frequent. There was a surprising amount of variables associated with meta-level aspects, such as *task related* (42) and *study related variables* (21), although some of these may be related to task success.

4.4.2 Use of Hypothesis Tests

We were especially interested in the frequency of hypothesis tests. In the studies in our sample, 7.30 hypothesis tests were conducted per study on average (SD=12.35), and given that a paper

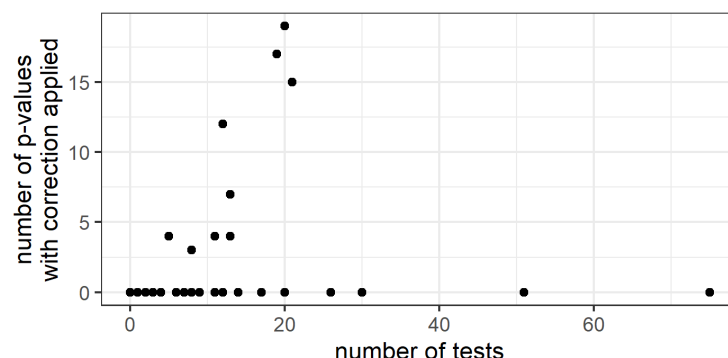


Figure 4.2 – Scatterplot showing the frequency of corrected p-values in relation to non-corrected p-values per study

could encompass multiple studies, the number of hypothesis tests per paper ranged from 0 (for purely descriptive papers, like [143], or qualitative papers, like [201]) to 74 (M=8.6, SD=13.2). One of the studies, which the authors identified as qualitative, nevertheless contained 26 statistical hypothesis tests. In this comparison of end-users' and administrators' mental models of HTTPS, Fisher's exact tests were used to compare the appearance of various concepts in the mental models of the two participant types [285].

Since some studies included a large number of hypothesis tests, with outliers at 75, 51, and 30 hypothesis tests in a single study, we explored whether p-value correction methods were used in our sample. Figure 4.2 shows that the majority of studies (31/41) did not use corrections for any of the reported p-values. This likely includes some studies where these corrections were not necessary since the hypotheses tested were about different outcomes or were explicitly considered exploratory [485]. However, all three studies with the largest amount of tests did not include any corrections that we could identify for their hypothesis tests. This means that the results presented may partially be false positives.

The frequency of each of the different hypothesis tests within our sample is displayed in the left half of Figure 4.3. The most-used test (123 times) in our sample of papers was the non-parametric Wilcoxon rank-sum (a.k.a. Mann-Whitney U test), followed by two tests used for categorical data, the Fisher's exact and chi-squared Tests, which appeared 59 and 32 times respectively. We categorized the different hypothesis tests used in our sample according to the number and type of dependent and independent variables, the study design, and, in the case of categorical variables, the number of levels in the variable, see Table 4.1. All of the tests used only one dependent variable. In the remainder of our meta-analysis, we focus on those hypothesis tests with one categorical independent variable with two levels.

4.4.3 Completeness of Statistical Reporting

Ideally, each of these hypothesis tests would be reported in sufficient detail to be able to conduct power analysis using the data reported in the paper. However, this is not the case, as shown in Figure 4.4. Of those tests for which we can make this classification, the paired t-test was reported with sufficient information in all cases. However, it was reported only once. For the other tests, the

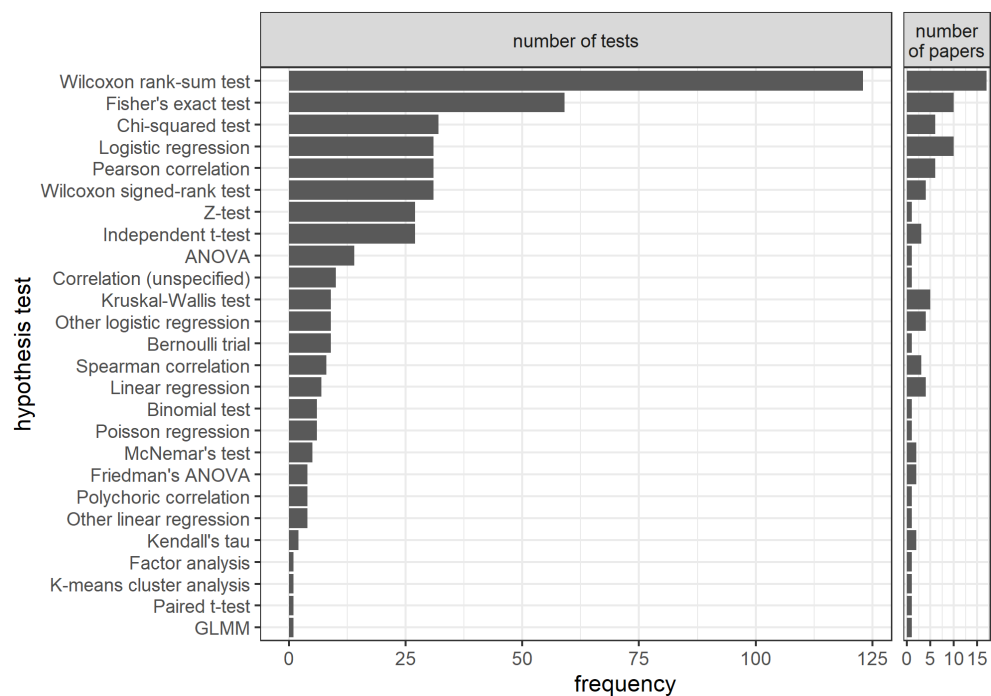


Figure 4.3 – Left side: Frequency of type of hypothesis test in the sample, Right side: Number of papers using this hypothesis test in the sample

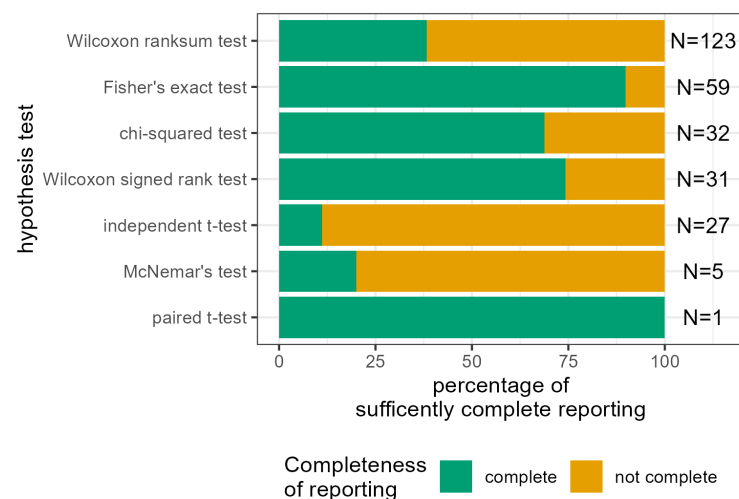


Figure 4.4 – Stacked Barchart depicting the proportions sufficiently complete reporting for those tests at the focus of our analysis

completeness of reporting varied between 11.1% and 89.8% of sufficient reporting (mean=50.3%, sd=31.9%). The most frequently reported test, the Wilcoxon rank-sum test, was reported with sufficient information 38.2% of the time, or in 47 of 123 cases. Overall, this shows that reporting practices, even for these simple tests, are not sufficient to do power analysis using them as a basis about half of the time.

4.4.3.1 Power Meta-Analysis

To assess whether the field of DCUS suffers from underpowered studies similar to other fields, we conducted a power meta-analysis based on Ellis [136, p.74]. This should not be confused with a post hoc power analysis since we did not use the effect sizes or p-values reported in the papers. Instead, we only used the reported sample sizes and used G*Power to calculate the power to detect small, medium, and large effects (Cohen's d equivalent of 0.2, 0.5, and 0.8 [91]). This was possible for five of our seven types of hypothesis tests. We excluded Fisher's exact tests and McNemar's tests from this analysis since G*Power required input of effect size based on concrete data from the study, i.e., success proportions for the Fisher's exact test and the total proportion of discordant pairs for McNemar's test. Since post hoc power analysis using values directly from studies like this is not useful [136], we concentrated on the other five types of tests. We set $\alpha = 0.05$ for all analyses and assumed two-tailed tests. For the non-parametric tests, we used the minimal A.R.E. setting in G*Power to get a conservative estimate of the achieved power. Next, we calculated an average over all achieved power values at each level of effect size per included study, and then an overall average [136]. We considered a power of 0.8 to be the lower bound of what should commonly be aimed for [91]².

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

4.5 Systematic A Priori Power Analysis

Conducting an a priori power analysis requires researchers to know or guess the standardized effect sizes (e.g., Cohen's d) they expect or want to detect. Alternatively, they can also know or guess a non-standardized effect size (such as 1 point on a 7-point scale) and additional information, such as the standard deviation of the groups. In a mature field with many studies examining similar variables (e.g., blood pressure), expected effect sizes might be common knowledge. However, in the

²Although deviations from this are perfectly fine when done consciously.

DVs in test	IVs in test	Partic- ipants	Test	ES	Descriptive stats	Paper
Attempted security (yes, no) security, behavior “participants who attempted to store user passwords securely, but struggled and then deleted their attempts from their solutions (this was coded as attempted but failed, or ABF). [...]”	Priming (priming, non-priming) study-related variable manipulated in experiment. “Priming - Participants were explicitly told to store the user passwords securely in the Introductory Text and in the Task Description.”	computer science students (N=40)	FET	OR= 19.02; d= 1.62	Proportion p1=0.7 (priming); Proportion p2=0.1 (non-priming)	[345]
Secure (secure, insecure) security “In addition, we used a binary variable called secure which was given if participants used at least a hash function in their final solutions and thus did not store the passwords in plain text.”	Warning displayed (yes, no) system type, study related variable whether a warning was displayed (could only happen in PyCrypto patch condition). “PyCrypto control condition, or the PyCrypto patch condition, where we tested our security warning”	Python developers (N=53)	FET	OR= 56; d= 2.22	Proportion p1=0.727 (yes); Proportion p2=0.0455 (no)	[184]

Table 4.2 – Examples of information from the database which can be used for power analysis. The two examples are tests, where there was sufficient data to conduct power analysis. ES=effect size, FET=Fisher’s exact test

absence of closely related prior work, as is often the case in relatively new fields, such as DCUS, a more realistic approach is for the researcher to decide what the smallest practically relevant effect size is, that they want to be able to detect [136], also called the smallest effect size of interest. This can be determined by theoretical considerations but also by juxtaposing the benefit of the desired outcome with the costs to achieve this outcome or by considering practical limitations, such as the number of available participants [292]. In any case, researchers need experience and deep knowledge of their field to estimate effect sizes and this is one aspect that makes power analysis difficult [92]. In addition, used effect sizes [136] and power analysis procedures vary between different types of statistical tests [149].

Our systematization of knowledge aims to ease this process. Ideally, related studies have already been published, and standardized effect sizes are reported or can be calculated. In this case, the researcher needs to be able to find them. For this, they can query our database introduced in Section 4.3. Some types of information systematized in the database which are helpful for the search include variable categories, variables themselves and participant sample type. The participant sample type can help researchers identify prior work with a similar demographic to their planned study.

Table 4.2 shows an excerpt of the data which can be returned by the database. The first row contains all the data needed to perform a power analysis for a Fisher's exact test. A researcher could find it by any of the keywords or categories listed. The second row contains all the information to perform a power calculation for a different Fisher's exact test. The full data set is available on our companion website³. For each test, variables, effect sizes, relevant information to calculate them, the participant sample, and the source paper are listed and sorted by variable category.

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

4.5.1 Effectsize Meta-Analysis

As a final step in our systematization, we conducted a descriptive meta-analysis of the standardized effect sizes from tests where we had enough information for power analysis. With this, we aimed at providing a broad overview of effect sizes in DCUS which can be used to check the plausibility of power analyses. While it is preferable to find exact or at least close matches, we also want to support researchers where this is not possible. Without related work, researchers basically have to guess standardized effect sizes or things like expected proportions or standard deviations. To at least give a frame of reference against which to judge these guesses, we examined the range of effect sizes present in the field of DCUS. We used the categories from Section 4.4.1 to aggregate the data for our guide.

To enable comparison and aggregation, we used the effect sizes directly reported in the paper where possible and converted them to Cohen's d , as this is one of the most widely used effect sizes in our sample. In other cases, we first used the provided data to calculate an effect size, which we then converted to Cohen's d . For converting d to OR, we used the formula from Haddock et

³<https://powerdb.info>

al. [193], and a correction factor of 1.09, which is an average over the correction factors Poom and af Wählberg recommend for sample sizes between 20 and 100 [399], since developer studies mostly feature smaller sample sizes. To convert between d and ϕ , we used Rosenthal's formula [432] as described by Burns et al. [72]. Finally, as described above, sufficient data was not reported for many tests, and we exclude such tests from our analysis.

When judging the size of effects, effect sizes of Cohen's $d=0.2$ are generally regarded as small, $d=0.5$ as medium, and $d=0.8$ as large effect sizes [91]. When converted to other effect sizes, this yields $OR=1.37, 2.21, 3.57$ as small, medium, and large effect sizes displayed as odds ratios, $\phi=0.10, 0.24, 0.37$ for the effect size ϕ used with χ^2 -tests. While we did not encounter the correlation coefficient r used as an effect size in this analysis, we nevertheless note for researchers encountering r , that effect sizes of $r=0.1, 0.3$, and 0.5 are considered small, medium and large effects respectively [91].

In our DCUS sample, the reported effect sizes ranged from equivalents of Cohen's $d=0$ to Cohen's $d=2.22$ ($M=0.55$, median= 0.47 , $SD=0.42$). We excluded two effect sizes from a Fisher's exact test in [345] and a McNemar's test in [510], where the reported success proportion in one of the groups was zero, and thus the effect size approaches infinity. Figure 4.5 shows the distributions of effect sizes separately for the variable categories of the dependent variables. So if researchers have to make educated guesses for their power analysis, they can compare their values with the violin plots to judge where in the spectrum they lie.

Four variable categories were not assigned to dependent variables with a present effect size: experience, system type, participant type, and task related. These categories were more frequently applied to independent variables. We did not plot data for variable categories where fewer than five effect sizes were reported. This is the case for three variable categories: usability ($N=2$), participant characteristics ($N=1$), and study related variables ($N=2$). All of the six remaining variable categories exhibit a large variance of effect sizes, which range from negligible and small to large and very large effects. Median effects for tests with artifact-related variables as the dependent variable are in the medium range, and for all other categories and overall, the median effects are in the small range. However, in the cases of behavior, security, and participant judgment as well as overall, they border on medium according to Cohen [91].

4.5.2 Publication Bias Correction

There is one final important warning. Irrespective of whether single entries from the database are used or our aggregation violin plots, researchers must be mindful of the effect of publication bias on reported effect sizes [234]. Since papers with statistically significant results are more likely to be published and studies in DCUS often have fairly poor power it should be expected that reported sample effect sizes are larger than true population effect sizes (see Ellis, p.79ff. [136]). Ideally, our database would also include work that is methodologically sound but did not get published due to statistically non-significant tests. However, since we could not think of any feasible way of including this at scale, we recommend taking this publication/sampling bias into account.

Consequently, when planning a study using effect sizes from related work or from our systematization or even a pre-study, it is recommendable to either correct the acquired effect size estimates [533, 238] or increase the desired sample size to be able to detect slightly smaller effects than ones reported in prior work. While this correction still requires some guesswork, it is a lot easier than having to guess blindly.

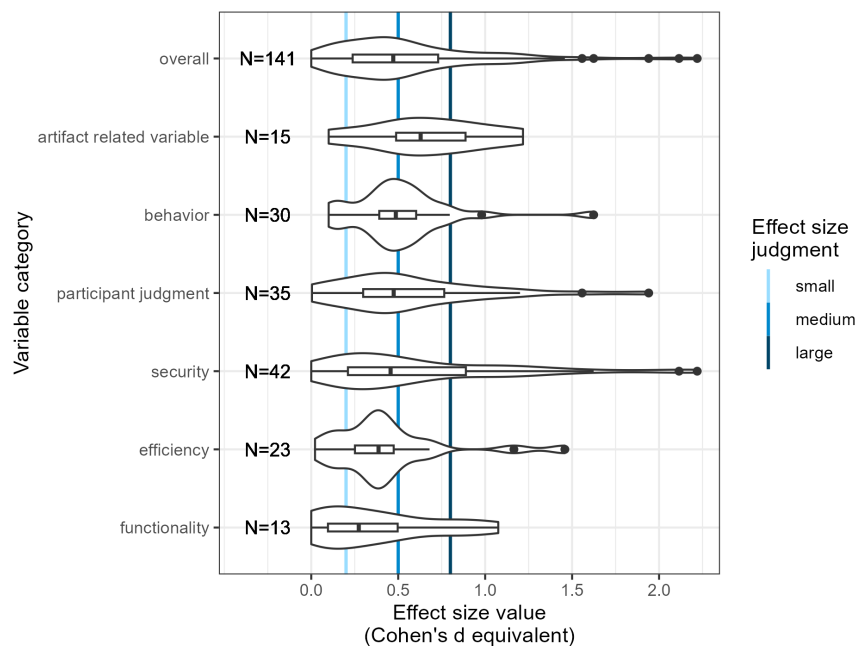


Figure 4.5 – Violinplots for the distribution of effect sizes for tests, faceted by the variable categories of the dependent variable

4.6 Limitations

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

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

Our work focuses only on a priori power analysis. Power, and as such, power analysis is situated within the NHST framework, and the analyses of the field and recommendations in this work apply within NHST. NHST results in point estimates about coefficients or effect sizes and a priori power analysis serves to determine the number of participants needed to detect such an effect at a specified significance level. The practice of focusing on the value of point estimates, rather than the precision of the estimates has also been criticized [324]. Even when planning a study with adequate power to detect an effect, the confidence intervals around it can be wide, leading to little precision of the effect size. Different analysis methods also exist, e.g., Bayesian analysis [257, 264]. Additionally, some statistical analyses, like regression, are not used merely in NHST to falsify hypotheses, but also to make predictions, and different judgment criteria would apply in these cases [75]. In predictive analysis, i.e., what is often known as machine learning, the influence and explanatory power of individual variables included in the model is not as important as the

accuracy of the prediction [67, 75]. Other criteria for what constitutes good reporting may apply in these cases than what we have covered in this work. However, NHST is commonly used in DCUS specifically and HCI in general. We did not in fact encounter any Bayesian analyses in our sample of analyzed DCUS papers and given that recruiting software developers is hard [7], having sufficient data for large-scale predictive analyses is likely rare in this field. In conclusion, we believe that our contributions align with common methods used in DCUS at the time.

Finally, we only conducted our meta-analysis on tests with a single categorical IV and a single DV. Thus, our work is limited to assisting in the power analysis of the following tests: Fisher’s exact test, chi-squared test, McNemar’s test, Wilcoxon rank-sum tests, Wilcoxon signed-rank tests, and paired and independent t-tests. We aim to extend this in future work. However, in combination with simplifications even the current data might offer benefits for the assessment of more complex tests[386, 296].

4.7 Conclusion

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

4.8 Future Work

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

measured objectively or subjectively using the categorization of variables, or to conduct sensitivity analysis, i.e., to analyze the power to detect effects.

Chapter 5

Power and Effect Size at CHI

Disclaimer:

The contents of this chapter were published as part of the publication “Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation” at CHI 2025. This paper is the result of a collaboration with my co-authors Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas and Matthew Smith and additional students involved in the data extraction, collection and analysis: Dorian Heidorn, Lysander Lenzen, Maximilian Mundt, Antonia Sistig and Lukas Wenz. The project idea was conceived by Matthew, Florin and me, and I organized and supervised student work on different aspects of the project. An initial extraction system was implemented by Lysander under Florin’s and my supervision. Theo implemented the download and filtering function for quantitative papers. Mischa and Theo built the final extraction process on top of the Cyber Reasoning System implemented by Mischa’s WASPS Team for the AIXCC competition¹. I conceived an initial prompt and Theo refined it further. Florin implemented a JSON formatter to aid in extracting a ground truth and Dorian implemented extraction of CCS categories. Antonia and Lukas implemented a plausibility checker to check presence of extracted values in the html file. I extracted a ground truth data set from 25 papers with feedback from Florin and Lisa. Maximilian, Lisa and I conducted an initial qualitative analysis of researchers’ interpretation strategies in a sample of quantitative CHI papers from 2023 and Lisa and I finalized and refined the qualitative analysis. I conducted the meta-analysis of effect sizes. I wrote the initial draft of the paper and revised it together with Florin (especially the abstract, introduction, conclusion and discussion), Lisa (especially the qualitative analysis, results and discussion), Mischa and Theo (especially the extraction method) and Matthew. I prepared the paper for publication with Florin, Lisa and Matthew.

5.1 Motivation

Researchers frequently emphasize statistical significance when presenting the outcomes of statistical tests. However, effect sizes are equally if not more important [136], but are reported and discussed far less frequently. This is problematic for two reasons: firstly, knowledge about effect sizes in related work is vital to understanding the practical significance of statistical results and to putting

them in context to what is common in the research field. Secondly, knowledge about common effect sizes offers valuable guidance for power calculations.

A priori power analysis is ideally used during the design phase of a study to estimate the sample size necessary to detect a desired effect. Eismayer et al. found that only 5 of 519 experimental papers at CHI used power analysis [134]. Underpowered studies present a threat to scientific validity since they increase the likelihood of both Type I and Type II errors, overestimate effect sizes, hamper future research and waste research resources [136, 233, 234]. Ideally all experimental studies should conduct an a-priori power analysis [136]. To do this researchers have to decide what the minimum effect size which they want to be able to detect is. Ideally, this is guided by effect sizes from previous similar work. However, since work published at CHI focuses on novel contributions [370, 264] there may not be enough prior work that is closely related. In this case, the power calculation can be guided by meta-analyses of effect sizes in the research subdomain. However, for CHI, this does not exist yet.

Similarly, these meta-analyses would be helpful for the interpretation of effect sizes in the context of the research area. Currently, results are often interpreted using a binary scale - there is an effect or there is no effect. This is usually done by using the 0.05 p-value cut-off criterion [47]. Further interpretation of effect size is mostly limited to using the labels of “small” or “large” from generic guidelines, if it is done at all.

However, generic, domain-agnostic guidelines for judging the size of effects like Cohen’s [91] are criticized for not taking into account domain-specific and study-specific information [465, 151, 91, 397]. To remedy this, alternative guidelines have been proposed for various domains [397, 58, 253], including some in HCI [355]. However, for a wide range of subdomains in HCI, no effect size guidelines are available. This is unfortunate since effect sizes can vary quite widely within a discipline, as has been shown for psychology [307]. There is also a lack of guidance on how authors can structure their effect size discussion. The CHI submission guidelines call for transparent reporting to enable replication, which would include effect sizes, but they do not make concrete recommendations or offer examples [512].

To tackle these problems this work makes several contributions:

- 1) We created a meta-analysis support tool using an LLM (GPT-4o [363]) to extract and categorize statistics from HTML CHI papers. We focused on those values that are helpful for interpretation, power analysis, and meta-analysis: effect sizes, participant numbers, confidence intervals, and p-values. We built in plausibility checks to ensure the LLM was not hallucinating. We provide our extraction and analysis code², so other researchers can replicate our results, or extend it to a subset of research that is relevant to them.
- 2) Using our tool, we extracted statistics from all CHI publications between 2019 and 2023 that contained statistical tests (1692) and manually checked for inconsistencies. For our meta-analysis, we classified publications into 12 different CHI subdomains and created overviews for standardized effect sizes for each of these research areas. We focus on standardized effect sizes which can be converted into each-other: Cohen’s d , Pearson’s r , R^2 , odds ratio, common-language effect size (CLES), the non-parametric correlation coefficients Kendall’s τ

²<https://github.com/Behavioral-Security/A-meta-study-of-effect-sizes-at-CHI>

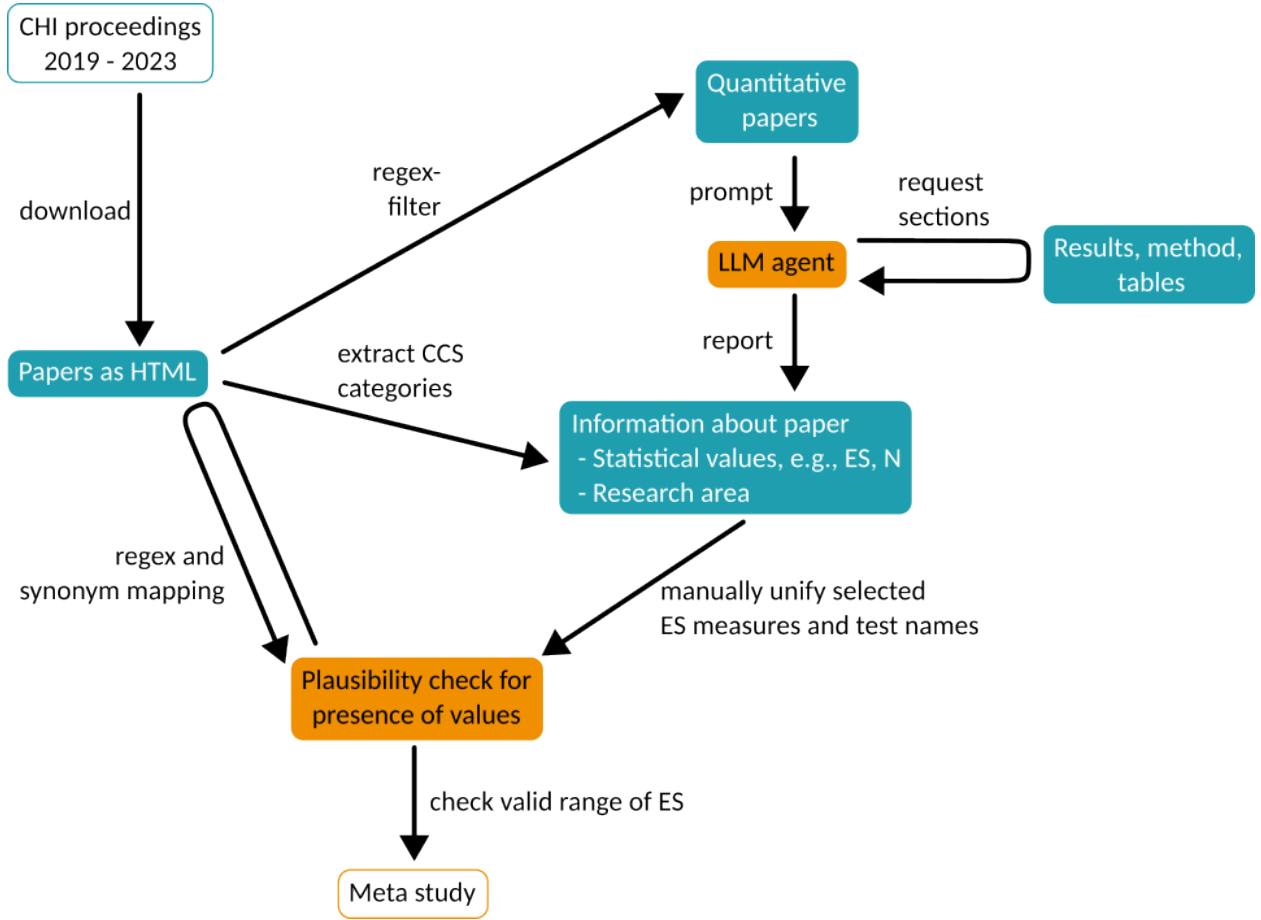


Figure 5.1 – Overview of Extraction Process

and Spearman’s ρ , Cramer’s V, and η^2 , including variants like η_p^2 or ω^2 . This data will aid researchers in interpreting their effects in the context of their domain and in planning power calculations.

5.2 Extracting Statistics

In this section we describe our data collection process to extract statistical information from five years of CHI publications. Figure 5.1 shows an overview of the extraction and analysis process.

We downloaded and filtered the CHI papers, so that only quantitative papers remained (see Section 5.2.1). For a given paper, we extracted identifying information, i.e. the title and DOI, as well as the abstract and structure of the paper from the HTML to pass to our LLM agent in the next step. As described in Section 5.2.3, the agent used our prompts and this information to request content from the paper and extract the relevant statistics described in Section 5.2.2. The final output per paper was converted to JSON and forms the basis of our meta-analysis, which we describe in Section 5.3. To enable re-use of our data, we provide the scripts used for analysis and

our analysis results on Github³.

5.2.1 Sample of CHI Publications

We selected five years of CHI publications for our analysis (2019 to 2023). CHI was first published in HTML format in 2019. This machine-readable format facilitates extraction considerably. Using optical character recognition (OCR) to generate machine-readable text from PDF is especially error-prone around tables [115], diacritics [57] and when mixing languages [377]. Statistics are often reported in a mix of Greek and Latin letters, with sub- and superscripts and frequently in tables, so using the machine-readable text from HTML directly is more reliable.

Concurring with prior work [77, 78], we define a quantitative paper as one reporting statistical tests. We used regular expressions to identify p-values, confidence intervals (CIs), and Bayes factors. P-values and CIs are specific enough to ensure that an actual test was conducted and reported, rather than theoretical papers discussing the merits of statistical testing methods, and at the same time not limited to a single type of hypothesis test in the way that e.g. effect sizes are. We configured our regular expressions to include variations in capitalization and plural and singular terms. For confidence intervals, we filtered for singular and plural of use of the term “confidence interval”. Since the often used abbreviation “CI” is a letter combination not uncommon in English, we did not use this as a filter. To get an idea of the frequency of use of Bayesian statistics for analysis, we filtered for inclusion of the term “bayes factor”, since Bayes factors are a common way to compare models in Bayesian analysis [454]. If a paper satisfied at least one of these criteria, we included it in our sample of quantitative papers. We evaluated our filtering criteria by manually checking a random sample of 145 papers evenly spread over all five years in our sample. In our evaluation, all but two papers were correctly classified, one each as false negative and false positive. False positives may occur due to meta-scientific or theoretical papers discussing statistics, but not using and reporting statistical tests, such as [70]. False negatives can be caused by papers using “p>”, “p=” or “p<” for values abbreviated p, which are not statistical p-values, although this did not occur in our filtering ground truth, or by papers reporting statistical tests, but providing statistical values only in images, e.g. [85] reporting the results of a Wilcoxon ranksum test in a Table included as an image.

Between 2019 and 2023, 3724 papers were published at CHI. Of these, we identified 1692 (45%) to be quantitative and used these as the basis for the rest of our work. The ratio of quantitative to qualitative papers is similar to findings in prior work [77, 78]. Of these quantitative papers, 13 (0.8%) were identified as using Bayesian methods, which we did not include in the meta-study.

5.2.2 Statistics of Interest

We focused on extracting the information most relevant when deriving domain-specific guidelines for effect size judgments: The name of the conducted hypothesis test, sample size N, the p-value, effect size and the confidence interval around the effect size. For some types of tests, like regressions with multiple factors or multi-way ANOVAs, multiple p-values, effect sizes and confidence intervals may be associated with a single test.

³<https://github.com/Behavioral-Security/A-meta-study-of-effect-sizes-at-CHI>

These statistics are generalizable in the sense that they apply to many tests in the NHST paradigm, which avoids the need for separate handling of every imaginable test. For this reason, we decided to focus on standardized effect sizes. Confidence intervals around the effect size can be used to gauge reliability of the estimates, however, reporting confidence intervals around effect sizes is currently not common at CHI. In their absence, sample sizes can be used to assess reliability. Test names are necessary to get information about the experiment design, as effect sizes may share a name, but be calculated differently for within group compared to between groups designs, as is the case for Cohen’s d , which has versions for both. Finally, while p -values are not directly useful for our further analysis, they seem to be the statistic reported most often and serve as an identification criteria of hypothesis tests conducted based on NHST.

5.2.3 Extraction Tooling

We focused our extraction on data available in text form, including the main text, tables and captions but excluding images. Text extraction from images requires OCR which can be inaccurate, especially for non-latin characters [377, 57], and if extracting values represented by visual elements, the accuracy depends on the scale of the image. We did not analyze supplemental material.

We used OpenAI’s GPT-4o [363] in combination with a LangGraph-based agent [298] to extract the statistics and used the lowest possible temperature settings to lower randomness in our extraction process. Our prompts are described in more detail in Section 5.2.4. Three tools were defined for the LLM to utilize: a section read tool, a table read tool, and a test report tool. The read tools allow the agent to request a section or table. The agent can specify the section or table by its index or heading. Sections are returned as plain text, while tables are returned in csv-format, accompanied by a prompt instructing the LLM to read the table row by row. The report tool provides the LLM with a set of predefined fields that must be satisfied with statistical test data, such as the p -value or the effect size measure type, or “UNKNOWN”. Upon receiving the initial prompt, the LLM determines which sections to request via the tool. Based on the section text, it identifies the statistical tests and makes use of the report tool. This process is subsequently repeated with the tables.

5.2.4 Prompt Engineering

We built on our experience extracting statistical information from papers, e.g. from creating the ground truth as described in Section 5.2.5, to derive an initial prompt. We improved on this prompt through manual experimentation, similar to prior work [508, 423, 588]. For example, in the extraction results based on our initial prompt, information from tables was largely missing. In subsequent iterations, we adjusted the prompts and tooling to consider tables separately. Our prompts consisted of a system prompt and a task prompt. The system prompt consisted of a persona description and general instructions including response format. The task prompt had a closer description of the task, including the statistics and information to be extracted and a description of special cases and how to handle missing information. Finally, the prompt included information about the paper which we extracted from the HTML beforehand: the title, abstract and structure, i.e. section indices and headings, as well as table indices and captions. We provide the final prompts in Appendix D.1.

5.2.5 Evaluation

We manually annotated a sub-sample of quantitative CHI papers ($N=25$) to compare our extraction results to. To create this extraction ground truth, we collected a random sample of five quantitative papers per year from our sample. One researcher with a research focus on methodology and statistics manually extracted the relevant statistics and entered them into a JSON data structure so they were human- and machine-readable. To reflect the data available to the LLM, they only used information from the main text, tables, and captions. The researcher corrected typographical mistakes (“ $p < 05$ ”, e.g. missing the decimal point). Context was used where possible to derive values, e.g. sample sizes for an individual test from tables reporting demographics, or from degrees of freedom and the number of conditions in a test [374]. In case of uncertainties the researcher discussed these with their co-authors and noted remaining uncertainties within the ground truth.

For evaluation based on the ground truth, we focused on the effect sizes and that they are reported together with the correct test and sample size. From the 25 papers, we extracted 528 tests of which 438 did not contain effect sizes. From the remaining 90 tests, we extracted 403 effect sizes. Of these, 130 were relevant for the meta-study. The remaining effect sizes consisted of 258 regression coefficients and 15 correlation coefficients where the type of correlation was not specified. We could not use regression coefficients for the meta-study because additional descriptive statistics, i.e. the standard deviation of the outcome variable, and group sizes, are necessary to convert them to an effect size useful for meta-analysis [308].

Overall, 86 of the 130 effect size values were extracted correctly by the LLM and 2 were extracted with a missing $-$, i.e. the wrong direction of effect was extracted, which is not critical, since we use absolute values in the meta-study. 42 were missing. Of these, 23 effect sizes were only reported in tables, and not individually addressed in the text, 12 were Cohen’s d s reported as part of a meta-analysis, but only in the form “ $d=$ ”, which might not be clearly associated with statistics, and one was reported in the method section, which might not have been requested by the section read tool. The remaining 6 missed effect sizes were clearly identifiable as statistics by us and we do not know why the LLM did not extract them. No effect size values were extracted incorrectly. Having no incorrect effect size values was our main goal, so we accepted the 32% false negative rate

Of the relevant 90 tests we extracted manually, the LLM extracted 64, of which 37 were extracted with correct and complete test names, 9 with correct, but underspecified names (e.g. “t-test” instead of “dependent t-test”) and 14 with incorrect names. The test names are only relevant to distinguish between within or between groups designs. For the 64 tests which were extracted by the LLM, participant numbers were extracted correctly in 49 cases and incorrectly in 12 cases, where the LLM did not recognize that a different subsample was used in the test. Three tests were not reported with determinable sample sizes. The sample sizes do not directly affect the effect size, but only the weighting of the effect sizes in the meta-study.

Given that we prefer missing data to wrong data, we considered this sufficient for use in our meta-study and used the extraction procedure on our full sample of CHI papers.

5.3 Constructing Effect Size Guidelines for CHI Research Areas

We converted the effect sizes extracted in the previous section to a common effect size to investigate effect size distributions in 12 research areas at CHI.

5.3.1 Categorization of Papers

To identify papers' research areas for our meta-study, we started with the top level CCS categories. A paper can fall into multiple different categories. *Human-centered computing* was unsurprisingly the most frequent research area by a large margin (3280, or 91.6% of all papers) with the next most frequent category being applied computing (10%). Other categories were also fairly general, making them less useful. To remedy this, we decided to additionally use the subcategories of these general three categories, *general and reference*, *human-centered computing*, and *applied computing*, for more concrete and thus interpretable research areas.

For our sample of quantitative CHI papers, this resulted in 141 different categories. In the following we describe the categorization for a subset of these, the 599 papers, which reported effect sizes which we used in our meta-study. For the most frequent CCS categories, we merged two pairs of categories which were conceptually similar. We merged *user studies* and *empirical studies in hci* as well as *empirical studies in collaborative and social computing* and *collaborative and social computing*. We then chose the top 10 most frequent categories, which were specific enough to allow distinction between papers and covered the largest possible number of these papers. There were two more categories which appeared the same number of times as the least frequent of these categories, in 27 papers, so we included these to avoid excluding categories due to alphabetic ordering, and not different frequency of appearance. This resulted in 12 research areas. While we extracted statistics and conducted our meta-analysis for all quantitative CHI papers, we focus analysis per research area on these 12. Only 12 papers of 599 do not fall into one of these areas, they are nevertheless included in our meta-analysis when we discuss effect sizes at CHI overall.

While we would have liked to use CHI-specific categories such as the subcommittees, data on subcommittee assignment is not available publicly. Of the 12 categories we discuss, *collaborative and social computing*, *interaction design* and *security and privacy* match well with the more CHI specific categorization into subcommittees of *Interaction Beyond the Individual*, *Privacy and Security* and *Computational Interaction* respectively. Some matches between these two categorization schemes are not as clear. Our categories of *interaction design* and *human computer interaction (hci)* also fit to some extent with the CHI subcommittees of *Blending Interaction* and *Interacting with Devices*. Our results for these categories are directly applicable for researchers submitting to CHI. The subcommittees of *Accessibility and Aging* and *Learning, Education and Families* have overlap with the category of *social and professional topics* and the same applies to the subcommittee of *Specific Application Areas* and *applied computing*. Our guidelines for these categories may be better suited than those for CHI overall, but researchers should take their own research specifics into account, even more than for some of the other categories. While our research areas of *virtual* and *mixed/augmented reality* are concrete and identifiable for researchers, they lack an equivalent within the subcommittees.

5.3.2 Converting Effect Sizes

For our meta-study, we focused on standardized effect sizes which can be converted into each-other: Cohen’s d , Pearson’s r , R^2 , odds ratio, CLES, the non-parametric correlation coefficients Kendall’s τ and Spearman’s ρ , and η^2 . In addition, η_p^2 (partial eta-squared) is equivalent to η^2 in one-way ANOVA, but not for multi-way designs [291], where all η_p^2 may add up to values of over 1 [424]. However, judging the size of η_p^2 is possible through effect size specific benchmarks derived from those set up for Cohen’s d [424, 91] and as such, we included it in our calculations. Similarly other variants of η^2 , such as generalized η^2 (η_G^2) or less biased variants like ω^2 may not always yield the same value when applied to the same data set, but they are measured on the same scale as η^2 and we treated them as such in our calculations. On the other hand, we excluded pseudo- R^2 used in logistic regression, such as McFadden’s or Nagelkerke’s pseudo R^2 , since they are not measured on the same scale, but their true maximum depends on the data used in the model [475, 431]. Similarly, we excluded ϕ , since it is only the same as Cramer’s V in 2x2 contingency tables, but different in larger tables [98], and based on our extraction, we cannot tell reliably which is the case.

Given that more conversion formulas were available for r , and its value range is clearly bounded in contrast to Cohen’s d , where the upper limit is ∞ , we converted all effect sizes to Pearson’s r . Since we were not interested in the direction of effects, but only the size, we used the absolute value of the reported effect sizes in the conversion. We applied the average correction factors to our conversions, which Poom and af Wählberg derive through Monte Carlo simulation to improve the accuracy of converted effect sizes for meta-analyses [399]. For individual cases, this can lead to values outside the valid range for r . We report and incorporate these in our meta-study as is to provide consistency for conversion back to the original effect size. Researchers using the guidelines should refer to Table 5.1 for the valid range of effect sizes and only interpret them within this range. The conversion formulas we used are also in Table 5.1.

5.3.3 Data Cleaning

We describe the process of data unification and cleaning in the following.

Some statistical tests have multiple names, a common example being the Wilcoxon rank-sum test, also known as Mann-Whitney U test or Wilcoxon-Mann-Whitney test. In addition, our LLM-extracted data often contained additional information on the variables involved in the test as part of the name. For further analysis, we unified the test names by mapping them to a list of test names we collected from a statistical text book [153]. We added the names of extracted tests which did not correspond with the tests on our list where applicable. We followed a similar procedure for the effect size measures, except that we focused specifically on effect sizes we would use in our meta-study, as described in Section 5.3.2.

In our analysis of the ground truth, we noticed that sometimes, effect sizes were reported twice: once as a factor, and once with the omnibus test, so for our final analysis we removed 6009 rows of data containing duplicates.

Since manual extraction and checking for a ground truth is time and labor intensive, and cannot tell us about performance in the full sample, we also implemented several plausibility checks using regular expressions on the HTML source, focusing on effect sizes and N s, as we primarily used

ES	equiv- alent ES	valid range	conversion formula	source	correc- tion factor
r	ω, η	-1 to 1	-	-	-
Cohen's d		$-\infty$ to $+\infty$	$r \approx d/(d^2 + 4)^{0.5}$	[161]	1.27
OR		0 to $+\infty$	$r \approx (OR^{0.5} - 1)/(OR^{0.5} + 1)$	[154]	1.46
Kendall's τ		-1 to 1	$r \approx \text{Sin}(0.5 \times \pi \times \tau)$	[269]	1.01
Spearman's ρ		-1 to 1	$r \approx 2 \times \text{Sin}(\rho \times \pi/6)$	[439]	1.02
η^2	$R^2,$ $\omega^2,$ ϵ^2	0 to 1	$r \approx \sqrt{\eta^2}$	[161, 153]	-
Cramer's V		0 to 1	$r \approx V$	[98]	1.44
CLES		0 to 1	$r \approx \sin((CLES - 0.5) \times \pi)$	[129]	-
Cohen's f		0 to $+\infty$	$d \approx 2f$; use conversion for Cohen's d	[91]	-
Cohen's f^2		0 to $+\infty$	$r \approx \sqrt{f^2/(f^2 + 1)}$	[91]	-

Table 5.1 – Overview over the effect size measures included in the meta-study, including valid ranges, formulas for conversion of effect sizes to r and sources for the conversion formulas. Correction factors are mean correction factors taken from Poom and af Wählberg [399]. Correction factors were not available for all effect size measures.

these values in the meta-study. Due to inconsistencies in how statistical values are formatted, a simple check that matches exact values in the HTML is insufficient. Therefore, we designed regular expressions to capture variations, such as leading zeros, white spaces, and different formula representations, and also created a synonym mapping for statistical measure names, as these are used inconsistently across different authors. These plausibility checks initially flagged 235 effect size values and 110 Ns, which we manually checked. We removed 34 wrong effect size values and 40 wrong Ns from the data set.

As an additional plausibility check on the effect size values, we checked that they were within the limits of a valid effect size of the reported effect size measure, as depicted in Table 5.1. We removed 43 of 6755 effects for invalid effect sizes. These invalid effect sizes can be due to a hallucination, but can also be introduced into the papers through human error, such as the ones we found during the creation of our human-annotated ground truth, where we saw a correlation coefficient larger than 1.

5.3.4 Data Analysis

We used R [408], specifically the `tidyverse` [568] for data manipulation and `weighted_quantile()` from the `modi` package [228] for the meta-study.

Effect sizes from within- and between groups designs should not be compared directly, since the error variances are calculated differently for different designs, especially in the case of η_p^2 [359], which was quite frequent in our sample (21.3% of the valid values). Due to our method of extraction, we used information from the extracted test names and in some cases also effect size measures to derive the experiment design, e.g. independent t-tests are conducted for between groups designs

and dependent t-tests for within groups designs. However, due to unclear reporting of tests as merely “t-tests” without further specification or various types of regression where determining the use cases in terms of experiment design required disproportional effort, 1379 (20.6%) of effect sizes could not be assigned to an experimental design for certain. Analysis results from within-groups designs have less random error and e.g. the same difference between groups results in a larger effect size. Wrongly counting an actually within-groups effect size from an unclear test as a between-groups effect size means it will likely be comparatively larger than the between-groups effects. Within-groups studies generally have less random error than between-groups studies. Thus, given an effect size reported with an unknown study design, classifying this as between-groups would represent a larger effect (e.g., mean difference) than classifying E as within-groups. Therefore, misclassification of a within-groups effect size as between-groups will shift the overall distribution of the effect sizes to be larger than they actually are. Overestimating effect sizes is already a problem due to publication bias [234]. We believe underestimating effect sizes to be less grave and wanted to retain the information value of the unclear effect sizes and so we counted them as within-groups.

For each of these analyses, we first calculated median effect sizes per paper, weighted using the sample size of each test. We chose to use the median as a measure of central tendency, since we are interested in typical effect sizes at CHI and robustness against outliers. Where sample size was not available, we calculated the median sample size for the paper as a stand-in, and if no sample sizes were extracted for a given paper, we substituted the overall median sample size, 40. For extreme values of sample size > 352.5 , using Tukey’s fences as the outlier criterion [519], we substituted the maximum non-outlier value: 352.5. Tukey’s fences identify outliers as values that fall beyond 1.5 times the interquartile range (IQR) above the 75% quantile or below the 25% quantile [519]. We used the sample size as weights, since larger sample sizes lead to more reliable estimates [136] and sample sizes were more consistently reported and extracted than for example confidence intervals, which also help judge reliability. Calculating a per-paper median first avoids undue influence of a single paper with an extreme amount of tests on the average overall [136]. We did this separately for the within-groups and between-groups designs. We grouped these per-paper measures of central tendency according to the categories assigned to each paper to get the distributions of effect sizes for different research areas at CHI. We followed prior work on domain-specific effect size guidelines, which approximated small, medium and large effects in a specific research area, by dividing the distribution into thirds and calculating the median effect size within these thirds as a threshold between size categories [355, 253]. To counter the bias towards large effect sizes [136], we again used medians weighted by sample size.

5.3.5 Limitations

Even though there has been criticism of NHST and prior work has argued that Bayesian statistics are better suited for CHI’s needs of gaining precise knowledge and estimates even from small sample studies [264], we focused our meta-study on NHST statistics and effect sizes, as this is still the most commonly referenced statistical paradigm at CHI. While Bayesian statistics may be more suitable, they also require a significant learning curve to get started [391]. Phelan et al. have investigated providing analysis templates to make it easier, but it was still hard for researchers without a statistics background to trust the priors recommended by the templates [391]. This

means that wide-spread adoption of Bayesian statistics could still be some way off, so we believe it is worthwhile aiding knowledge accrual when NHST is used in the meantime.

Our LLM extraction process, like any use of LLM, can include hallucinations [243] or other mistakes. We did our best to mitigate their influence as much as possible. We used the lowest possible temperature settings to lower randomness in our extraction process. Additionally, we used automatic and manual plausibility checks especially on numeric values to check whether they actually exist in the papers. For the effect sizes in our meta-study, we reviewed that the effect size measure fit with the test it was reported with and additionally screened out values not in the valid range of values for the specific effect size measures. On the whole we designed our approach to err on the side of caution, rather accepting that effect sizes were missed than incorrect ones were included.

Our effect size meta-analysis is naturally only based on those tests which report effect sizes. Unfortunately these are not always reported. Additional information such as exact test/experimental design and sample sizes also are not always clear and we had to make inferences. Our ground truth analysis highlighted that our LLM missed effect sizes not clearly referenced as such in the text. Thus, our meta-analysis only provides a view on a subset of effects at CHI. Our LLM also did not always correctly identify the correct subsample N , leading to possible overemphasis of certain effects. Finally, the LLM did not always extract the correct test names, leading to uncertainty, regarding the between or within groups study designs. In practice, researchers also often interpret effect sizes from within and between groups designs in the same way [448]. However, despite these issues, since all extracted effect sizes in our ground-truth evaluation were correct, we believe our data set is large and robust enough to serve as a valuable starting point for the community. We believe this is a useful step forward over the state of the art. In future work, we hope to establish automatically generated and machine readable supplemental material containing all necessary information on tests to remove the above issues.

Converting effect sizes to a different effect size measure can also introduce biases. Poom and af Wählberg showed in their simulations that common conversion formula may lead to systematic misestimations [399]. While some effect sizes are derived from the same basis, they are not all numerically equivalent - being based on different experimental designs [91]. Wherever possible, we use the average correction factors provided in their work, but they were not available for all effect sizes we incorporated [399]. We nevertheless wanted to include as many effect sizes as possible, since selective exclusion may also bias the results of the meta-study.

For the most part, effect sizes in our analysis were sample based effect sizes, that aim to provide an estimate of the effect in the population. However, they can be biased estimators [424, 291]. Some less biased effect sizes measures, like ω^2 for ANOVA-Type analyses or Hedge's g for mean comparisons are recommended [291], and were also used to a small extent in our sample, but for our overview of effect sizes at CHI, both less and more biased effect size measures were incorporated, to the extent that we identified conversion formulas for them. When considering only published results, as we do in our work, publication bias has an influence on the size of reported effects [234]. Papers with statistically significant results are more likely to be published, but papers with small sample sizes (such as many of the studies in our sample) have poor power to detect small effects, so it is more likely that published effects are larger than true population effect sizes [136]. We fully expect that updates to our recommendations will be necessary as the CHI community moves towards more detailed effect size reporting practices. However, to provide assistance to researchers now, we have to make use of the data we have now, biased as it may be.

Finally, because the CCS we used to categorize our sample are used by the whole ACM, they are not specific to CHI. As such some of the discussed research areas, like *security and privacy* are more granular and useful than other more general ones, like *empirical studies in hci*. To enable researchers to make usable inferences for the categories they view as most interesting, we have published our dataset, including the extracted statistics and categorization per paper on OSF. Through matching by doi, researchers with access to different categorization metrics could include these.

5.3.6 Results

We present an overview of the distribution of effect sizes reported in different research areas at CHI. Our sample for the guidelines consists of 599 quantitative CHI papers from 2019 to 2023, which reported effect sizes we could clearly identify and convert to Pearson's r . The most frequently reported tests were ANOVA (18.0%), followed by repeated-measures ANOVA (14.5%) and Pearson's correlation (5.1%). Correspondingly the most frequently reported effect sizes were η_p^2 (21.3%), Pearson's r (16.7%), and η^2 without any specifiers (14.4%), followed by Cohen's d (12.7%) and OR (10.8%). Among all of the valid effect sizes, 6.4% were reported without any hypothesis test specified.

Figure 5.2 shows the distribution of effect sizes calculated over all 599 quantitative papers (incorporating 4381 tests and 6712 effect sizes) as well as for the individual research areas. We also differentiate between within-group and between-group experiment designs. In Table 5.2 we report the median sample sizes and the number of papers⁴ used to calculate our thresholds. The higher these numbers are, the more robust the thresholds will likely be for a given research area. In our estimation, more caution is advised for the areas of interaction design, mixed / augmented reality, social and professional topics and computing methodologies. While we still find the thresholds useful for putting results in context, since it is the best context we have, for power analysis, we recommend adding some buffer because effect sizes might be inflated and/or unduly influenced by a small number of outliers.

Guidelines for all effect size measures included in our meta analysis can be found in Table 5.3. The research area specific size guidelines are provided in Table 5.4 for Pearson's r . We calculated the thresholds using Pearson's r , but for convenience, we provide the size guidelines converted for the other effect size measures, η^2 , Cohen's d , OR, CLES, Cohen's f , Cramer's V , f^2 , Kendall's τ and Spearman's ρ , in the appendix.

According to Cohen's guidelines, $r=0.1$, 0.3 , and 0.5 would be considered small, medium and large effects respectively [91]. For the overall between group effects in our CHI sample, these thresholds would be 0.14 , 0.27 , and 0.61 respectively, and for within group effects, 0.12 , 0.32 and 0.72 .

While the overall small and medium effect threshold for between groups at CHI is relatively similar to Cohen's guidelines, the threshold for large effects is higher. This is even more pronounced for within groups effects. But most importantly we see clear variation between the different research areas, e.g. with effect sizes generally lower in *security and privacy*, compared to *collaborative and social computing*.

⁴Papers can contain both within and between group effects thus the numbers do not add up to 599

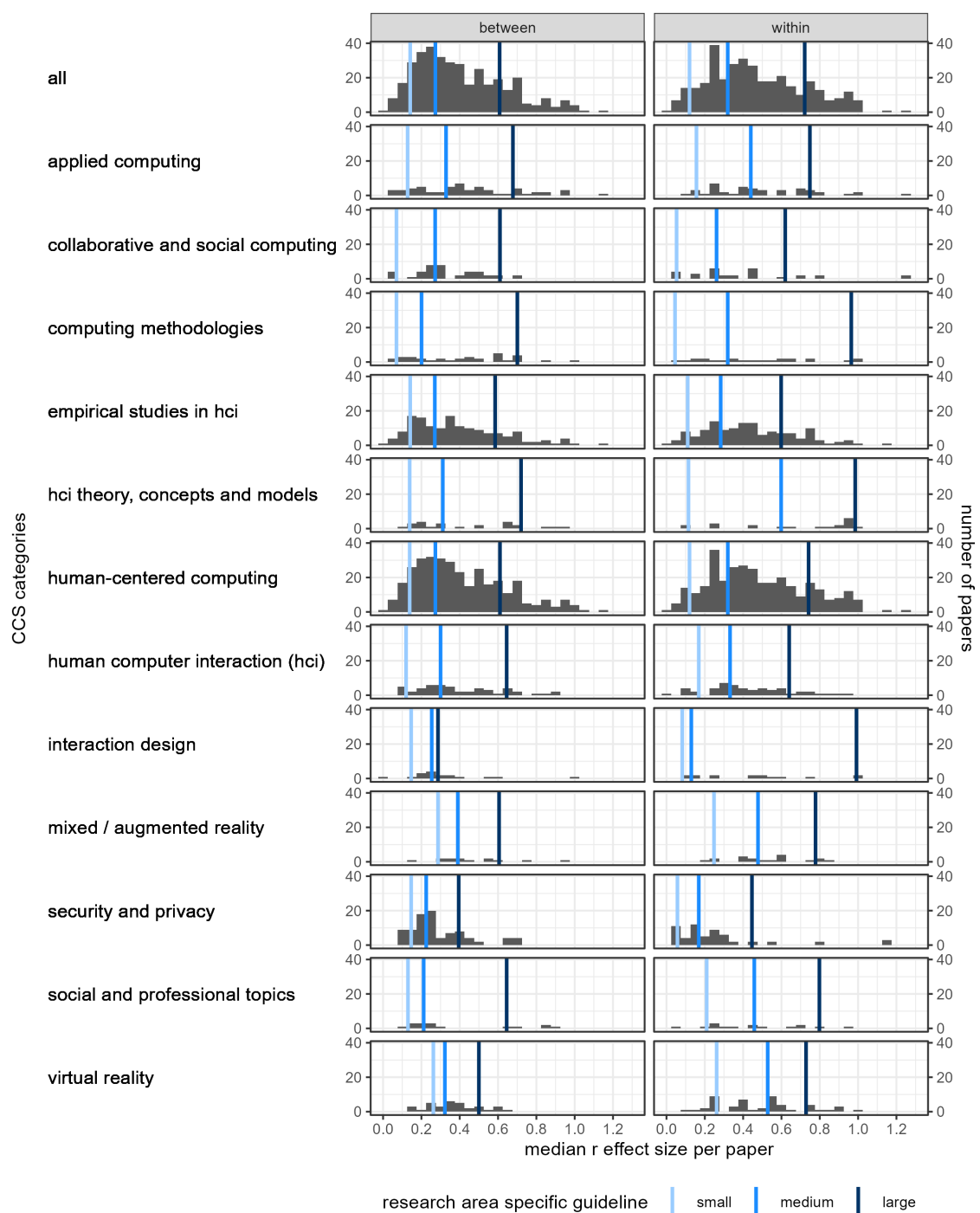


Figure 5.2 – Histograms of median effect sizes (Pearson’s r) per paper, reported separately for different research areas and experiment designs. Research area specific size guidelines superimposed on the histograms.

research area	between groups				within groups			
	# Paper	sample size			# Paper	sample size		
		min	median	max		min	median	max
all	359	1	50.0	289 mill.	346	4	28.5	7010271
applied computing	57	3	67.5	289 mill.	43	6	40.0	100000
collaborative and social computing	41	20	195.0	289 mill.	30	27	70.0	7010271
computing method- ologies	33	20	44.0	2871	23	11	40.0	12000
empirical studies in hci	155	5	65.0	289 mill.	138	6	33.0	65903
hci theory, concepts and models	24	8	82.0	1690	20	10	33.0	1132
human computer interaction (hci)	50	12	61.0	1690	51	9	28.0	65903
human-centered computing	332	1	48.5	289 mill.	322	4	27.0	7010271
interaction design	17	12	66.0	1506	15	8	27.0	6690
mixed / augmented reality	13	10	24.0	66	18	7	20.0	40
security and pri- vacy	89	19	244.0	11953	58	16	437.0	7010271
social and profes- sional topics	16	24	577.5	26174	15	10	69.0	3155
virtual reality	33	8	32.0	240	50	9	21.5	9860

Table 5.2 – Median and range for sample sizes and number of papers per research area in the meta-study. Papers categorized as multiple research areas are included for all that are applicable.

Since sample size influences the size of effect that is detectable by a significance test, we also plotted the sample size distributions in Figure 5.3. We used both box plots to show summary statistics, and violin plots to give more information about the shape of the distribution. Plots display the sample sizes as used for our meta-analysis of effect sizes, e.g. the maximum sample size considered was 352.5, the cut-off for outliers according to Tukey [519] and all larger sample sizes were interpolated as this maximum sample size. This is visible in the slight overemphasis of distributions at that sample size. For most research areas, sample sizes are skewed towards small sample sizes, and are smaller for within groups designs, which makes sense since these have more power to detect effects, so fewer participants are needed. For *collaboration and social computing*, *security and privacy*, and *social and professional topics*, this skew is not as obvious and sample sizes are evenly spread or even skewed towards higher sample sizes. *Mixed/augmented reality* and *virtual reality* are also somewhat special cases, in that they exhibit comparatively lower sample sizes than other research areas, likely due to hardware and special equipment needed to conduct such studies, representing

effectsize measure	between groups			within groups		
	small	medium	large	small	medium	large
r, ω, η	0.14	0.27	0.61	0.12	0.32	0.72
$\eta^2, R^2, \omega^2, \epsilon^2$	0.02	0.07	0.37	0.01	0.10	0.52
Cohen's d	0.22	0.44	1.09	0.19	0.52	1.38
OR	1.48	2.13	5.90	1.39	2.44	8.72
CLES	0.54	0.59	0.71	0.54	0.60	0.76
Cohen's f	0.11	0.22	0.55	0.10	0.26	0.69
Cramer's V	0.10	0.19	0.42	0.08	0.22	0.50
f^2	0.02	0.08	0.59	0.01	0.11	1.08
Kendall's τ	0.09	0.17	0.41	0.08	0.20	0.51
Spearman's ρ	0.13	0.26	0.58	0.11	0.30	0.69

Table 5.3 – Guidelines including all research areas and for all effect size measures. The number of included papers is 359 for between-groups designs and 346 for within-groups designs. The median sample size is 50 for between-groups designs and 28.5 for within-groups designs.

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	0.14	0.27	0.61	0.12	0.32	0.72
applied computing	0.13	0.33	0.68	0.16	0.44	0.75
collaborative and social computing	0.07	0.27	0.61	0.05	0.26	0.62
computing methodologies	0.07	0.20	0.70	0.04	0.32	0.96
empirical studies in hci	0.14	0.27	0.58	0.11	0.28	0.60
hci theory, concepts and models	0.14	0.31	0.72	0.12	0.60	0.98
human computer interaction (hci)	0.12	0.30	0.64	0.17	0.33	0.64
human-centered computing	0.14	0.27	0.61	0.12	0.32	0.74
interaction design	0.15	0.25	0.29	0.08	0.13	0.99
mixed / augmented reality	0.29	0.39	0.60	0.25	0.48	0.78
security and privacy	0.15	0.22	0.40	0.06	0.17	0.45
social and professional topics	0.13	0.21	0.64	0.21	0.46	0.80
virtual reality	0.26	0.32	0.50	0.26	0.53	0.73

Table 5.4 – Guidelines for separate research areas, using Pearson's r

a limit on the number of participants it is possible to recruit.

Combining insights from both effect size and sample size distributions, it seems that effect sizes in some fields, like *security and privacy* are smaller than in others, even though these fields have access to comparatively large sample sizes, which would enable them to detect both small and large results. However, there is no clear pattern based on sample sizes, given that the effect size thresholds for *mixed/augmented reality*, which primarily has smaller sample sizes, are not higher than for the rest of CHI. Figure 5.4 shows the relationship of sample size and effect size for all papers included in the meta-study. There are negative correlations which would be judged as between medium and small in the context of all analyzed CHI papers, see Table 5.3. This may be because in some papers, effect sizes were only reported for significant results, so small effect sizes were possibly under-reported, effect sizes from small studies might be inflated, and it could also be an indicator for publication bias. However, around the median sample size (50 for between-groups and 28.5 for within-groups designs) effect sizes are distributed across the whole spectrum from very small to very large. In comparison, an analysis of 447 papers in psychology found a correlation of $r = -.48$, $CI_{95\%} = [-.56, -.37]$ ⁵. When compared to the results in Figure 5.4, the negative correlation in HCI is noticeably weaker than in psychology [287], such that the CIs hardly overlap.

For use in a priori power analysis, these guidelines should only be a starting point for researchers' estimation. Researchers in areas where large effect sizes are common and thus small samples are sufficient, should consider whether detecting smaller effect sizes might also be of practical relevance. Researchers in areas where small effect sizes are common should not be discouraged from conducting studies with small sample sizes if they believe the effect they aim to detect is large, but of course if possible go for larger samples.

While our research area specific guidelines can help researchers interpret their results in the context of typical effect sizes of their area better than Cohen's guidelines, we do not suggest using them blindly. They should serve as context but further qualitative interpretation is very valuable.

5.4 Application of Findings

In the following we describe how the guidelines and findings in our work can be used by researchers.

- 1) Researchers planning a study have the study design figured out and now need to decide how many participants they want to recruit.
- 2) They know which statistical tests they want to use to analyze the data, and which type of effect size to report with it. They can use various ways to decide on an appropriate sample size, including resource constraints, prior work or deciding on the minimal effect size of interest [293].
- 3) For this example, let's assume the researchers' topic is very novel, so there is no direct related work they can use to determine what effect size to expect.
- 4) Now there are two options. One, if based on their expert judgment, they have a minimum relevant effect they want to find, they can use our tables to check whether this effect size is in line with other effect sizes in the area of the planned study. Two, if the researchers do not have this

⁵The paper originally reported a Spearman correlation of $r_s = -.45$, $CI_{95\%} = [-.53, -.35]$, which we converted using the formula in Table 5.1.

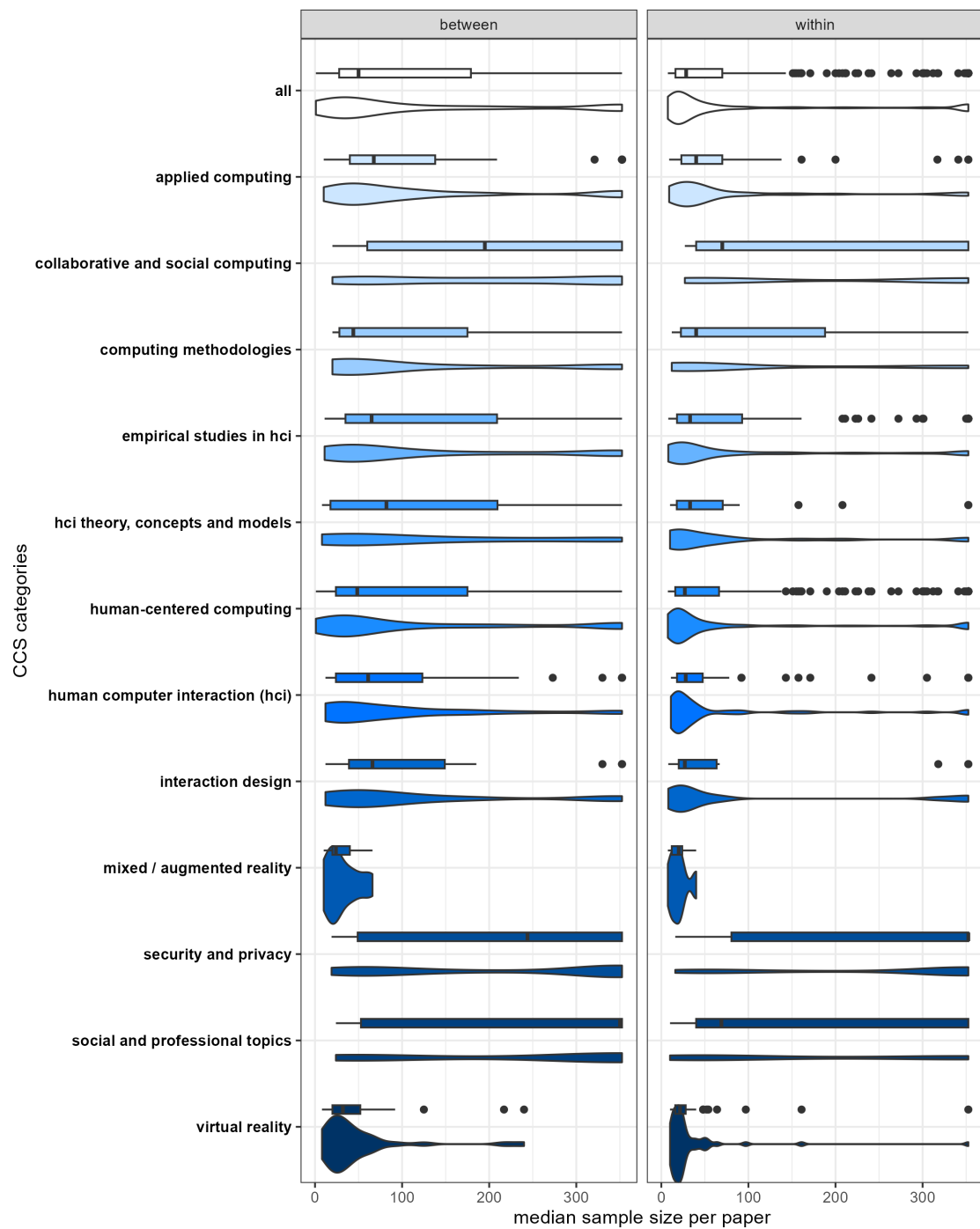


Figure 5.3 – Violin plots and box plots displaying the distribution of median sample sizes per paper, as used in the meta-study, separately for different research areas and experiment designs.

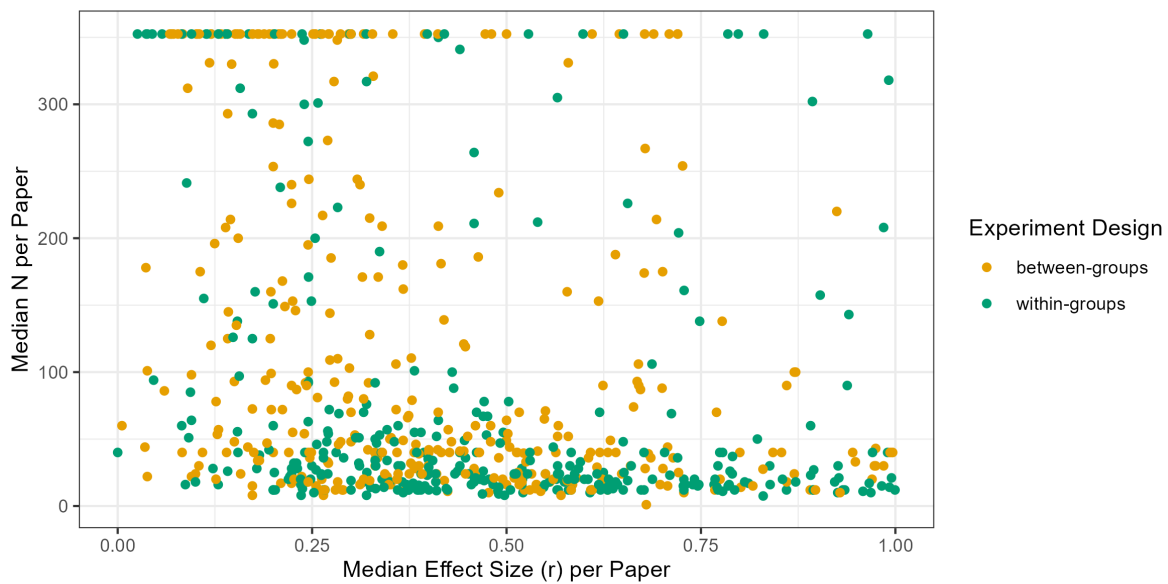


Figure 5.4 – Effect size plotted against sample size per paper and experiment design. For between-groups designs, the Pearson correlation was $r=-0.29$, $CI_{95\%}=[-0.38, -0.19]$, and for within-groups designs, it was $r=-0.24$, $CI_{95\%}=[-0.34, -0.14]$. Outliers in sample size are plotted at 352.5 (value used as maximum weight in meta-study).

intuition, they can simply use our tables to see what small, medium and large effect might be in their area. If the researchers' study domain is not specifically covered by the guidelines, e.g. *health*, they could instead look at a more general category, like *empirical studies in hci*. The idea is that researchers always take the most specific guideline as possible.

5) Should researchers see that the effect size they are targeting is smaller than what is typically found in their field, they need to carefully assess whether their experimental set-up is capable of detecting such small effects, e.g. by including mechanisms to reduce random errors or ensuring their sample size is large enough.

6) Based on this, the researchers can now conduct a power analysis, e.g. as explained for simple tests in the tutorials accompanying [369] ⁶ or for more complex designs by [71, 294].

7) Once the study is completed, and the authors have their effect sizes, they can use our tables to discuss them in the context of the related effect sizes in their field. While this step is the last step in the tutorial, we believe it is the most important step, because this significantly eases interpretation of results. Using different interpretation strategies highlighted in Chapter 6, they can e.g. assign size labels to contextualize the effect size or discuss practical consequences of the effect they found on their population of interest.

Should researchers be conducting studies with different effect size measures, the table with the most similar effect size measure can be used as a starting point. E.g. for Hedge's g , Cohen's d values can be adjusted, given that Hedge's g is usually a bit lower than Cohen's d , since it is less biased [291]. However, this is less reliable than our empirical analysis for the effect sizes in this paper. Future work can look at explicitly extending our work to more effect size measures.

⁶<https://powerdb.info/tutorials>

5.5 Conclusion

In this work, we extracted statistics from quantitative CHI papers published between 2019 and 2023 and used them to derive effect size guidelines specific to 12 research areas at CHI. Researchers can use these guidelines to interpret their results within their research area or as a starting point for a-priori power analysis.

Our guidelines for interpreting effect sizes in the different CHI research areas need to be updated on a regular basis to reflect new developments in the field. To ease this process and make it more robust, we recommend that statistical libraries generate machine-readable and complete data of statistical tests, including subgroup sample sizes, effect sizes, confidence intervals, test names, test statistics and descriptive statistics, which can be published as supplemental material with little extra effort for the authors. With this, future general meta-studies could be fully automated.

Chapter 6

Interpretation of Effect Sizes

Disclaimer:

The contents of this chapter are part of two publications:

1. “Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation”, published at CHI 2025. This paper is the result of a collaboration with my co-authors Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas and Matthew Smith and additional students involved in the data extraction, collection and analysis: Dorian Heidorn, Lysander Lenzen, Maximilian Mundt, Antonia Sistig and Lukas Wenz. Maximilian, Lisa and I conducted an initial qualitative analysis of researchers’ interpretation strategies in a sample of quantitative CHI papers from 2023 and Lisa and I finalized and refined the qualitative analysis. I wrote the initial draft of the section on effect size interpretation and revised it with Lisa (especially the qualitative analysis, results and discussion) and Matthew. I prepared the paper for publication with Florin, Lisa and Matthew.
2. “A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen’s d Effect Sizes”, published at CHI 2025. This paper is the result of a collaboration with my co-authors Julia Angelika Grohs, Simon Lenau, and Matthew Smith. Matthew and I conceived the research questions and idea. Julia and I designed the interview and survey study. The study design was discussed at a group meeting with Matthew and other members of our research group. I conducted the interviews and implemented the survey. Julia assisted me with recruitment and conducting the survey and interview studies at the CHI and SOUPS conferences. Julia and I analyzed the qualitative data and I analyzed and visualized the quantitative data. I conducted an additional interview with Simon in his role as a participant. Julia and I analyzed this interview and incorporated it in the results. I wrote the paper and iteratively revised it with Julia, Simon and Matthew to prepare it for publication.

6.1 Motivation

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 HCI in general [253, 272], or USP [189] 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 effect sizes available for a statistical test like an ANOVA: e.g. generalized or partial η^2 , ω^2 , and f^2 [291]. 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 [32], 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 [32], 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 [136] and may be a reason why they are not used frequently [189, 253]. 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 [91]. 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.

In this chapter, 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 [430, 421]. 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 [122]. 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.

We investigate effect size interpretation from multiple angles. To assess how effect sizes are interpreted in published work, we analyzed effect size discussions in papers from CHI’23 and identified five categories of approaches taken by researchers.

In addition, we explore how researchers interpret effect sizes they encounter when reading about study results. We conducted interviews and surveys at CHI and 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 [69, 200, 328, 433], but not used as much. For the interview and survey studies, we started by investigating the following research questions:

RQ-ES-1: Which factors influence researchers' judgments of effects

RQ-ES-1.1 regarding the size of effects?

RQ-ES-1.2 regarding the importance of effects?

We use aspects of constructive grounded theory [81], e.g. adapting our survey and interview methods to investigate new questions arising during the data collection. The surveys focused on RQ-ES-1.1 and RQ-ES-1.2 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 RQ-ES-1. In the second interview study, we focused on a new research question:

RQ-ES-2: 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 [91]. 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. In addition, we discuss the approaches to effect size interpretation taken by the authors of the CHI papers we analyzed and offer a framework to help authors choose which approaches are appropriate for their statistical analysis.

6.2 Effect Size Interpretation in Published Work

While the research area specific guidelines presented in Chapter 5 can help researchers interpret their results in the context of typical effect sizes of their area better than Cohen's guidelines, we do not suggest using them blindly. They should serve as context but further qualitative interpretation

is very valuable. We conducted a qualitative analysis of effect size interpretation in quantitative CHI papers published in 2023 and identified five categories of approaches employed by authors to interpret effects and effect sizes. Currently, effect sizes are rarely used to interpret results in depth. The most common form of effect discussion is the use of simple comparisons of descriptive statistics. In the rare cases of a more qualitative interpretation e.g. a discussion of practical relevance, the incorporation of effect size values is even less frequent.

6.2.1 Sample Selection

For our qualitative analysis, we chose CHI'23 as its proceedings were the most recently published ones available at the start of our analysis. We randomly sampled quantitative papers as categorized by the filtering described in Section 5.2.1. We employed saturation regarding the interpretation approaches as a stopping criterion, i.e. we continued sampling as long as we were identifying new approaches. We coded all approaches we found, even if they only occurred once. We stopped after not identifying any new approaches in 5 papers. We then analyzed 5 additional papers to validate the approaches we had identified. Three papers that contained statistics but were *largely qualitative papers* were excluded from this process. This resulted in 67 analyzed papers. We provide a list of our sample in Appendix E.1.

6.2.2 Data Analysis

In our coding process, we focused on two main goals. We refer to the first goal as *reporting coding*. In this step of our process, we used deductive coding to answer the following questions for the papers we analyzed:

- Are (standardized) effect sizes reported with hypothesis tests?
- If yes: Which effect sizes are reported?
- Is there a size judgment reported with the effect size?

Two coders (C1 and C2) established a codebook based on these research questions and an initial coding of 15 papers. We coded p-values, standardized effect sizes and size judgments for the papers. We considered this simple coding, as there are clear true and false answers [365]. As such, after establishing a shared codebook for this aspect of the coding process, a single researcher analyzed each paper. If they were unsure about something, they discussed it with the other researchers.

We refer to our second goal as *interpretation coding*. In this step, we used inductive coding to answer the question

- How are effect sizes interpreted?

C1 and C2 established an initial codebook based on the same 15 papers. They and one additional coder C3 then coded individually. We consider this aspect to be complex coding, as it leaves more room for interpretation [365], so C1 and C3 went over all the coded instances of interpretations,

verified them and grouped them to identify approaches. We set a rather low threshold for what we considered interpretation. Authors had to go beyond merely reporting the standardized effect size as is, or alongside descriptive statistics. Any further elaboration, e.g. specifying the direction of an effect or rephrasing to explain the effect size, was considered interpretation.

We additionally collected the following information about each paper:

Type of paper Within the papers considered quantitative by our filters, we identified four flavors: *NHST-quantitative*, *method*, *largely qualitative* and *Bayesian*.

Has descriptive statistics For each paper we noted whether it reported descriptive statistics with at least one hypothesis test. We considered descriptive statistics to be e.g. means, standard deviations or occurrence counts, depending on the hypothesis test.

6.2.3 Limitations

Our qualitative evaluation of effect size interpretation is limited by the expertise of the analyzing authors in the various research areas at CHI. While we use NHST in our own research and thus can interpret statistical results, due to our random selection of papers from across research areas, we are not equally familiar with the common measures, terminology and goals of each research area. When analyzing interpretation approaches, the extent of interpretation necessary to arrive at a conclusion was thus not always straightforward. However, we derived approaches to interpretation that were applicable across research domains.

6.2.4 Types of Papers

During our qualitative analysis, we identified four types of paper within our sample. *NHST-quantitative* papers were the expected default. These papers reported quantitative or mixed-methods results for HCI related research questions. They contained a substantial amount of hypothesis testing following the NHST paradigm. We applied both reporting and *interpretation coding* to these papers. *Method* papers developed new methods in HCI, such as development or validation of new measurement scales or discussing of analysis methods, without applying them to a content-related research question. *Bayesian* papers used only or mainly Bayesian statistics in reporting the results for content-related research questions. As we focus our contributions on providing guidance for choosing effect sizes for power analysis and interpreting results in HCI, i.e. for content-related studies in contrast to *method papers*, and as *Bayesian* papers report different measures of effect sizes, we only applied the *interpretation coding* to these papers and did not apply the *reporting coding*. *Largely qualitative* papers only report hypothesis tests which are not relevant for answering the main content-related research questions, e.g. Jacobsen et al. [239] employed two tests to detect shifts in interaction frequency across households, but otherwise focused on findings from their qualitative analysis. The tests were not discussed further, and we did not consider these papers in our analysis.

6.2.5 Overview of Effect Size Reporting

We analyzed a total of 67 papers. Of these, we categorized 57 as *NHST-quantitative*, 6 and 1 as *method* and *Bayesian* papers respectively and 3 as *largely qualitative*. Of the 57 *NHST-quantitative* papers, 30 reported at least one standardized effect size, and 10 reported effect sizes accompanied by a size judgment at least once. 2 of the papers pointed out a source for their size judgments. 42 of 57 *NHST-quantitative* papers contained descriptive statistics alongside at least one test.

6.2.6 Interpretation Approaches

We identified five categories of approaches used by CHI'23 authors in interpreting effects and effect sizes in their work. We situate them in a spectrum from descriptive approaches, which provide additional, often numerical and objective information to interpret effect sizes, to interpretative approaches, which include in-depth reasoning beyond the immediate scope of the work in question. Figure 6.1 shows this spectrum, while Table 6.1 provides specific examples from our sample of analyzed papers for each of the five categories. In the following, we describe these approaches in more detail.

On our spectrum from descriptive to interpretative approaches to interpreting effect sizes, the most descriptive category is *comparing values*, which are part of the output from the statistical analysis. The simplest is stating the *direction of the effect*, e.g. when comparing groups, which group was faster or made less mistakes. Reporting the effect size *in units of the study* is a bit more involved than just reporting descriptive statistics, i.e. reporting the difference in group means in the units in which it was measured instead of just the group means. This is a form of non-standardized, or simple effect size. Differences can also be compared *in relative units*, using phrases like “20% faster”. This abstracts away the units of the simple effect size, but is still close to the measurement methods of the study. However, using such relative units, e.g. percentages, can be confusing as well, as Dragicevic has elaborated specifically for reporting of speed differences [124]. He recommends reporting ratios or percent differences if relative comparisons are required [124]. If standardized effect sizes are reported, then *comparing effect sizes* within a study, or to a set threshold, is another approach we saw. In this approach, the effect sizes themselves are included in the comparison, rather than descriptive statistics.

Assigning labels can be considered descriptive or interpretative, depending on how researchers address them. ‘Labels’ in this context refer to size judgments, e.g., researchers denoting an effect as having a specific size. *Making size judgments using guidelines* was the more common approach, using guidelines for determining a small, medium or large effect, such as those by Cohen [91], even if these guidelines were not always explicitly named. This approach falls on the descriptive end of the spectrum. However, when *making individual size judgments*, the judgments are not obviously based on published guidelines. If they are explained in the context of the study, we deem these judgments to be interpretative.

Some researchers are *using a statistical and methodological reference frame*, i.e., they reference statistical or methodological decisions or consequences in their interpretation of effect sizes. When *following up on non-significant effects* researchers use other descriptive approaches of interpreting effect sizes to get around the binary notion of statistical significance and highlight differences in their results despite non-significant tests. Researchers *referencing sample size* interpreted effect size

Approach category	Approach	Example Quote
Comparing test-specific values	Direction of effect	“However, writing with multiple suggestions leads to slightly longer texts.” [104]
	In absolute units of the study	“[...] significantly increased the number of words in a sentence after which suggestions were requested – by about 1.5 words (Means: $s = 10.93$, $s = 9.48$; i.e. a relative increase of 15.3%)” [104] [emphasis added]
	In relative units	“[...] significantly increased the number of words in a sentence after which suggestions were requested – by about 1.5 words (Means: $s = 10.93$, $s = 9.48$; i.e. a relative increase of 15.3%)” [104] [emphasis added]
	Comparing effect sizes	“Also, the correlation coefficient between profile type and recommendation ratings was higher (and sometimes statistically significant) for museum users.” [583]
Assigning size labels	Making size judgments using guidelines	“Both plank (static core) and press-ups (dynamic upper) showed only non-significant, small effect sizes (Cohen’s $d = 0.239$ and 0.298).” [87]
	Making individual size judgments	“These differences are rather small, about 6-10 words” [104]
Using a statistical and methodological reference frame	Following up on non-significant effects	“Finally, it is worth nothing that in both cases the effect is in the right direction ($F > P$), however the effect sizes are small, meaning that the study would have been under-powered to detect such small changes” [87]
	Referencing sample size	“However, since our sample size is relatively small ($N=24$ after two removed samples), there are limitations regarding the certainty of this observation” [453]
	Making methodological decisions	“This means that the model is able to explain 99.98% of the shared variance between the Hexad-12 and the Hexad-24. This result clearly shows that the two variable sets are not independent. Thus, we can continue analysing the results of the dimension reduction analysis to check whether the predictor variables load on the same CF as the criterion variables.” [283]
Comparing different observations	Comparing with related work	“Our findings replicate [41] that individual mood does not change much with effective peer counseling, but contrasts with those of [3]” [549]
	Triangulating within work	“for concurrent joint attention, the charts area of interest (Figure 6a) all tree conditions perform equally well [sic]. This result is consistent with the Q6 question of the questionnaire, which indicates no statistical difference across the experimental conditions. Qualitative analysis of interviews suggests that this result could be due to the effectiveness of verbal communication in refining and specifying the area of interest ” [61]
Interpretation for the big picture	Possible explanations for effects	“We believe there are two potential explanations. First, our study procedure is straightforward. An improved consent form reading may not have a strong effect in preparing the participants for the later study. Second, our study is low-stake compared to medical trials.” [579]
	Practical consequences or limitations of effects	“Although our MR system helped to improve the success rate and precision of the deformation, 4Doodle has limited accuracy compared to machine printing because of the uncertainty associated with human performance.” [503]

Table 6.1 – Examples of interpretation approaches. All citations within the quotes refer to the bibliography of the original paper.

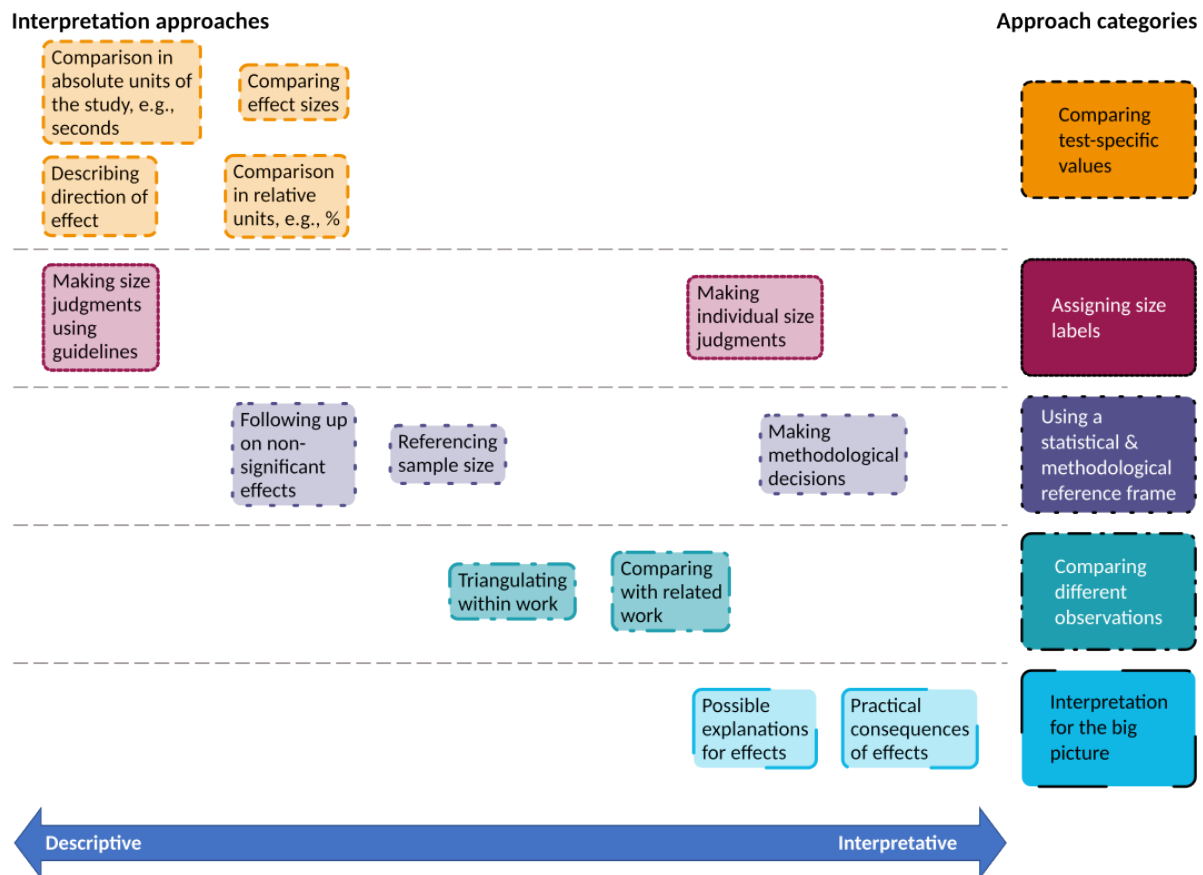


Figure 6.1 – Visualization of the spectrum from descriptive to interpretative interpretation approaches. We display approaches and their respective category on a single horizontal level each. The distance to the x-axis does not carry meaning otherwise.

in the light of the sample size of their study, especially relating to under-powered or over-powered hypothesis tests. Finally, effect sizes were used to *make methodological decisions*, e.g., regarding model choice, predictor selection and merging of data sets within a study, which we consider more interpretative.

Comparing different observations involves comparing results from a different data set, in contrast to comparing individual values within a data set. This can happen as *triangulation within work*, where other data sources presented in the same study are used to back up quantitative results. In our sample, these sources were mainly qualitative or descriptive. Another approach is *comparing with related work*, where the comparison is outside the study currently being reported, with published related work. Both comparison approaches can show similarities, where the different observations support the current result, or comprise of contradictory results. We allocate these approaches to the middle of the spectrum between descriptive and interpretative approaches. Comparisons can reproduce binary judgments about outcomes, e.g. if an intervention is effective or not, in which case they are more descriptive, or they can go into more interpretative depth regarding connections between the observations and comparing actual sizes of effects.

On the interpretative side of the spectrum, researchers provide *interpretation for the big picture*,

i.e. beyond the current study, as context around effects and effect sizes. They speculate on *possible explanations for effects*, e.g. based on other statistical results, common knowledge, or prior work. Researchers also discuss *practical consequences of effects* for further research and theory, but also deployment of newly developed techniques in practice or in real-world scenarios.

In considering which approach to use, we do not believe that one category or one side of the spectrum is better than the other. Instead, they are suitable for different use cases. The more descriptive approaches in our spectrum are easier to implement, as values to compare may be provided automatically from statistical analysis software, or can be calculated based on existing data. As such these approaches can be used for all or most of the statistical tests reported in a paper. Since descriptive approaches refer to a single concrete effect, they should be reported close to the hypothesis tests.

More interpretative approaches require more effort and knowledge of a field and are thus more dependent on individual researchers' backgrounds and judgments, in addition to needing more space to describe implications. They often bring up prior work and implications for the future or practicality of deployment to assess the relevance of not just a single effect size, but multiple test results taken together for a big picture. Anvari et al. provide a thorough summary of influencing factors on practical relevance with examples from psychological research [19]. When judging practical consequences of effects, even small effects can be immensely beneficial if many people are affected [19] and no larger effect interventions exist or are practical. On the other hand, the importance of large effects can be diminished through practicality or interactions with other effects or not being as generalizable as anticipated [19]. This is especially important for the main hypotheses and contributions of a paper. Interpretative approaches fit in the discussion, where a broader context can be taken into account. However, to maintain a grounding of the interpretation in the statistical results, numbers from effect size calculations should be referenced in such discussions, instead of a binary statement that an effect exists or not based on statistical significance.

In addition, not all approaches are equally applicable to all types of research. This applies especially to the category *Using a statistical & methodological reference frame*, wherein e.g. effect sizes can only be used to *make methodological decisions*, if testing was conducted for this purpose and if results are statistically significant, then *following up on non-significant effects* is not necessary. When deciding which approach to use, researchers should thus take into account one that is suitable for their goal.

6.3 Studies to Investigate How Researchers Interpret and Understand Effect Sizes

To investigate effect size interpretation from a different angle than from published work, we conducted two iterations of a study, each consisting of a survey (-s) and an interview (-i) study, to get multiple angles of insight into how USP and HCI researchers interpret and understand effect sizes, specifically 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 6.2 has

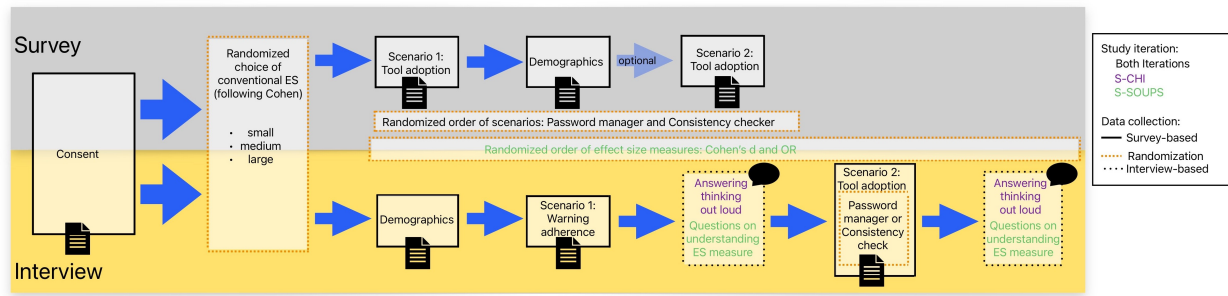


Figure 6.2 – Process of survey and interview studies

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 6.2 – Overview of the studies presented in this work

an overview of the study iterations and the research questions at the focus of each and Figure 6.2 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.

6.3.1 Positionality

Researchers' experience, knowledge, and beliefs contribute to the interpretation and planning of research [444, 163]. 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 work. 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.

6.3.2 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.

6.3.3 Measurement

We were interested in several concepts¹. We asked participants for a **judgment of importance** about the presented effect, distinguishing whether it was *important* or *not important*. We also measured 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 6.3.4). To see whether participants' views on **criticality** matched our own, we asked

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

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 [91]. We used these despite criticism of the guidelines [448, 265], 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.

6.3.4 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 [488, 203], 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 6.3 and the full text of all varieties can be found in the supplemental material.

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. 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. 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. 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.	Research question and study design Descriptive statistics Inferential statistics Effect size explanation - only present in S-CHI

Figure 6.3 – 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.

6.3.5 Conventional Effect Size

We chose values for small, medium and large effect sizes according to Cohen [91], 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 [193] with 1.18 as the recommended average correction factor [399].

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 [181]. 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.

6.3.6 Surveys and Interviews

The study process for survey and interview participants is shown in Figure 6.2. 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 (RQ-ES-1.2), we removed the threshold questions for the judgment of size and the corresponding confidence questions from the survey.

6.3.7 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 [474], 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.

6.3.8 Participants

Information on the participants is in Table 6.3. 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.

6.3.9 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 [407], 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.

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 6.3 – demographics of study participants

6.3.10 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 6.3.13.

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 (RQ-ES-2): *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 (RQ-ES-2), 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 RQ-ES-1.2. 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.

6.3.11 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 6.3. 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.

6.3.12 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.

6.3.13 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 (RQ-ES-1.1), factors influencing participants' judgments of the importance of effects (RQ-ES-1.2) and comprehension of the effect size measures (RQ-ES-2). For the analysis of the survey data for RQ-ES-1, 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 (RQ-ES-2) we retained all 60 responses.

6.3.13.1 Understanding of Effect Size (RQ-ES-2)

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 6.3. 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, ϕ , 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 smaller than 1. If there is no effect at all, $OR=1$ and Cohen’s $d = 0$.

Among the three thresholds each participant submitted, there were 26 values that indicate misconceptions. These include five 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 of these 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 6.4. 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 right away, instead of later in the interview, to avoid causing frustration.

In S-SOUPS-i, we recognized some misconceptions from the surveys when prompting our partic-

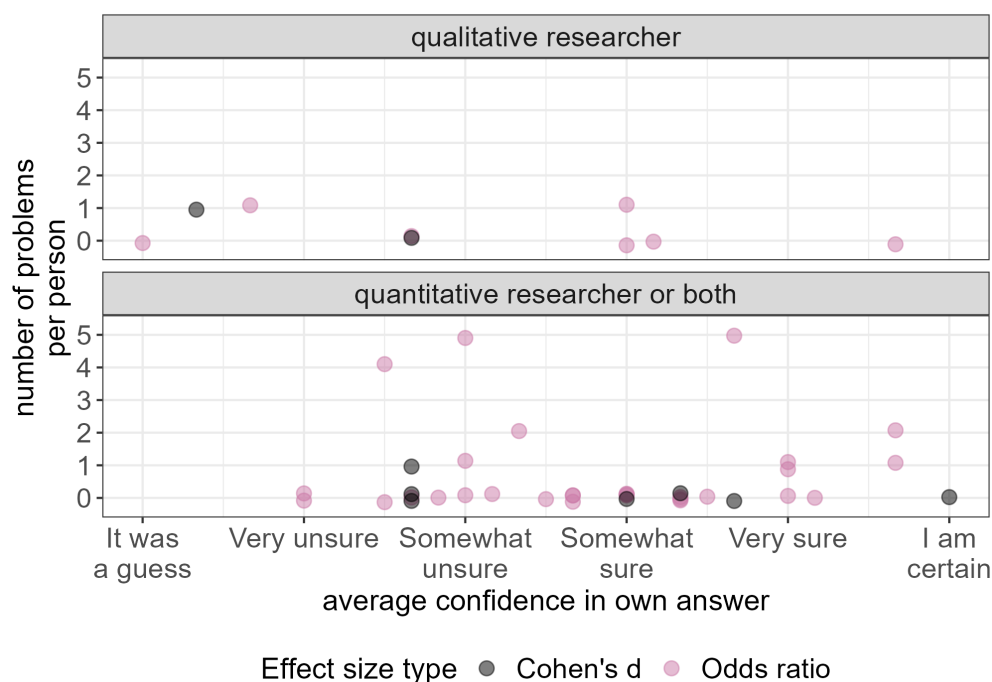


Figure 6.4 – Number of identified problems in submitted Cohen’s d or OR thresholds per participant vs. average confidence rating per participant over all submitted thresholds, grouped by participants’ research focus. We juxtaposed qualitative researchers with all other research focuses. This includes 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.

ipants 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 $\frac{2/2}{1/3} = 3$, while the risk ratio is $\frac{2/4}{1/4} = 2$. Higher base rates of outcomes lead to larger differences between odds ratios and risk ratios [385]. This misconception is well-known in medicine and public health research [222, 385]. 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 [295]. For each effect size measure, at least one participant also saw a direct relationship between the

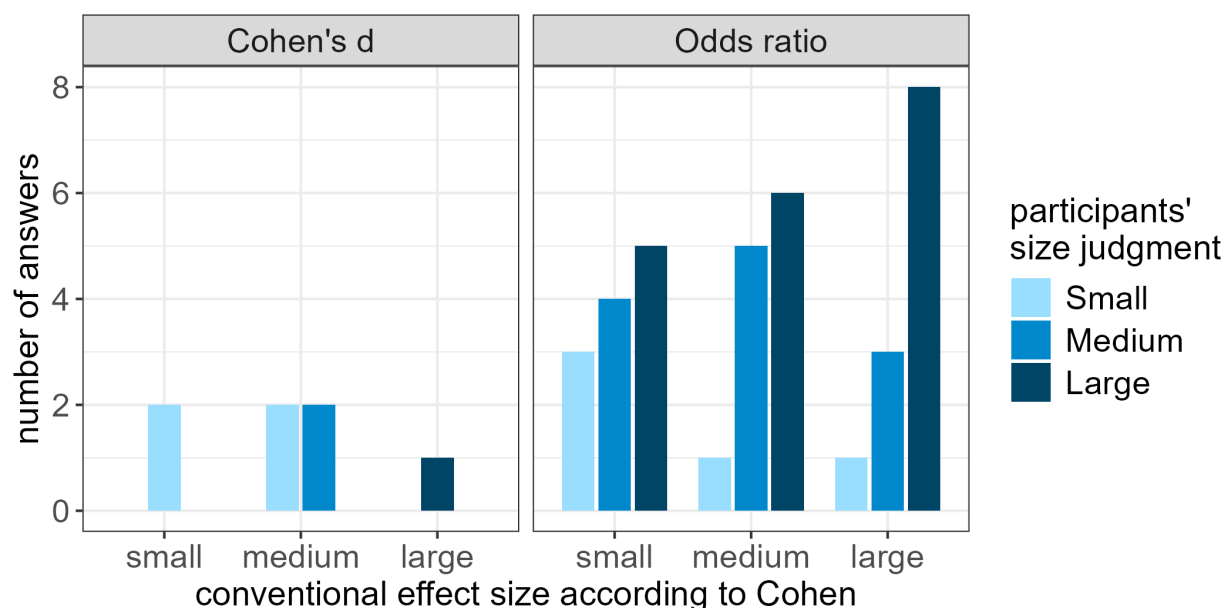


Figure 6.5 – Comparison of two ways to categorize effect size: participants’ judgment as selected from the options in the survey (depicted by different color, see legend), and Cohen’s conventions (x-axis), converted for OR (right)

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 behavioral sciences [212], and especially for specialized populations like software developers, administrators, researchers, etc. [344, 5, 39]. 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.

6.3.13.2 Judgment of Size of Effects (RQ-ES-1.1)

Participants’ judgment of size of effects in the surveys is compared to Cohen’s effect size conventions [91] in Figure 6.5. 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

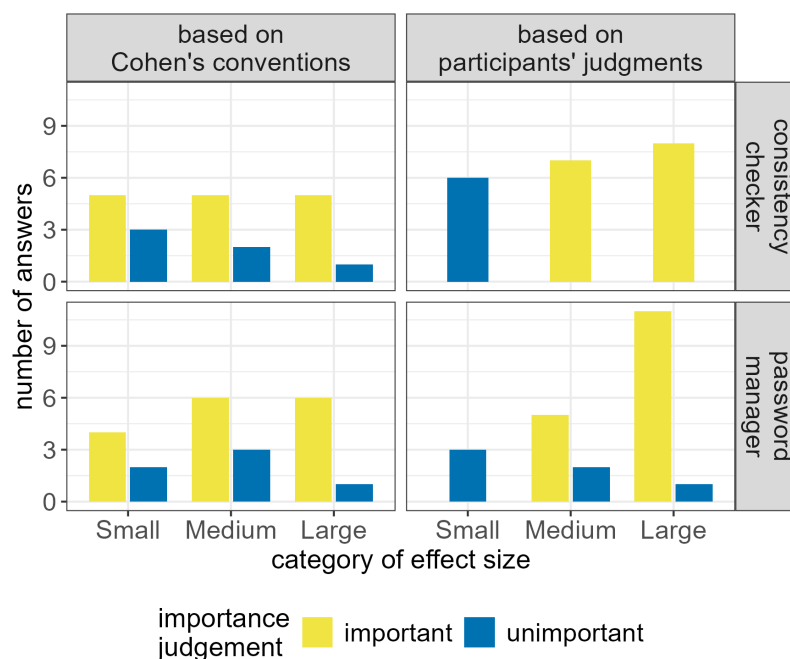


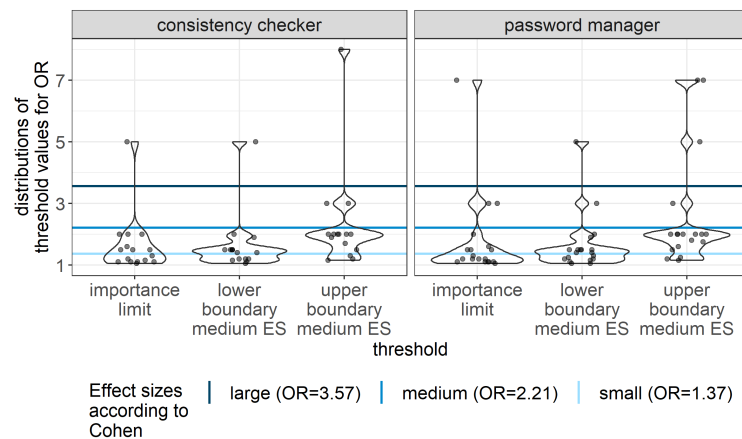
Figure 6.6 – Comparison of judgment of importance based on Cohen's conventions (converted for OR, on the left) and participants' judgment of size (right), grouped by scenario

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 [91], p.532.

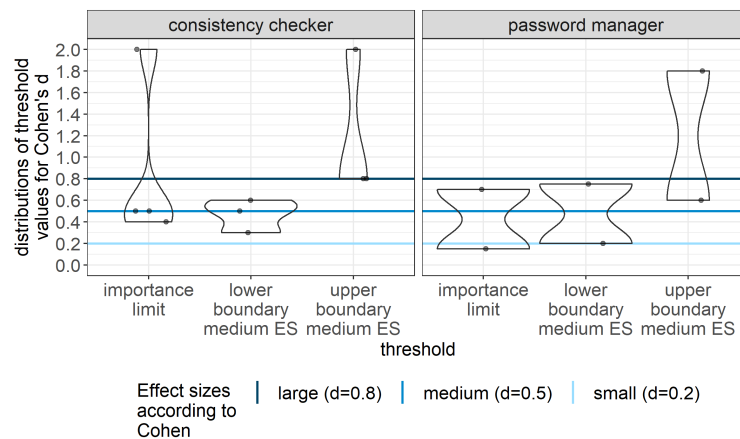
6.3.13.3 Judgment of Importance of Effects (RQ-ES-1.2)

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).

Size Based on the surveys, the right side of Figure 6.6 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 6.8), 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 6.6, too. The data



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

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 6.7a) and Cohen's d (Figure 6.7b), 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 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.

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 6.6 shows that our participants saw things differently and highlights our personal bias in the choice of scenarios. For both scenarios, 6 participants 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 6 in the password manager scenario, but 6 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 6.8). 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 6.7, 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.

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.

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



Figure 6.8 – Participants judgments of effect size and importance, grouped by scenario and research background. Sample: Only participants with misconceptions

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.

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 [181] 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 [26, 451]. 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.

6.3.13.4 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.

6.3.14 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 [93]. 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, being USP researchers ourselves. For example, this influenced our choice of scenarios. Nevertheless, we gained insights on factors influencing participants' judgments in different scenarios.

6.4 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 [499, 194, 574, 47]. 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 [447]. 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 [358]. Critique of the status quo of effect size reporting and interpretation is common across fields however, e.g. in exercise science [110], psychology [166] and education [36, 472], 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. The five categories of interpretation approaches we identified in our analysis of effect size reporting can provide a starting point for interpretation.

Our study also 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 [220], or using effect sizes which are easier to interpret [200, 69]. The binomial effect size display (BESD) [433] and the common language effect size (CLES) [328] 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 [69]. 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 [69]. Where effects are at risk to be overvalued, such as in research communities generally suffering from publication bias [234], this may be less desirable. Hanel and Mehler conducted a survey to compare subjective informativeness of significance statements without effect sizes to Cohen's U_3 , the probability of superiority, the overlapping coefficient, Bayes Factors and Cohen's d [200]. Cohen's U_3 was rated most informative, but surprisingly still less informative than significance statements without effect size information [200]. The authors attributed this finding to an exposure effect, given that participants were likely more familiar with mere significance statements rather than effect sizes [200]. 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 [91] 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.

6.5 Conclusion

We investigated interpretation of effect sizes in publications, more specifically in quantitative papers published at CHI'23. We identified five categories of approaches for interpretation and recommend that researchers provide additional descriptive interpretation for the majority of their reported hypothesis tests, and more in-depth qualitative interpretation including practical consequences of the main result. Additionally, 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.

Chapter 7

A Re-replication Study About Experts' and Non-experts' Security Practices and Advice

Disclaimer:

The contents of this chapter were published as part of the publication “Replication: "No one can hack my mind" - 10 years later: An update and outlook on experts' and non-experts' security practices and advice” at SOUPS 2025. This paper is the result of a collaboration with my co-authors Jenny Tang, Arthi Arumugam, Daniel Huschina, Lisa Geierhaas, Florin Martius, Luisa Jansen, Kolja von der Twer, Lilly Jungbluth, and Matthew Smith and additional students involved in the data collection and analysis: Carina Lübcke, Charlotte Starke and Charlotte Mädler. The project idea was conceived by Matthew and me, and I organized and supervised student work on different aspects of the project, and managed collaborators' contributions. The adaptations to the existing study design of the studies we replicate were discussed by Matthew and me for the surveys and by Florin, Lisa, Luisa and me for the interviews. I adjusted the surveys and interview guidelines to incorporate these changes. Florin and Luisa, Lisa and Charlotte Mädler and Lisa and Daniel conducted the interviews at three different venues. Daniel and Lilly initially transcribed the interviews and I corrected and anonymized the transcriptions. Daniel and Lilly conducted an initial analysis of the interviews and Daniel and I refined this analysis for the publication. Carina, Charlotte Starke, Matthew and I used social media and snowball sampling among our contacts to recruit experts for the survey. Recruitment of non-experts was managed by me through Prolific. Arthi, Kolja and I analyzed the freetext advice submitted in the non-expert survey. Carina, Charlotte Starke and I analyzed the freetext advice submitted in the expert survey. Jenny and I conducted descriptive and inductive analysis on expert and non-expert survey data, with Jenny focusing on amounts, frequencies and hypothesis tests and me focusing on data preparation, visualization, and effect sizes. I wrote an initial draft of the paper, and revised it with Daniel, Kolja, Lilly, Luisa and Florin contributing to the related work section, Arthi, Daniel, Florin, Kolja and Lisa writing the parts of the methods which were most relevant to their respective contribution, Jenny and Florin working on the results, and Matthew and Jenny providing further feedback throughout the revision. Analyses for the meta-scientific results were conducted and written down by me after the publication of the paper.

7.1 Motivation

Staying secure online is still a challenge to users today as in the early days of USP [566, 9, 389, 17]. Advice on staying secure has therefore been a much researched topic in our field [215, 140, 416, 171, 417, 414, 420, 63, 49, 338, 437]. In 2015, Ion et al. published a study comparing expert and non-expert security advice and behavior and found that experts and non-experts differ in their behavior, but also that experts do not strictly agree on a set of advice [236]. Their study was later replicated by Busse et al. and published in 2019 [74]. The replication showed that many of the challenges in the first study were still apparent 4 years later, but also suggested some development [74]. Research has shown that advice is still fragmented—making it difficult for non-expert users to prioritize between different advice, as well as actually putting advice into practice—as usability is still a hurdle to achieving security [417]. Now, almost ten years after the original “No one can hack my mind” study was published, we replicated the study again and took a look back at ten years of security advice and behavior by experts and non-experts. Specifically, we investigated the following research questions:

RQ-Rep-1: How has security behavior and advice developed over the past ten years?

RQ-Rep-2: How do experts and non-experts differ in their security behavior?

RQ-Rep-3: What are security experts' perspectives on the development of security behavior in the past and in the future?

Our work identifies a larger shift in security behavior and advice than the first replication from 2019 [74]. Over the duration of ten years, some behaviors, e.g., using 2FA, have seen a strong rise in popularity, while others, such as the use of antivirus, have become less relevant. This establishes a time frame for researchers to observe such developments. In addition to this main contribution, we provide an updated overview of current end-user security practices and discuss both challenges and successes for the USP community. Password manager use is an example of an on-going challenge, where expert advice diverges from end user practice. On the other hand, our study also revealed USP success stories, e.g., increased adoption of 2FA. Looking ahead, experts express a hope for more user-friendly and context-specific security advice in the future, such as a shift in focus away from computers towards the now more prevalent phones or tablets. They also support future developments moving toward security measures that are built into systems by default, circumventing user error, e.g., automatic updates. The USP community should monitor the changing status of security advice in these areas in longitudinally-oriented research. From a meta-scientific perspective, we additionally describe the development of study methodology, recruitment and analysis methodology over the three iterations of this study and showcase on-going and developing challenges in these areas, e.g. with increasing availability of AI changing recruitment of experts and non-experts. We discuss how future replications could deal with these challenges and support the goal of further replications in this area by providing a replication package on OSF.

7.2 Replicated Studies

We base our work on two studies, the original “No one can hack my mind” study from 2014 by Ion et al. [236], published in 2015, and its replication by Busse et al., conducted in 2018, published

in 2019 [74]. Each consisted of interviews with security experts in addition to surveys with security experts and non-experts on the topic of security behavior and advice. A comparison of study design and goals between these publications and our re-replication is in Table 7.1.

The original study by Ion et al. [236] highlighted that experts and non-experts approach online security differently: Experts rated some of the non-expert practices positively, but non-experts did not comply with other practices that received high ratings from experts, the most prominent of which being installing system updates, using a password manager and using two-factor authentication. They discussed more detailed results for four categories of advice: *software updates*, *antivirus software*, *password management*, and *mindfulness*. The replication by Busse et al. [74] showed that many of the topics relevant in the original work were largely unchanged in 2018. One development was that non-experts preferred to use browser extensions to block cookies rather than deleting them manually, compared to the original study. In their evaluation of the compound question asking for ratings of effectiveness and realism within a single question, they identified several areas where experts' effectiveness and realism ratings diverged: password security, two-factor authentication, links and attachments, and application updates. These areas emerged as areas for improvement for the USP community.

7.3 Method

We based our replication on materials published in prior work, as well as additional analysis materials, such as codebooks from Busse et al. [74] and the numbers from Figure 1 in the original study by Ion et al. [236], which the authors had subsequently shared with Busse et al. In the following, we describe the data collection and analysis process, highlighting differences to the published studies. We provide study and analysis materials, including code books, analysis scripts and anonymized data on OSF to support further replication efforts.¹

7.3.1 Positionality

Eight of the authors are part of a group doing research on usable security at a European university, including a professor and research group leader, and multiple PhD students and undergraduate students of computer science or IT security. Two other PhD student authors belong to different research groups, one at a North-American university in a group studying security and privacy, and one at a different European university, studying psychology of security and privacy. The group of authors collectively had experience with collecting and analyzing interview and survey data and conducting replication studies in USP before this study. All of the authors have experienced situations in which they took on an expert role to give security advice to non-expert friends or family members. As such, their personal opinion on advice may have influenced the coding process of security practices and advice.

¹https://osf.io/j43ws/?view_only=d2ab51af654e4cf78f35149f5a578fb7

		Ion et al. [236]	Busse et al. [74]	This paper
Expert Interviews	Recruitment	BlackHat, DefCon, USENIX Security	CeBIT international trade fair	BlackHat London, C3, USP event
	Security experience for inclusion	5+ years	not specified	1+ years
	Sample N	40	40	35
	Study goal	develop survey	evaluate survey design	update survey questions
Expert Survey	Recruitment	Google Security Blog, personal contacts	social media, personal contacts	social media, personal contacts, Freelancer.com
	Security experience for inclusion	5+ years	1+ years	1+ years
	Survey design	compound	both compound and separate (A/B design)	compound; added questions on new behaviors
Non-Expert Survey	Recruitment	Mturk	Mturk	Prolific
	Target demographic	US	US	US representative
	Survey design	separate	separate	separate; added questions on new behaviors

Table 7.1 – Methodological comparison for all studies by Ion et al. [236], by Busse et al. [74], and in this paper. Compound = compound questions used for advice rating; separate = separate questions used for rating effectiveness and realism of advice

7.3.2 Ethics

The studies were approved by the ERB of our institution. To ensure compliant handling of participants' data and adherence to the GDPR, we developed and followed a data management plan. All personal identifiers were either removed or pseudonymized to protect participant privacy. Data were stored on secure institutional servers with controlled access, deleted when no longer necessary, such as removing audio recordings after submission, and no personally identifiable information was shared beyond the research team.

7.3.3 Study Design

7.3.3.1 Expert Interviews

We conducted expert interviews at three security and privacy related events in Europe. The interviews were designed to take less than five minutes. All interviews began with the same question,

to ensure consistency and comparability across all participants: “What are the top 3 pieces of advice you would give to a non-tech-savvy user to stay safe online?”

The follow-up questions varied between the interviews depending on the event. The first event took place before our replication of the survey study. The participants (N=18) were shown results from the non-expert surveys from the prior studies [236, 74], compiled into a single figure. Then, we asked for an open-ended description of their first impression, and if applicable followed up with questions on their opinion of the current status quo and possible future developments.

We used the results of these interviews to refine our survey. After collecting the results from the survey, we conducted more interviews at two further events with N=11 and N=6 participants, respectively. This time, we showed participants the updated results, and again asked for their first impression. Follow-up questions asked for their opinion on the presented advice and for a perspective on possible changes in the future.

Following data minimization principles, we did not collect demographic information beyond participants’ years of experience in IT security-related fields, as this was our main criterion for sample selection, and the source of their experience in IT security. We ensured that all participants had at least one year of experience.

7.3.3.2 Surveys

The survey structure was similar to prior work, so we only provide a brief overview here. All participants gave informed consent before starting the survey. Non-expert participants then reported their top three security practices to stay safe online, and expert participants additionally submitted the top three pieces of advice they would give to a non-tech-savvy user. The second block of questions contained multiple-choice questions on participants’ own security-related practices, and in the third block, participants rated security advice intended for a non-tech-savvy audience. Non-experts rated advice separately based on effectiveness and realism to implement the advice, while experts gave one overall rating of advice quality, incorporating effectiveness and realism. Finally, all participants answered demographic questions.

Based on the initial round of expert interviews and prior work on new authentication practices [380, 178, 578], we expanded the second and third survey block. We primarily added questions about new authentication methods such as hardware tokens, authenticator apps, passkeys and biometrics but also a question on sharing personal information, and checking permissions in mobile applications. In the third block, we added a question about using large language models (LLMs) such as ChatGPT to get security advice. Based on more recent best-practice, we adjusted the question on gender to include a non-binary option and the option to self-describe [480].

Finally, we added measures for bot detection. This was a necessary addition to our surveys, given the rising popularity of LLMs like ChatGPT², which can also be used by crowd workers to answer survey questions [86, 241, 196] and evade existing methods to detect them, such as CAPTCHAs [393]. We attempted to detect such responses by including several experimental methods for those participant groups which were paid for their participation. These methods included questions hidden through either font color or JavaScript, similar to measures that had been tested in prior work [482, 221].

²<https://chatgpt.com/>

Non-expert participants received direct compensation for their participation, and thus might be more incentivized to use tools like LLMs to efficiently answer surveys. To counter this, we used additional optional knowledge questions—which would be hard to answer for a human—as a bot detection measure. We explicitly stated that participants should not use other resources for these questions, and that they did not have to answer them.

For the expert survey, we reverted to using the compound question to ask experts for quality judgment of advice, as in the original study [236], since the additional questions about security behavior made the survey more lengthy overall.

7.3.4 Recruitment

We followed Busse et al. [74] in including security experts with at least one year of experience in an security related field.

For the interviews, we recruited security experts from three venues in Europe. The first was a USP community event, which we choose not to name in this work to prevent potential deanonymization of our participants. The other two venues were large enough that this concern does not arise, namely the BlackHat Europe, and the 38th Chaos Communication Congress (38C3), which are both hacker conferences. Participants were not offered monetary compensation, but they received a stuffed toy as an incentive to take part.

For the non-expert surveys, we intended to recruit 300 non-expert participants on Amazon Mechanical Turk (MTurk) to replicate the prior studies as closely as possible, and additionally recruit a larger representative US-sample on Prolific to compare to the MTurk sample, given reports of declining data quality on MTurk [501, 496, 260]. However, due to the lack of data quality in our MTurk pilot sample (N=10), we elected to stop recruitment on MTurk and focus on Prolific instead. Non-expert participants were compensated with US-\$3.

Similar to Busse et al. [74], we used social media to recruit for the expert survey. We published a call for participation on the most senior researcher's LinkedIn. We attempted to use different subreddits for recruitment, but mods of subreddits like *r/computerscience* and *r/sysadmin* had implemented rules against posting surveys or posts not strictly fitting the topic of the subreddit, so we were only able to post on *r/takemysurvey*. In addition, flyers were distributed at the interview venues. We also distributed our call for participation among contacts with a background in IT security and asked them to forward it further. We attempted to recruit additional security experts for a more diverse sample on Freelancer.com, which has been recommended for recruiting software developers in prior work [260]. To ensure data integrity, we added screening questions on basic security knowledge to the beginning of the survey. However, few developers passed these questions, leading to only four additional participants. Participants on Freelancer.com were compensated with 15€. All other expert participants were eligible to participate in a raffle for three times 100€.

7.3.5 Participants

Of the 35 interview participants, 18 were interviewed at the USP community event (16 university, 1 industry, 1 government background), 11 at BlackHat Europe (9 industry, 1 university, 1 government background) and 6 at the 38C3 (3 university, 3 industry background).

	Ion et al. [236]	Busse et al. [74]	This paper
<i>N</i>	231	75	75
Gender			
Women	4 %	9 %	13 %
Men	~ %	79 %	84 %
Other	~ %	3 %	0 %
No answer	~ %	9 %	3 %
Age			
18 - 24	~ %	4 %	23 %
25 - 34	30 %	40 %	55 %
35 - 44	32 %	35 %	16 %
45 - 54	18 %	12 %	4 %
55 - 64	~ %	3 %	1 %
65 or older	~ %	0 %	1 %
No answer	~ %	7 %	0 %
Education			
Bachelor's or higher	73 %	75 %	85 %
Location			
US	47 %	27 %	5 %
Other	53 %	73 %	95 %
Work place			
Industry	69 %	51 %	37 %
University	15 %	21 %	43 %
Corporate research lab	7 %	9 %	5 %
Government	11 %	1 %	8 %
Self-employed	13 %	3 %	5 %
Other	~ %	~ %	1 %

Table 7.2 – Demographic comparison for expert demographics of Ion et al. [236], Busse et al. [74], and this paper. ~ indicates unknown percentages.

	Ion et al. [236]	Busse et al. [74]	This paper
<i>N</i>	294	288	990
Gender			
Women	40 %	48 %	50 %
Men	~ %	52 %	48 %
Other	~ %	0 %	1 %
No answer	~ %	0 %	0 %
Age			
18 - 24	19 %	9 %	11 %
25 - 34	50 %	45 %	17 %
35 - 44	19 %	25 %	17 %
45 - 54	~ %	14 %	16 %
55 - 64	~ %	6 %	26 %
65 or older	~ %	2 %	12 %
No answer	~ %	0 %	0 %
Education			
Bachelor's or higher	47 %	52 %	57 %

Table 7.3 – Demographic comparison non-expert demographics of Ion et al. [236], Busse et al. [74], and this paper. ~ indicates unknown percentages.

A comparison of demographic data of the survey participants across the three iterations of the study is in Table 7.2 for experts and Table 7.3 for non-experts, while more detailed participant information for the current iteration is in Table 7.4. Following the prior studies, we excluded participants who made more than one mistake on the three attention checks in the non-expert survey ($N=169$), and two attention checks in the expert survey ($N=2$).

The non-expert survey originally had 997 participants after filtering based on attention checks. Six further participants were filtered out because they submitted the exact same answers to open questions and one participant did not complete the full survey, resulting in a sample of $N=990$.

Our 75 expert survey participants came from the following recruiting channels: 4 through flyers distributed at our expert interview venues, 18 through LinkedIn, 1 through Reddit, 48 through personal contacts and 4 through Freelancer.com.

7.3.6 Limitations

Even though Busse et al's replication had shown that expert ratings for realism and effectiveness differ from responses to the compound question encompassing both [74], we reverted to Ion et al's original question phrasing [236]. Since we wanted to learn about experts' perceptions of new security practices, we had to balance effort and time needed for experts and the distinction between these two aspects of advice quality. The focus in this study was on new developments in security advice and as such we chose to accept the limitation of the compound question.

		NE	E
Gender	Women	50.1 %	13.3 %
	Men	48.4 %	84.0 %
	Non-binary	1.0 %	0.0 %
	Self-described	0.1 %	0.0 %
	No answer	0.4 %	2.7 %
Age	18 - 24	11.3 %	22.7 %
	25 - 34	17.3 %	54.7 %
	35 - 44	17.5 %	16.0 %
	45 - 54	16.1 %	04.0 %
	55 - 64	26.0 %	1.3 %
	65 or older	11.8 %	1.3 %
	No answer	0.1 %	0.0 %
Education	Profession. Doctorate	1.6 %	1.3 %
	Doctoral Degree	1.7 %	5.3 %
	Master	15.3 %	56.0 %
	Bachelor	38.4 %	22.7 %
	Associates Degree	9.7 %	10.7 %
	Some college, no degree	18.1 %	0.0 %
	Tech./Trade School	3.3 %	4.0 %
	Other	11.4 %	0.0 %
Occupation	No answer	0.5 %	0.0 %
	Employed full-time	47.1 %	
	Employed part-time	13.9 %	
	Self-employed	11.0 %	
	Care-provider	0.3 %	
	Homemaker	4.8 %	
	Retired	11.4 %	
	Student	2.3 %	
	Looking for work	6.7 %	
	Other	1.9 %	
	No answer	0.6 %	
Work place	Industry		37.3 %
	University		42.7 %
	Corp. research lab		5.3 %
	Government		8.0 %
	Self-employed		5.3 %
	Other		1.3 %
Security experience	1 - 4.5 years		28.0 %
	5 - 9.5 years		45.3 %
	10 - 14.5 years		13.3 %
	15+ years		13.3 %
Location	US	100 %	5.3 %
	Europe	0.0 %	92.0 %
	Africa	0.0 %	2.7 %

Table 7.4 – Full Demographic information for expert (E; $N = 75$) and non-expert (NE; $N = 994$) survey from 2024 sample

We attempted to replicate the non-expert survey recruitment on MTurk, following prior work. However, in our pilot sample of 10 participants, responses to open-ended questions appeared suspicious and contained patterns that suggested use of LLMs or bots, leading us to discontinue MTurk as a recruitment channel. Like the prior replication [74] our expert sample is smaller than in the original study [236], and recruited from a different population. It is a largely European expert sample with more university background. This might bear the risk of biasing the sample due to cultural differences, since Europe is widely believed to be generally more privacy conscious than the U.S. However, a recent study compared general population samples of both Germany and the U.S. and found that at least when it comes to government surveillance, there were only small differences in their privacy attitudes [172]. Using snowball sampling in the expert recruitment process means the responses may not be generalizable. We discuss development in recruiting practices as part of the methodological considerations in Section 7.5.

We based our coding of security practices and advice on the prior publications [74, 236], drawing from figures, tables, text, and additional coding material we received from the authors of the replication study. Nevertheless, qualitative coding is influenced by researchers' background and experiences [444, 163], and as such our application of the codes may have differed. In some instances, the prior iterations of the study used different names for codes. To increase readability, we unified these names when presenting the results.

7.3.7 Data Analysis

Given the wide range of advice evaluated and no clear hypotheses on expert vs. non-expert differences for specific advice, we chose descriptive statistics over null hypothesis testing. Nevertheless, to enable direct comparison to prior replications, we report on Fisher's exact tests comparing expert and non-expert behavior, and Wilcoxon signed rank test comparing non-experts' realism and effectiveness ratings, in the Tables 7.6 and 7.7. We report the effect sizes Cramér's V for the Fisher's Exact tests and rankbiserial r for the Wilcoxon signed rank tests, to enable comparisons even when sample sizes are different.

For qualitative analysis, we coded top-three responses (things-you-do for all, advice for experts) based on prior work [74, 236] and additional materials we received from Busse et al. After initial expert interviews, we expanded the codebook to include new advice, informing additional survey questions. Different sets of coders analyzed each of three types of data: Freetext answers from expert and non-expert surveys, and expert interviews. The main author took part in coding all types of data and ensured that the advice codebook was used consistently across data sets. For the survey answers, a subset of responses was coded independently by multiple coders, IRR was calculated and any differences in coding were discussed and resolved. This process was iteratively repeated until a sufficient IRR of at least $\kappa = 0.8$ had been achieved for the independent coding. The remaining data was split among coders. During this phase, coders marked responses for discussion if they were uncertain, and these were resolved together with all coders involved with the analysis of this type of data.

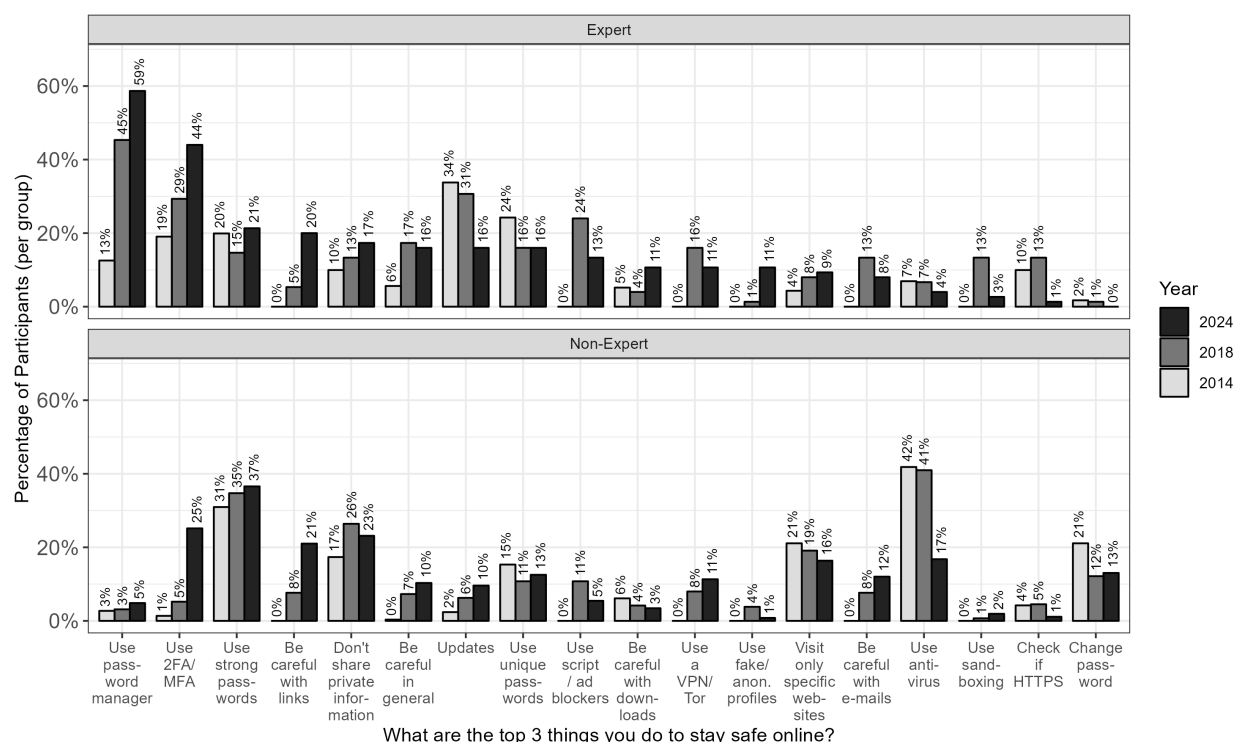


Figure 7.1 – Behaviors from the top 3, where at least one sample had over 10%

7.4 Content-level Results

We describe the content-level findings from our re-replication, beginning with the development of security behavior across the three iterations of the study, from 2014 to now. Next, we describe differences between experts and non-experts based on the surveys, focusing on newly included behavior and advice. Based on the expert interviews, we then describe an outlook into the future of security advice. We translated expert quotes to English where applicable. Quoted participants are represented by an identifier reflecting the interview venue: **CE** for the USP community event, **BH** for BlackHat and **C3** for the 38C3.

7.4.1 Development of Security Behavior over Time (RQ-Rep-1)

In our surveys, we asked both experts and non-experts to indicate the top 3 security enhancing behaviors that they carry out online. Figure 7.1 shows the development of experts' and non-experts' top 3 security behaviors across the three iterations of the survey. We only list behaviors which at least 10% of one sample (experts or non-experts in 2014, 2018 or 2024) named in their top 3, since there was a long tail of behaviors in the top 3 of barely over 5%.

We made some changes in naming the categories of the top 3 security practices, as compared to the works we build upon: Some respondents specified that they used multi-factor authentication, not only two-factor authentication, so we adjusted the code to include this: *Use 2FA/MFA*. Given the short responses, it is not clear if participants really use more than two factors or if they are

using words they have encountered during their practice. Similarly, many respondents did not clarify whether they meant updates on a (operating) system level, or of applications, so we summarized both as a single practice: *Updates*. We unified *Be careful/suspicious* to *Be careful in general* and use the description *Be careful with downloads* from Busse et al. [74] rather than *Use verified software* from the original study. Participants used other subjective criteria to decide which websites they would visit, including not only *known* and *trusted*, but also *professional* or *secure* websites, and avoided *risky* ones. To summarize these, we call the practice *Visit specific websites* and compare it to the practices *Visit known/trusted websites* from prior iterations. We shorten *Be careful with emails/attachments* to *Be careful with emails*.

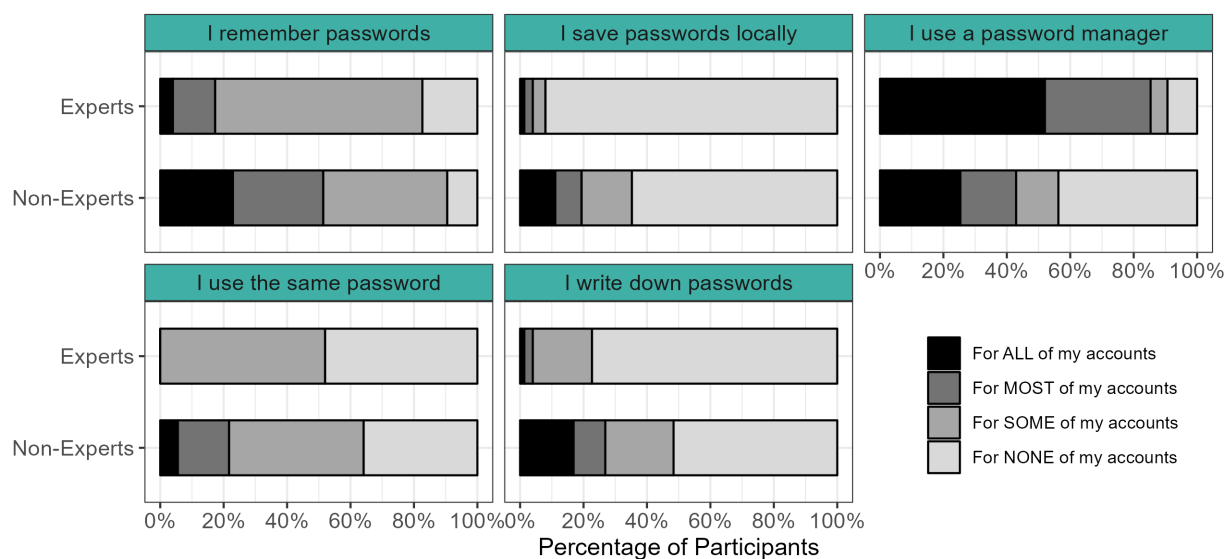
In Figure 7.1, some trends are immediately apparent, most notably the decrease in use of anti-virus software, and increase in use of 2FA. Also, non-experts increasingly report being careful with links. In the expert sample, the reported use of password managers and 2FA has increased, and the focus on using unique passwords and updating as security relevant behaviors has decreased. This is not to say that these measures are now deemed less important - they may also be seen as more self-evident or automated. Both the increased reported use of 2FA and the decrease in antivirus software was stronger in the non-expert sample compared to the expert sample. This leads to an overall more similar reported behavior between the two groups in the current sample, with a difference of 19 percent points in the use of 2FA as compared to the 24 percent points five years ago, and correspondingly 13 percent points difference compared to 34 in the use of antivirus.

We briefly compare this to current self-reported behavior (Figure 7.2b) and rating of these pieces of advice. 2FA was reported to be used for at least one account by both groups (E=99%, NE=93%). Non-experts rated the advice higher than experts. 44% of experts rated the advice as *very good*, while 74% of non-experts rated it as *very effective* and 66% considered themselves *very likely* to follow it. While using antivirus was among the top 3 self-reported security behaviors among non-experts in both 2014 and 2018, it now ranks fifth overall, with 17% of non-experts naming this in their top 3. Despite it being less popular than in previous years, non-experts still tend to be more favorable towards antivirus software than experts. While only 23% of experts rated using antivirus as *very good*, 50% of non-experts rated it *very effective* and 63% considered themselves *very likely* to follow the advice. Similarly, over twice as many non-experts reported using antivirus as experts (E=33%, NE=73%).

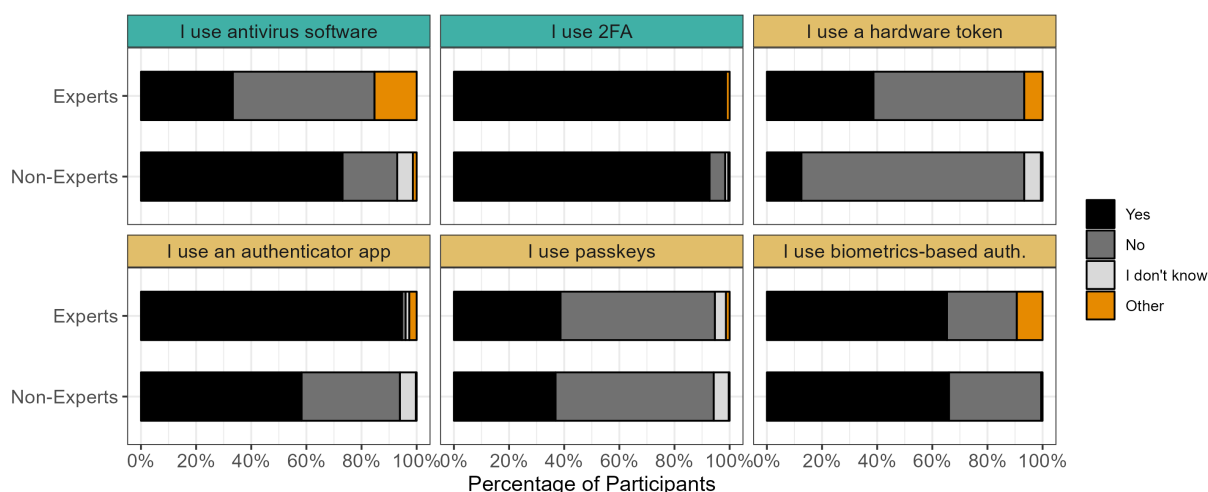
We identified seven new behaviors, which have not been discussed in prior work, but were reported by over 5% of either the expert or the non-expert sample: In our expert sample 11% considered *using fake/ anonymous profiles* online to stay anonymous among their top 3 security practices, while 1% of non-experts did this. Expert behaviors that were not included in Figure 7.1 were using *security settings*, e.g. in the browser or applications, *use encryption*, *use token-based authentication*, *self-host* and *limit internet use* (all >5%). For non-experts, one new security practice that had not been recorded in prior iterations was *use secure connections* (5%). This was often related to refraining from using public wifi, or at least being careful when using it.

7.4.2 Differences between Experts and Non-Experts (RQ-Rep-2)

In the surveys, security advice was rated on a scale of 1 to 5. Experts were asked how good (combining effectiveness and realism) they would find the advice for a non-tech-savvy user. Non-experts were asked to indicate how likely it was that they would follow that same advice, and how



(a) Comparison of password behavior



(b) Comparison of antivirus and authentication behavior

Figure 7.2 – Self-reported expert and non-expert security behavior, related to authentication practices.

We use background color to mark whether behavior was newly introduced: yellow for new, blue for old.

effective they felt it would be. We asked experts and non-experts to indicate whether or not they engaged in various security practices. We first discuss differences in advice ratings between experts and non-experts. A visualization of the advice ratings discussed in the text is in Figure 7.3. We report the percentage of participants who rated advice as *good* (i.e., experts who rated it as *very good* or *good*), *effective* (i.e., non-experts who rated it as *very effective* or *effective*), and *realistic* (i.e., non-experts who stated that they are *very likely* or *likely* to follow the advice). When comparing over the whole range of the scale, we report means. An overview over the rating differences between experts and non-experts and breakdown of responses into specific rating levels is in Table 7.5.

Advice questions	Expert – Goodness						Non-Expert – Effectiveness						Non-Expert – Realism					
	1	2	3	4	5	IDK	1	2	3	4	5	IDK	1	2	3	4	5	IDK
Use anti-virus	17	15	20	25	23	0	2	4	15	28	50	1	2	5	9	20	63	1
Install latest OS updates	0	4	12	27	57	0	1	4	13	30	51	1	1	3	9	23	63	0
Automatic updates	0	3	8	28	61	0	3	6	18	29	42	2	3	5	10	25	56	1
Update applications	0	8	23	35	35	0	1	5	16	33	45	1	1	3	9	27	58	1
Clear cookies	25	37	19	8	11	0	3	11	22	26	34	3	3	6	15	23	52	0
Check app permissions	9	25	31	23	12	0	1	5	17	28	48	2	1	4	12	26	56	1
Use unique pws	1	8	13	23	55	0	2	3	8	20	67	0	4	5	12	19	60	0
Use hard to guess pws	4	5	16	20	55	0	1	2	4	16	78	0	2	2	6	19	72	0
Don't write down pws on paper	12	20	24	13	28	3	14	12	16	15	41	3	19	10	12	13	44	2
Save pws locally	45	32	11	9	1	1	38	19	17	11	11	3	35	14	16	13	20	3
Use pw manager	1	1	13	28	56	0	7	8	20	26	35	3	11	9	17	19	41	3
Write down pws on paper	28	17	31	13	5	5	36	15	17	13	16	4	37	11	15	13	21	4
Check if HTTPS	5	9	16	32	37	0	2	4	11	27	55	2	2	3	11	20	64	1
Be careful in general	1	11	27	23	39	0	0	3	10	21	66	0	1	2	9	19	68	0
Be careful with links in e-mails	0	7	8	27	59	0	1	2	4	13	81	0	1	2	5	13	81	0
Visit only known websites	16	27	24	26	7	0	2	5	20	28	45	1	6	8	19	25	41	0
Use 2FA	3	8	16	29	44	0	1	2	5	17	74	1	2	3	9	21	66	0
Don't share private info	3	17	24	23	32	1	1	1	5	16	77	0	1	1	5	15	78	0
Use hardware token	11	25	27	17	17	3	5	4	19	23	33	16	11	11	23	21	23	10
Use authenticator app	0	8	20	40	29	3	2	3	11	25	50	8	5	6	15	22	47	4
Use passkeys	3	13	17	27	24	16	5	5	16	27	37	11	9	8	19	22	36	6
Use biometrics	4	8	20	29	35	4	4	3	11	25	55	4	8	6	11	18	55	3
Don't click on links from unknown	0	16	23	28	33	0	1	2	5	16	76	0	1	2	5	13	79	0
Don't enter pws on links from e-mail	0	7	16	25	52	0	1	2	3	14	79	1	1	1	4	14	80	0
Check URL	1	8	15	37	39	0	1	2	5	18	74	0	1	2	8	15	73	0
Don't open e-mail attachments from unknown	0	4	16	27	53	0	0	1	5	13	81	0	1	2	5	13	79	0
Use AI security advice	20	29	23	20	5	3	21	17	28	15	12	7	25	13	24	15	16	6

Table 7.5 – Expert and non-expert ratings of advice. Differences of expert “goodness” ratings from non-expert realism and effectiveness ratings are marked in green, if the difference is less than 5 percent points and in orange, if the difference is more than 10 percent points per individual response category. We abbreviate “password” as **pw**.

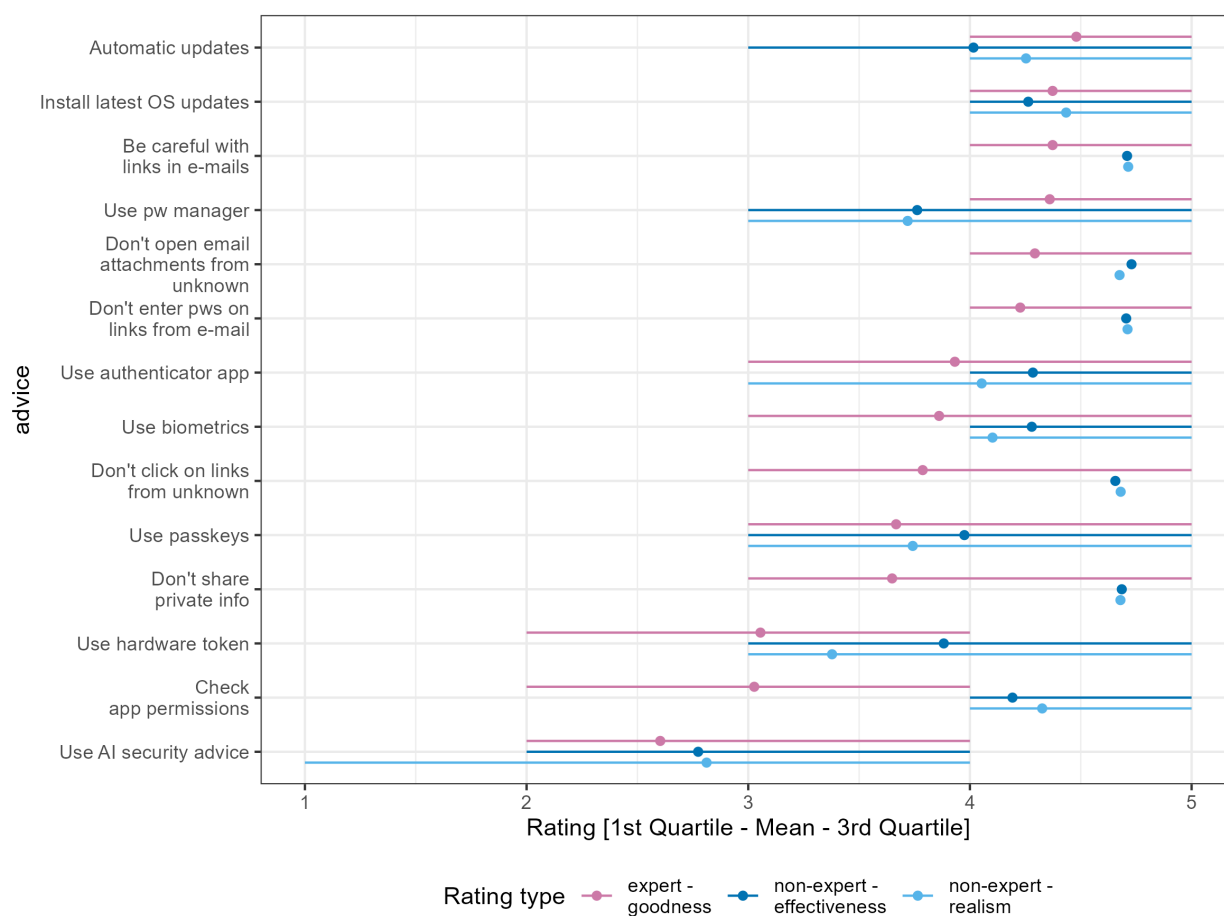


Figure 7.3 – Means (dots) and Interquartile Range (lines) of experts' and non-experts' advice ratings for advice discussed in the text, sorted by expert ratings.

We also compare expert and non-expert behavior. Responses to the behavioral questions are in Figures 7.2, 7.4 and 7.5. Across all categories of practices, the response option *Other* was quite common for experts. They had more nuanced explanations on their behavior, which we will reference where appropriate in the following.

7.4.2.1 Rating of New Advice

We added new advice to be rated based both on the expert interviews at the USP community event and new practices identified in the first replication [74], which had not previously been included in the advice to be rated. Expert and non-expert participants did not appear to be enthusiastic regarding using AI (such as ChatGPT) for security advice (Figure 7.3). Non-experts were more positive than experts about limiting sharing of unnecessary personal information (Figure 7.3), and also self-reporting that they did this (Figure 7.5). Furthermore, 65% of experts reported not sharing information, while 89% of non-experts did so, revealing diverging attitudes and practices between the two groups. However, this may be due to differences in what experts and non-experts determine

to be “unnecessary” personal information or knowledge about what information is gathered and what can be inferred from it. This could potentially lead to higher rates of experts indicating they share such information. Two experts commented, one that they were not sure “what qualifies as unnecessary” and one that they restricted which people could access private information that they share. Meanwhile, though both groups check app permissions (often or sometimes) at similar rates (E=71%, NE=74%) (Figure 7.5), experts were much less positive about checking app permissions as security advice than NE (Figure 7.3).

Authentication Methods We also asked a set of new questions regarding authentication methods, and found that non-experts tended to be less familiar with advice regarding authentication methods, in particular hardware tokens and passkeys. Figure 7.2b compares non-expert and expert behavior related to 2FA and use of different second factors and Figure 7.3 contains corresponding advice ratings.

Both experts (83%) and non-experts (87%) viewed securing accounts with 2FA positively, with high reported adoption (E= 99%, NE= 93%). Regarding the choice of the second factor, experts showed a clear preference for authentication apps, with 69% rating them as *good* advice, compared to only 34% for hardware tokens. This preference was reflected in their reported usage: 95% of experts reported using an authentication app for at least one account, whereas only 39% reported using hardware tokens. A similar pattern was observed among non-experts, though their overall reported adoption rate was lower: 58% reported using an authentication app, while only 13% used a hardware token. Interestingly, we found that non-experts struggled to understand hardware tokens, with 16% responding *I don't know* when asked about their effectiveness. However, non-experts appeared more familiar with passkeys than with hardware tokens. Rates of reported use of passkeys between experts and non-experts were similar (E and NE = 39%), though interestingly, as were rates of answering *I don't know* when asked if they used passkeys (E=4%, NE=5%). Despite the higher reported use rates of passkeys compared to hardware tokens for non-experts, this group did not tend to rate it as *effective* more frequently (64%) though there was an increase in the proportion of non-experts saying they would *likely* follow this advice (58%). Interestingly, experts appeared to be more uncertain about passkeys than non-experts, with more experts (24%) responding with *I don't know* regarding the goodness of advice on using passkeys than non-experts (11% *I don't know* for effectiveness, 6% *I don't know* for realism).

64% of experts thought using biometrics was *good* advice, while 80% of non-experts considered it *effective* and 73% said they were *likely* to follow this advice. Rates of use were similar between the two groups (E=65%, NE=66%). Experts responding *other* used biometrics to unlock their password manager or passkeys and were unsure whether this counted in the sense of our question. Two also explicitly mentioned using biometrics on their phones.

These results suggest that non-experts are less familiar with some of these authentication methods. It may also be that experts may feel that while the advice is effective (especially as a large percentage of these experts reported using these methods), it may not be realistic, which we were unable to distinguish given the compound format for our questions for experts. However, we did not see large discrepancies between the realism and effectiveness ratings from non-experts for this category of questions.

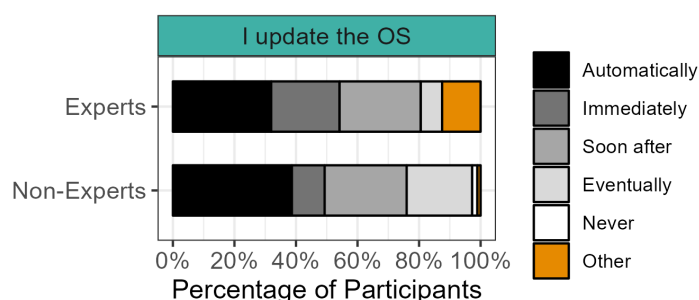


Figure 7.4 – Comparison of expert and non-expert update behavior. This question was included in prior iterations of the study.

7.4.2.2 Remaining Usability Discrepancy

Password managers were consistently favored by experts, in terms of use and advice, but much less so by non-experts (Figure 7.2a and Figure 7.3). Experts were more positive towards advice on using password managers (56% *very good*) than non-experts (35% *very effective*, 41% *very likely*). It also is the most frequent of the top 3 actions that experts take to manage their own security (59%), though listed rarely by non-experts (5%) (Figure 7.1). The behavior of experts matches their evaluations of the advice, with 52% of experts reporting they use password managers for *all* of their accounts, while 25% of non-experts use it for *all* accounts. On the other hand, 44% of non-experts did not use password managers, as opposed to only 9% of experts. It was the only piece of advice rated as good ($M=4.33$) by experts but not realistic by non-experts ($M=3.72$). In fact, its realism score was the sixth lowest of all the pieces of advice according to non-experts, but the fourth best piece of advice from the mean rankings by experts. At the same time, expert interviewees acknowledged usability problems with password managers, e.g. P-BH-4 explained “[...] you set it up correctly. It actually works really really well. But [...] sometimes I’m like why the hell is this not working on my mobile phone?” Thus, we see a mismatch between what experts believe non-experts should do in terms of security, and what non-experts believe.

7.4.2.3 Best Rated Advice

We find discrepancies between what experts and non-experts think is good advice, and how much the advice is self-reported to be followed in practice by both experts and non-experts. A visualization of the rating differences is in Figure 7.3.

Experts’ Best Rated Advice Experts ranked advice relating to software and OS updates highly, with two pieces of advice related to updates in the top 3 in terms of mean ratings of advice goodness. Automatic software updates was ranked the “best” piece of advice by experts in terms of mean goodness rating (4.48). Experts were somewhat more favorable towards this advice than non-experts. Despite being more positive than non-experts towards automatic software updates, fewer experts said that OS updates were installed automatically than non-experts ($E=32\%$, $NE=39\%$) (Figure 7.4). Nonetheless, experts appear to be a little more timely at installing OS updates in general than non-experts. This aligns with the advice to install the latest OS updates, which was

tied for second rank in terms of mean goodness (4.37) according to experts. Regarding self-reported behavior, 81% of experts reported that they either updated their OS *automatically*, *immediately*, or *soon after* (with a further 13% responding with *other*), while a slightly lower proportion of non-experts (76%) install the latest OS updates. This may be because while experts value keeping their OS updated, they prefer to have more control over their own OS, and thus think of automatic updates as good for non-expert users rather than for themselves. Among the 13% of experts providing freetext answers to the question on their OS update behavior, some experts discussed reasons for their update frequency, e.g. delaying updates to monitor stability or updating on a rolling release base.

In some cases, experts and non-experts had similar evaluations about what is good security advice, as both groups ranked the advice to be suspicious of links in emails and messages highly. It had the second highest (though tied) mean rating for goodness (4.37) according to experts as well as the highest for realism (4.71) and second highest for effectiveness (4.71) for non-experts.

Non-experts' Best Rated Advice The advice to not enter passwords on sites linked in emails is rated the second highest in realism (M=4.71) and third highest in effectiveness (M=4.71) for non-experts. In fact, 77% of non-experts say they follow this advice and never enter their passwords on sites linked from emails (Figure 7.5). Experts were less positive about this advice. While 77% of experts rated the advice as *good*, a smaller percent of experts (60%) say they follow this advice. 20% of experts respond with *other*, suggesting more nuance to expert practices than a clear-cut yes or no. Not opening email attachments from unknown senders ranked the highest in mean effectiveness for non-experts (4.73). It is followed by 56% of non-experts. Despite 80% of experts rating it as good advice, only 11% of experts (with a further 6% of experts answering *other*) actually follow it (Figure 7.5). Not clicking on links sent by unknown senders was rated by non-experts as the third most realistic advice (M=4.68). Experts were not as positive (61% = *good*) towards this advice compared to non-experts (rated *effective* and *realistic* by 92% each of non-experts), nor did they report to follow it in practice (65%) as much as non-experts (82%) (Figure 7.5).

From the trends in the most highly rated advice from experts and non-experts, we see diverging opinions of what is best for one's security. In general, both groups believe that being aware of potential risks (e.g., be suspicious of links in emails and messages) is effective and realistic. However, the groups differ in what specific actions to take (or not take). The results suggest that experts think *positive* security advice (e.g., installing updates) to take an action that increases security is good, while non-experts think *negative* security advice (e.g., not opening emails, not entering passwords on unknown sites) to not take an action that potentially decreases security are most effective or realistic.

7.4.3 Experts' View on the Current State of End User Security Advice Practices (RQ-Rep-3)

In the interviews that we conducted, we asked IT security experts two main questions: firstly, we asked for their top 3 pieces of security advice that they would give to non-experts. Then we showed them either the results of the two survey studies preceding ours, or in later interviews, we additionally showed them the results of our replication of the survey.

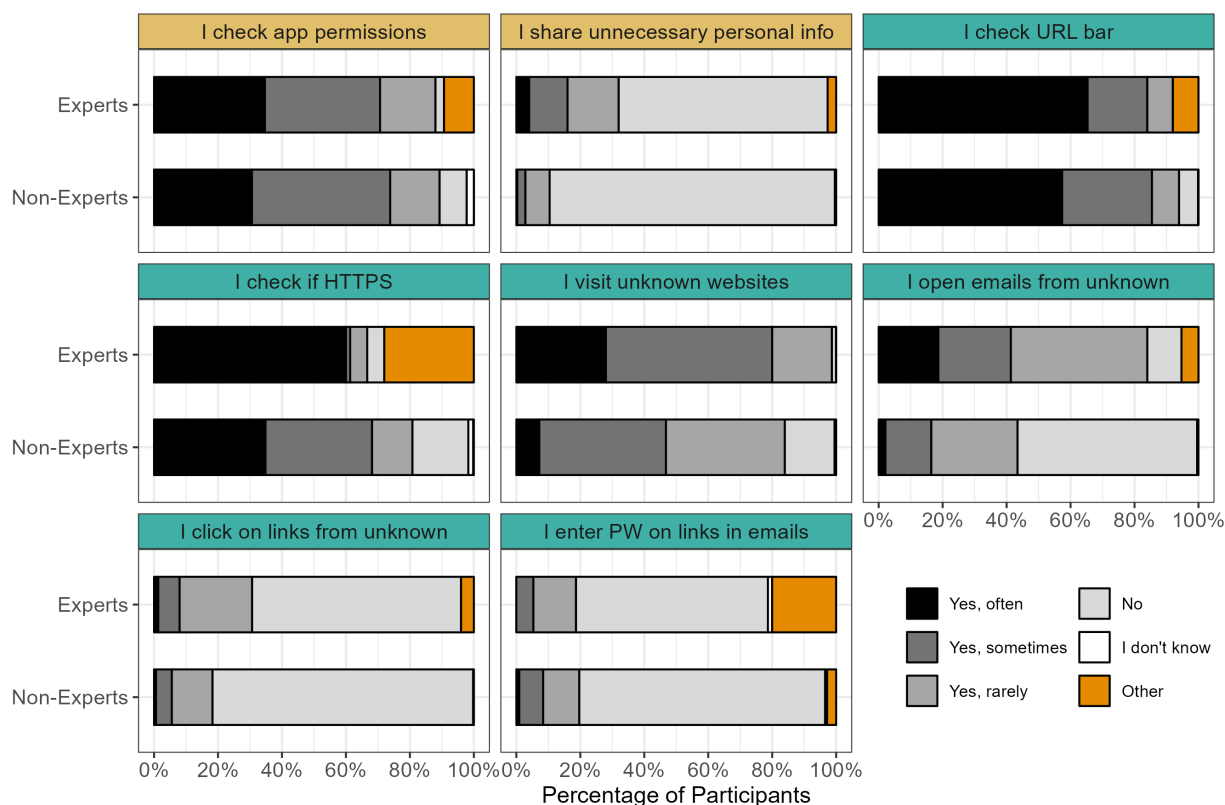


Figure 7.5 – Self-reported expert and non-expert security behavior, related to mindfulness. We use background color to mark whether behavior was newly introduced: yellow for new, turquoise for old.

7.4.3.1 Security By Default

Many interviewed experts felt that IT security should not be the responsibility of end users. They discussed examples of how responsibility for security can be kept away from users, some of which are already implemented and others they would like to see in the future. This included the automation of software updates, use of anti-virus software, and link security.

A common view was that keeping software up to date is crucial for security. Many experts supported automatic updates, as they reduce the users' responsibility. However, both expert discussion and usable security research emphasize the need for clear communication and some user control to avoid frustration [556]. As one interviewee put it, "It's simply updated whether you like it or not. [...] so I wouldn't have to put so much effort into it, because there are simply updates" (P-CE-11). A similar view was expressed in relation to anti-virus protection. Some experts pointed out that solutions such as Windows Defender have improved significantly, making third-party anti-virus software unnecessary for most users. As one expert put it, "Use antivirus, oh yes, and thankfully this is a default in Windows" (P-CE-18). Link security was another topic that came up frequently in the interviews. Several experts noted that links are meant to be interacted with, so it is important that browsers and email clients provide better protection. As P-CE-18 explained, "Because links are made to be clicked overall, so it's about the infrastructure to help with that, [...] the browser

or the email client". This is a topic where experts still saw the need for improvement. Overall, many of the experts interviewed shared the view that security should be built into systems by default, reducing the need for user awareness and manual intervention. However, there was also an emphasis on maintaining transparency and giving users a degree of control to avoid frustration.

7.4.3.2 Risks and Potential of AI

When speaking of AI, our expert interviewees largely discussed it in terms of risks. These risks referred to privacy, with user data being used to train models, when users interact with AI and AI-powered services. One interviewee also worried about AI-driven possibilities for deepfakes. These concerns lead the experts to advise strongly that users avoid sharing personal information, specifically with AI, but also more generally online, given that training data can also stem from social media and the broader internet [29]. These expert views are in line with non-experts' views from a recent multi-country survey study [267]. However, P-BH-4 also sees chances in the use of AI: "Threat actors are using AI, but defenders need to actually get better at using AI."

7.4.3.3 Focused and Specific Advice

Some interviewees pointed out gaps in the advice presented from prior and current results based on non-expert data, for example in the way that advice seemed overly geared towards computers: "This is a time, where in many countries almost nobody has a computer anymore but they are working with tablets, they are working with smartphones. And adding to that all the verbal devices [...]" (P-CE-6). The interviewees saw the need for device and situation-specific advice, e.g., advice focused on mobile or internet of things devices. This rise in different types of devices may provide chances to aid secure behavior online, for example, using encrypted messaging apps was perceived as a low-barrier method to communicate securely on mobile devices. But interviewees also pointed out risks, e.g., an increased risk of shoulder surfing when using mobile devices in public spaces, or lack of transparency for data transmission processes on internet of things devices.

Interviewees also mentioned the need for advice being focused on specific user groups. The interviewees saw some technological solutions improving security, like password managers or 2FA, as complex for non-experts, but more accessible for experts: "Use of firewall is irrelevant for most people. They have a router in the back, then they don't need anything else" (P-C3-4).

7.4.4 Discussion

7.4.4.1 Changing Patterns of Security Practices

Over the ten years since the original "No one can hack my mind" study was published, we observe development in non-expert and expert security practices. In this series of studies, new advice appeared in both replications, e.g., using script and ad blockers or VPNs in 2018 and using security settings or fake profiles in 2024. Non-experts have adopted and value security practices that were previously mainly valued by experts. In 2024, one of the most extreme changes was in using 2FA. While this was among the pieces of advice that were rated as effective by experts in 2018, now it is

the second most frequently named behavior that is seen as important by non-experts, with 97% of non-experts reporting using it for at least one account. Adoption of security technology, like 2FA, can be driven by usability improvements [578] but mandating the technology use may also be an influencing factor [1, 112]. Mandates are however a double-edged sword and may lead to security fatigue [379, 1, 112], and difficulties in deviating from in-grained practices, should advice change over the course of time, as was the case for recommendations on regularly changing passwords [177].

On the other hand, some security practices decreased in importance, such as the use of anti-virus software in the perception of non-experts, and updates in the perception of experts. In our expert interviews, automation was seen as a reason that these practices may have become less visible. Since they are often activated by default, their users may not be aware of them and thus their importance is sidelined. This aligns with a long-standing debate in the USP community about the trade-off between automation and user control [132, 99, 536, 333]. Automating security responsibility away from users also carries the risk of lack of transparency and feelings of loss of control and trust, leading to users circumventing automated security [556, 322, 141, 536]. Also, automation requires predefined rules for decisions: Their accuracy can be limited within flexible social situations and changing contexts, resulting in failures of the automation and, thus, involvement of users [132]. In turn, security design should balance automation with clear communication, minimizing responsibility for non-technical users by implementing secure defaults [379] while retaining control for those willing to exert it and keeping users in the loop [99, 132]. We propose to continue to monitor the state of advice perception among non-experts and experts to identify challenges and further fields of action.

7.4.4.2 Divergences Between Experts and Non-Experts

While for some advice, expert and non-expert behavior became more similar, there were still many differences between the two groups. One of the most striking was the attitude towards using password managers, which was the single most frequently reported behavior within experts' top 3, and received the fourth best rating. However, this advice was rated as unrealistic by non-experts with the sixth lowest mean rating. This can be considered a remaining challenge for the USP community. On the other hand, some advice like not clicking on links from unknown sources was perceived as realistic by non-experts, but not good advice by experts. Additionally, even experts are not consistent in their perception of advice. Two of the top 3 pieces of advice about updates received high ratings from experts, but were not perceived as realistic by non-experts and experts did not really follow the advice themselves. A prior study from 2018 similarly found that self-declared experts reported fewer security behaviors [76]. However, this likely stems from experts' more nuanced view on advice, which they explained in free text answers. They have the confidence or knowledge to deviate from advice when they see it as not necessary, but nevertheless experts themselves also diverge in their behavior and recommendations, as shown in related work [417].

7.4.4.3 Advice Should Be Context-Specific

While the primary question in this study asks about “staying safe online”, the meaning of online has changed in the past ten years. The internet is now more than ever used through and by multiple devices and in different contexts, from internet cafés in Kenya [338] to smart home environments

encompassing interactions between multiple devices and multiple people [470, 357], and connected cars [59]. Similar to these fragmented online use patterns, the status quo of security advice is also confusing for users, with a wide variety of advice available and experts not able to agree on prioritization of advice, leaving non-experts confused and without clear actionable advice [417]. Perhaps the way forward is not to try and find one set of the universal best pieces of advice, but rather identify the advice most salient for individual users or user groups. USP research has already embarked on this endeavor by investigating advice for older adults [335], activists [63], LGBTQ+ folks [171] or software developers [8], and how, where and why it is circulated and implemented or dismissed. For other user groups, such as users with visual impairments, their difficulties in achieving security and privacy online have been documented [346, 209], but there has been no specific investigation of advice directed at these users.

Another way to focus advice is by providing advice for specific devices. Two examples mentioned in our expert interviews were mobile devices and internet of things devices. Research on mobile security focuses on specific security or privacy related functionality, such as permission management [493, 12, 400] or authentication [204, 84], but there is no evaluation of mobile-focused advice. Much of the available security advice for internet-of-things devices was not actionable, and not understandable for non-expert users [41, 524].

7.5 Meta-scientific Results

We discuss differences and changes between the three iterations of the study across three different areas of methodology: study design, participant recruitment and statistical analysis.

7.5.1 Differences in Study Design

The main differences in study design are shown in Table 7.1. While the non-expert survey remained largely the same over the three iterations, with the exception of the addition of new questions in the most current iteration, the expert survey underwent more changes. Busse et al. investigated the effect of the use of the compound question on experts' advice ratings and found differences between realism and effectiveness ratings for experts for some pieces of advice [74]. Due to questions we added to the 2024 iteration of the study, which made it even longer than the original study, we expected similar constraints on experts' willingness to participate as the authors of the original study [236]. So, despite knowing about the limitations of the compound question, we decided to move back to using it instead of two separate questions, to cut back on survey length at least a little bit. Even so, the median duration for the experts in our final sample was 19.4 minutes. While actual durations were not reported in prior work, Busse et al. advertised their 2018 study, which was a close replication to the original study, as taking 10 minutes [74]. For the original original study in 2014, there were already two comments on the survey being too long on the Google Security blog used for expert recruiting [419]. In our 2024 replication, there was a comment on the duration in the freetext fields of the expert survey as well: "Your 10-minute time estimate is a bit optimistic :) Or maybe I'm just overthinking all my answers".

We propose some options to deal with this conundrum, especially in the event of further replications of this study. From Busse et al.'s work, we know the pieces of advice where realism and effectiveness

ratings differed most, e.g. “Use unique passwords”, “Use strong passwords”, and “Use two-factor authentication”, all with mean differences of over 1.5 between experts’ effectiveness and realism ratings. We can interpret experts’ goodness ratings in the light of these findings, assuming that the goodness ratings for these pieces of advice are likely somewhere in the middle of what the effectiveness and realism ratings would be. Of course, as time passes, this divergence of effectiveness and realism may not apply to all the advice for which Busse et al. had identified these differences anymore. In addition, for the questions added to the survey in the 2024 iteration, we do not have this information at all.

Another option would be to try to reduce the length of future surveys, likely taking out some questions, to be able to move back to separate questions for rating advice for the expert surveys. One possibility would be to remove behavioral questions where experts and non-experts hardly differ. An example of this is the question on use of 2FA. However, this question could alternatively be rephrased in future iterations to provide more informational value, i.e. moving it from a binary choice of using 2FA for at least one account to a phrasing similar to the questions on password behavior, asking for the number of accounts the technology is used for (from “ALL of my accounts” to “NONE of my accounts”).

Future research could reduce the number of questions by only including behaviors and pieces of advice that we as a community perceive to be helpful. However, given that there is variety in expert opinion and difficulty in deciding which advice to prioritize [417], reaching agreement on such a set of advice to include is a challenge in and of itself. Additionally, monitoring the frequency of behaviors which are not seen as helpful or even dangerous is also useful to inform communication about security advice.

Finally, another option is to recruit additional expert participants and to administer two varieties of the survey, each with questions on different behavior and advice to make the length of the individual surveys more manageable. However, as we discuss in Section 7.5.2, recruiting, especially expert participants, is also difficult.

7.5.2 Development of Participant Recruitment

The differences in recruitment for non-experts and experts are summarized in Table 7.1. In the age of widely available AI tools, like ChatGPT, recruiting presents a challenge both for non-expert and expert samples, as these tools can be used to generate responses quickly, and make screening based on prior knowledge or competence harder [538, 86].

The two prior iterations both used MTurk to recruit participants for the non-expert surveys, with Busse et al. adjusting payment after participation to remain fair and account for inflation. In the original study, 6 participants were screened out during quality control, and in the 2018 replication, 12 participants were screened out. In the 2024 replication on Prolific, 169 participants were screened out based on too many failed attention checks, and 6 submissions were removed because they had participated multiple times, with one submission being removed as incomplete.

When we started the 2024 replication, we had originally intended to collect data both from MTurk and Prolific. MTurk had been used for recruiting in the USP community for a long time [266, 415] but recent evaluations discussed bot prevalence [501]. We wanted to get a more direct comparison to the prior studies, using MTurk, in terms of responses but also data quality. In addition we wanted

to compare the findings from this sample to a census-representative US sample along gender, age and ethnicity collected from Prolific. Due to issues with data quality in MTurk, we decided to solely recruit on Prolific, but we discuss findings on bot detection in the Prolific and the MTurk sample in the following.

Even though incentives to cheat by using bots or an LLM to answer the non-expert survey are likely not as high as for the higher paid expert survey, they are still there. In our survey, we experimented with strategies to detect bots and use of AI, in addition to filtering out non-attentive participants using the attention check questions also present in prior iterations. While these questions are described in Section 7.3.3.2, we now describe what we learned from the use of these questions in our Prolific sample. There were no answers at all to the question hidden by JavaScript, so it did not lead to filtering out any bots in the Prolific sample. For the question which included hidden text with a white font on a white background, 13 participants responded to the hidden text. We checked their other answers for signs of bot use and retained all of them, with one being removed for failing attention checks later. The white-on-white text can also be detected by humans when selecting text with a mouse or finger.

We had designed the additional optional knowledge questions in a way that they would be hard to answer by a single non-expert, spanning different topics in IT security, e.g. asking for the definition of spin-locks in the Linux operating system or an explanation of salting in the context of password storage. Even though we had explicitly stated that these questions were optional, could be left blank and participants should only answer them if they knew the answer, and should not use any other resources to help them answer the questions, many participants nevertheless entered a response. Often this was “I don’t know” or similar but some participants also tried to give an answer. Some of these answers were very vague and brief, others were verbose and in our opinion likely the result of using additional resources. Only 120 out of 990 valid participants left the field completely empty for all three of the knowledge questions. Anecdotally, in answers to our final question asking for any remaining thoughts and feedback on the survey, some participants disclosed using an LLM or searching to answer these questions because they did not know the answer, e.g. stating “I did use Chat GPT on those last 3 questions. I am not an IT pro”. One participant withdrew from the study after encountering these questions, likely because they were scared that their submission would be rejected. On Prolific, the amount and frequency of rejections are used to judge the reliability of participants, and if their responses are rejected too frequently, they may be removed from the platform[403]. By returning a submission, participants can avoid being rejected. We manually approved this participant after checking their responses. Other participants mentioned they were relieved to see our debrief after these questions, where we assured them they would not be judged negatively if they had not been able to answer these questions. From these observations, we concluded that using these kind of optional knowledge questions was not a very effective way of screening out non-human responses, since crowd workers seem to expect rejection if they do not or cannot answer all questions, even if the questions are explicitly marked as optional. This makes a manual review necessary, not only of the answers to the optional knowledge questions but also of other freetext answers of that same participant. Due to the reasons described above, we did not use the answers to the optional knowledge questions to screen out participants in the Prolific survey.

In our pilot on MTurk with 10 participants, we noticed several signs pointing towards bot participation in our survey. Many participants’ answers to the question about their top three behaviors, were highly similar to each other. This was especially suspicious due to the exact similarity and

use of correct terminology in their responses compared to our Prolific sample. For example, 6 out of 10 MTurk participants had a close variation of “Enable Two-Factor Authentication (2FA)” in their responses. All of these used this exact phrasing, with slight variation in capitalization or the inclusion of the abbreviation in parentheses at the end. In comparison, we coded a wider range of phrasings as “Use 2FA/MFA” for the Prolific answers, e.g. “Two step authentication”, “Multi-factor”, “Two way authentication” or “Two-Factor authorisation”, most without additional verbs. One MTurk participant’s responses, while not falling into these categories of similarity, were instead non-sensical, e.g. “safety” or “good information” as a response to the question about their top three security behaviors. Another point of notice was the length of responses to freetext answers. While the median numbers of characters in answers to the question on where participants learned about their top three behaviors was 56 (range=3 to 802) for the final Prolific sample, it was 85 (range=22 to 1145) for the MTurk pilot. We also saw evidence of participants truncating long LLM responses after a certain amount of characters, likely to avoid detection. Again there was similarity in content for the MTurk responses, with four responses centering on learning from a “combination of trusted sources” and two from “books, websites, articles”. Three further responses did not answer the question but repeated the top three security practices, submitted what sounded like a promotional text about Bitdefender or submitted a general statement about the importance of IT security. We also observed answers to different questions not fitting together, i.e. one participant answered that they do not use a password manager for any of their accounts, but nevertheless answered that they use Bitdefender when asked which one they use. Finally, the question hidden by JavaScript was answered by all of the MTurk participants, with four of them answering in the requested style of “as if you were my grandmother, tucking me into bed”. None of the responses in the MTurk pilot would have been usable for the main study, and so we decided against collecting further data on MTurk.

In terms of bot detection, freetext questions were helpful in both the MTurk and the Prolific sample but required manual evaluation to detect some of the likely bots. They can also be used to detect lower quality answers in general and if a survey requires manual coding of freetext responses anyway, this step will not add additional analysis effort. The question hidden by JavaScript was only answered by participants in the MTurk sample, no participants in the Prolific sample answered it. Since here, answering this question at all is an indicator towards a non-human participant, making the analysis straight-forward, and it does not place an additional burden on human participants, it could be included in future surveys, at least as a fall-back mechanism. Most participants in the MTurk sample did not respond to the question part in white-on-white font but we received a few responses from participants in the Prolific sample whose answers we judged to be valid otherwise. As such this question was not very good at differentiating between humans and bots, and we do not recommend further use for now. Finally, for researchers intending to test bot detection questions for their survey, we recommend collecting data on MTurk, as there seems to be a high prevalence of them there.

For expert recruitment, the original study had used a post on the Google Developer Blog [419] and personal contacts to recruit a final sample of 231 security experts. For their study, a person counted as an expert if they had at least five years of experience working or studying computer security. Since the authors of the 2018 replication were not able to use the Google Developer Blog as a resource, they resorted to using social media, including X, then Twitter, various security related subreddits, and mailing lists, as well as asking personal contacts to share and pass on the study advertisement. They were able to recruit 75 security experts, although to achieve this number, they

adjusted the inclusion criteria to include people if they had at least one year of experience working or studying computer security, rather than five. Recruiting security experts for empirical research remains a persistent methodological challenge and we managed to recruit 75 security experts, applying the inclusion criteria of the 2018 replication, using a variety of recruitment channels.

In 2024 the social media landscape had changed. Especially in academia, in addition to what had been Twitter, several other platforms were in use for the purpose of microblogging, e.g Mastodon³ or Bluesky⁴. The main author of the 2024 iteration was not active on any of these platforms. In addition, many subreddits had implemented stricter policies in regards to publishing calls for study participation on them, so the majority of these were not available to use in our recruiting strategy in 2024. Again, this may have been different if the primary researcher had had a prior reputation of active participation in these communities but this was not the case. Only one valid expert participant was recruited through Reddit, and an additional 18 valid participants through LinkedIn. Since our recruitment for interviews at security related venues worked well, we also distributed flyers including calls for survey participation but this only led to the recruitment of four additional valid participants over two events used for recruiting. We hypothesize that this is due to less personal interaction with participants when distributing flyers and assume that survey recruitment at these venues would have been more successful if we had handed the flyers directly to individual attendees, rather than mostly leaving them communal areas.

For the 2024 iteration, the most effective recruiting strategy was through sending messages to known security expert personal contacts and asking them to forward the call for participation. This ties in with prior work on recruiting software developers, which suggests that they preferred personal contact for being recruited [459]. However, this approach is likely to introduce sample selection biases and recruiting through social media, as well as personal contacts are not reproducible recruitment strategies, as they depend on the researchers' social networks. For the recruitment strategies discussed above, experts were able to enter a raffle for three times 100€ at the end of their survey participation. This compensation strategy is commonly used with recruitment on social media but also offline in USP [258, 330, 54].

However, we also tried to recruit additional security experts through Freelancer.com⁵, which was recommended as a recruiting channel for recruiting software developers for security studies [260], which we also consider a type of expert participant. Since participants through Freelancer.com are paid, and at a higher rate than non-experts, an additional challenge becomes verifying participants' expertise, as non-experts are financially incentivised to pose as experts. This problem is compounded by the growing sophistication of LLMs, which lower the barrier to deception. This may explain the low success rate of our recruitment efforts on Freelancer.com, where we were only able to recruit four additional valid participants. Though ChatGPT-resistant screening instruments exist for participants with programming expertise [461], there are varied forms of security expertise, making it hard to pinpoint and verify. As screening becomes a cat-and-mouse game, future studies may need to rely more heavily on settings where participants' identity and expertise can be verified through direct interaction.

We believe the USP community needs to continually re-evaluate the recruitment channels used for expert (of any kind) as well as non-expert recruiting. This is the case even when prior work

³<https://mastodon.social>

⁴<https://bsky.app/>

⁵<https://www.freelancer.com/>

Behavior	$N_{Experts}$	$N_{Non-Experts}$	Cramer's V	95% CI	raw p	adj. p
Do OS update	63	939	0.12	[0.01,1]	0.003	0.023
Use anti-virus	61	882	0.22	[0.17,1]	<0.001	<0.001
Remember passwords	75	990	0.17	[0.108,1]	<0.001	<0.001
Write down passwords on paper	75	990	0.14	[0.073,1]	<0.001	<0.001
Save passwords locally	75	990	0.14	[0.073,1]	<0.001	<0.001
Use password manager	75	990	0.21	[0.154,1]	<0.001	<0.001
Re-use passwords	75	990	0.13	[0.059,1]	<0.001	<0.001
Use 2FA	74	974	0.06	[0,1]	0.028	0.181
Use hardware token	70	924	0.20	[0.143,1]	<0.001	<0.001
Use authenticator app	72	930	0.20	[0.142,1]	<0.001	<0.001
Use passkeys	71	933	0.00	[0,1]	0.802	1
Use biometric auth	68	985	0.00	[0,1]	0.425	1
Check app permissions	68	968	0.03	[0,1]	0.226	0.903
Check URL bar	69	989	0.07	[0,1]	0.036	0.181
Check Https	54	973	0.22	[0.159,1]	<0.001	<0.001
Visit unknown websites	75	985	0.22	[0.161,1]	<0.001	<0.001
Enter password on link from e-mail	59	956	0.00	[0,1]	0.668	1
Open e-mail from unknown	71	986	0.30	[0.246,1]	<0.001	<0.001
Click on link from unknown	72	988	0.07	[0,1]	0.026	0.181
Share unnecessary personal info	73	986	0.21	[0.148,1]	<0.001	<0.001

Table 7.6 – Fisher's Exact tests comparing expert and non-expert security behavior from the 2024 sample. We excluded "other" responses from the analysis and used the Holm-Bonferroni correction to adjust p-values for multiple testing.

has shown suitability of a certain platform for expert recruiting, e.g. in the case of software developers [496, 260] or non-expert recruiting [415, 501], as this may change over time [501].

7.5.3 Differences in Statistical Analysis

Finally, we discuss the different statistical analyses conducted in each of the iterations of the study, specifically their suitability for the collected data, and we compare effect sizes for the different iterations. The original 2014 study used chi-squared tests to compare frequency of expert and non-expert behavior, and also to compare non-experts' effectiveness and realism advice ratings. P-values were corrected using the Holm-Bonferroni method. In the 2018 replication, when the assumptions (an expected frequency of at least five in each cell of the contingency table) were satisfied, chi-squared tests were still used to compare expert and non-expert behavior. If the assumptions were not satisfied, Fisher's exact tests were used for this comparison. Additional tests comparing non-experts' ratings were omitted. P-values were again corrected using the Holm-Bonferroni method. Finally, in our 2024 replication, we did not have concrete hypotheses for which behaviors we expected expert and non-expert behaviors and advice ratings to differ for, so we focused our analyses on describing data without hypothesis testing. Nevertheless, we repeated the statistical analyses to enable effect size comparisons to prior work.

The chi-squared and Fisher's exact tests comparing experts' and non-experts' behavior do not take into account the order of the answer categories, e.g. the inherent order when asking for the amount of accounts participants use a certain practice for. However, given that many of the behavioral questions included an "I don't know" answer or sometimes an option to give a more detailed

Advice	effectiveness		realism		V	r_{rb}	95% CI	raw p	adj. p
	M	SD	M	SD					
Use anti-virus	4.24	0.95	4.40	0.97	18054	-0.39	[-0.448,-0.325]	<0.001	<0.001
OS updates	4.27	0.90	4.45	0.86	16566	-0.43	[-0.489,-0.371]	<0.001	<0.001
Automatic updates	4.02	1.08	4.27	1.04	18984	-0.49	[-0.542,-0.432]	<0.001	<0.001
Update applications	4.16	0.95	4.40	0.86	17799	-0.52	[-0.567,-0.461]	<0.001	<0.001
Clear cookies	3.79	1.15	4.15	1.08	24534	-0.54	[-0.585,-0.481]	<0.001	<0.001
Check app permissions	4.19	0.95	4.33	0.92	24802	-0.28	[-0.343,-0.209]	<0.001	<0.001
Use unique pws	4.49	0.88	4.27	1.10	41798	0.42	[0.356,0.475]	<0.001	<0.001
Use strong pws	4.68	0.70	4.58	0.80	19986	0.29	[0.227,0.359]	<0.001	<0.001
Don't write down pws on paper	3.58	1.48	3.54	1.58	46832	0.04	[-0.029,0.117]	0.419	1
Save pws locally	2.37	1.40	2.68	1.55	35896	-0.35	[-0.409,-0.279]	<0.001	<0.001
Use pw manager	3.78	1.23	3.73	1.39	44956	0.08	[0.009,0.155]	0.13	1
Write down pws on paper	2.59	1.50	2.69	1.60	26445	-0.17	[-0.239,-0.095]	0.005	0.056
Check if https	4.32	0.94	4.43	0.90	16520	-0.31	[-0.37,-0.238]	<0.001	<0.001
Be careful in general	4.49	0.81	4.52	0.83	12146	-0.11	[-0.179,-0.037]	0.118	1
Be careful with links in e-mails	4.71	0.70	4.71	0.67	6330	0.00	[-0.077,0.067]	0.956	1
Visit only known websites	4.08	1.02	3.89	1.20	50616	0.35	[0.283,0.41]	<0.001	<0.001
Use 2FA	4.62	0.76	4.47	0.88	22838	0.46	[0.397,0.511]	<0.001	<0.001
Don't share private info	4.69	0.67	4.68	0.72	8094	0.04	[-0.033,0.111]	0.633	1
Use hardware token	3.89	1.16	3.43	1.32	60442	0.55	[0.496,0.606]	<0.001	<0.001
Use authenticator app	4.29	0.96	4.10	1.14	34700	0.38	[0.311,0.439]	<0.001	<0.001
Use passkeys	3.98	1.11	3.81	1.27	38542	0.28	[0.212,0.353]	<0.001	<0.001
Use biometrics	4.28	1.03	4.13	1.24	26838	0.31	[0.24,0.373]	<0.001	<0.001
Don't click on links from unknown	4.66	0.72	4.68	0.73	8733	-0.10	[-0.166,-0.024]	0.218	1
Don't enter pws on links from e-mail	4.71	0.69	4.71	0.66	5548	-0.05	[-0.117,0.026]	0.602	1
Check URL bar	4.63	0.72	4.58	0.81	14439	0.20	[0.129,0.267]	0.006	0.061
Don't open email attachments from unknown	4.73	0.64	4.67	0.74	8390	0.24	[0.171,0.307]	0.005	0.056
Use AI advice	2.79	1.30	2.83	1.42	23319	-0.07	[-0.149,0]	0.222	1

Table 7.7 – Wilcoxon signed rank tests comparing non-expert effectiveness and realism ratings of security advice from the 2024 sample. We abbreviate “password” as **pw**. We used the Holm-Bonferroni correction to adjust p-values for multiple testing.

explanation, both of which are hard to incorporate into an ordinal variable, this remains a sensible choice of hypothesis test. To avoid having to specify different types of tests when reporting the results and since the assumptions for the chi-squared test were not met in eight cases, we chose to consistently use the Fisher’s exact test, since limitations regarding computational power that caused the chi-squared test to be recommended instead, do not apply anymore for the sample sizes included in our tests. Due to our fairly large sample size, any small loss of statistical power in comparison to the chi-squared test does not matter.

The use of chi-squared tests for comparing non-experts’ effectiveness and realism ratings in the original 2014 study does not take into account that, since each non-expert rates both the effectiveness and the realism of each piece of advice, the data for this analysis is dependent instead of independent. In addition, since the ratings are given on a scale from 1 to 5, even though there is also an “I don’t know option”, the data can more clearly be considered ordinal data than nominal, when omitting the uncertain answers. Consequently, we used the non-parametric Wilcoxon signed rank test for these comparisons.

For the data from the 2024 iteration of the study, we calculated Cramér’s V effect sizes for the chi-squared tests. It can take on values between 0 and 1 and is interpreted like a correlation coefficient. However, Cramér’s V does not indicate the direction of the effect. For the Wilcoxon signed rank tests, we calculated matched-pairs rankbiserial correlation coefficients (r_{rb}), which take on values between 0 and |1|. The sign of the effect sizes shows the direction of the effect. For our analyses, negative r_{rb} means that mean realism ratings were higher than mean effectiveness ratings and positive r_{rb} means that it is the other way around. As such, even though the type of effect size measure differs, it can be interpreted on the same scale in terms of the strength of the relationship, although directionality of the effect cannot be inferred from Cramér’s V.

Prior work reported test statistics, degrees of freedom and the sample size for each chi-squared test, which can be used to calculate the Cramér’s V effect size (Equation 7.1). In the equation, k represents the number of columns, which for our data are always two, since the two groups of experts and non-experts are compared. Meanwhile, r represents the number of rows and can be derived from the degrees of freedom through $r = df + 1$. This measure is adjusted to account for bias with smaller sample sizes using Equation 7.2, from Bergsma [46].

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}} \quad (7.1)$$

$$\tilde{V} = \sqrt{\frac{\max\left(0, (\chi^2/n) - \frac{(k-1)(r-1)}{n-1}\right)}{\min\left(\left(k - \frac{(k-1)^2}{n-1}\right) - 1, \left(r - \frac{(r-1)^2}{n-1}\right) - 1\right)}} \quad (7.2)$$

The Cramér’s V effect sizes of the comparison between expert and non-expert behavior are in Table 7.8 and the full statistics for 2024 are in Table 7.6. For many of the behaviors, the effect size has become smaller over all three of the iterations, or in some cases, since the second iteration in 2018. Exceptions to this are the behavior to “save passwords locally”, where there was no difference in the prior replications and now a, albeit small, difference was measured, the behavior to “do OS updates”, where the effect size remained similar to 2018, and the behavior to “check https”, where the effect is smaller in 2024 than in the 2014 iteration, but larger than in the 2018

Behavior	2014	2018	2024
Do OS update	0.38	0.11	0.12
Use anti-virus	0.24	0.46	0.22
Remember passwords	0.42	0.30	0.17
Write down passwords on paper	0.20	0.22	0.14
Save passwords locally	0.00	0.00	0.14
Use password manager	0.49	0.38	0.21
Re-use passwords	0.26	0.23	0.13
Use 2FA	0.21	0.25	0.06
Use hardware token			0.20
Use authenticator app			0.20
Use passkeys			0.00
Use biometric auth			0.00
Check app permissions			0.03
Check URL bar	0.32	0.23	0.07
Check Https	0.50	0.08	0.22
Visit unknown websites	0.34	0.35	0.22
Enter password on link from e-mail	0.00	0.41	0.00
Open e-mail from unknown	0.43	0.49	0.30
Click on link from unknown	0.30	0.19	0.07
Share unnecessary personal info			0.21

Table 7.8 – Cramér’s V effect sizes comparing expert and non-expert behavior for the three iterations of the study. For behaviors introduced in the 2024 iteration of the study, there are no effect sizes for earlier iterations.

iteration. Comparing to effect sizes common in USP studies published at CHI, using Table D.7, many of the effect sizes reported in the prior iterations would be judged as medium or large, as they are larger than 0.16 and 0.27 respectively. In contrast, for the 2024 iteration, the only large effect by this measure would be for “opening e-mails from unknown senders”. In summary, this suggests that non-expert and expert self-reported behavior is becoming more similar.

The Cramér’s V and r_{rb} effect sizes of the comparison between non-experts’ effectiveness and realism ratings are in Table 7.9, while the full statistics for the analysis of the 2024 data are in Table 7.7. We calculated two versions of Cramér’s V for the tests from 2014: One using the sample size ($N=294$) as reported by Ion et al. [236] and one using the number of data points in the analysis with each participant contributing two datapoints, one per type of rating, amounting to a total of $N=588$. The latter version underestimates the effect size because the analysis does not take into account the reduced variance based on individual differences from the within-subjects design. While the former version using a smaller N leads to a higher effect size, it is unclear how much this represents the actual effect size, since the amount of within-person variance is unknown. Both versions do not take into account the order of the rating categories from 1 to 5. As such, comparisons between the 2014 and 2024 iteration are hard. For most of the advice, the absolute value of the effect size is larger in 2024 than in either of the two versions for 2014. However, as described above, the tests used in 2014 were meant for independent instead of dependent samples, thus leading to less power

Rated Advice	Cramér's V (2014) N=294	Cramér's V (2014) N=588	r_{rb} (2024) N=990
Use anti-virus	0.19	0.13	-0.39
Install latest OS updates	0.31	0.22	-0.43
Automatic updates	0.39	0.27	-0.49
Update applications	0.34	0.24	-0.52
Clear cookies	0.44	0.31	-0.54
Check app permissions			-0.28
Use unique pws	0.23	0.16	0.42
Use hard to guess pws	0.08	0.06	0.29
Don't write down pws on paper	0.18	0.12	0.04
Save pws locally	0.23	0.17	-0.35
Use pw manager	0.23	0.16	0.08
Write down pws on paper	0.18	0.13	-0.17
Check if HTTPS	0.16	0.11	-0.31
Be careful in general	0.07	0.05	-0.11
Be careful with links in e-mails	0.07	0.05	0.00
Visit only known websites	0.26	0.18	0.35
Use 2FA	0.20	0.14	0.46
Don't share private info			0.04
Use hardware token			0.55
Use authenticator app			0.38
Use passkeys			0.28
Use biometrics			0.31
Don't click on links from unknown	0.00	0.00	-0.10
Don't enter pws on links from e-mail	0.05	0.04	-0.05
Check URL	0.00	0.00	0.20
Don't open email attachments from unknown	0.00	0.00	0.24
Use AI security advice			-0.07

Table 7.9 – Effect sizes for comparisons of non-expert's effectiveness and realism ratings in 2014 and 2024. There are two variants given for Cramér's V from the 2014 iteration: One using the number of participants (N=294) and one using the number of data points in the analysis (N=588). For pieces of advice introduced in the 2024 iteration of the study, there are no effect sizes for 2014.

and likely lower effect sizes. For the advice newly added in 2024, the r_{rb} for the advice regarding authentication methods were positive, meaning the effectiveness was rated higher than the realism of the advice, while the advice “check app permissions” received higher ratings in terms of realism. The other two pieces of newly added advice received similar ratings for effectiveness and realism.

7.6 Conclusion

Ten years after Ion et al.'s study comparing experts' and non-experts' security-related behavior and advice, we conducted a second replication study to monitor the development of security practices over time. In addition to the replication, we investigated how newer advice, such as more recent advances in passwordless authentication, are perceived by experts and non-experts. We found that experts and non-experts are beginning to adopt new security practices, but overall experts' and non-experts' security practices continue to differ. We additionally describe differences and challenges in study design, recruitment and analysis across the three iterations, highlighting the need for researchers to update the methods used, even when conducting a replication study. Ten years are a period that yields relevant change in security behavior both for experts and non-experts and we call for longitudinal studies monitoring such developments in USP, as well as in the methods used.

Chapter 8

Recommendations

Disclaimer:

Parts of this chapter were previously published as part of the following five publications:

1. “Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of Interview and Survey Study on Security Advice”, published at CHI 2023. This paper is the result of a collaboration with my co-authors Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz and Matthew Smith. This chapter uses text from this paper’s recommendations section. An initial draft was written by me and it was iteratively revised by Matthew and me.
2. “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”, published at SOUPS 2023, together with my co-authors Christian Tiefenau and Matthew Smith. This chapter contains text from the paper’s recommendation section. The first version was written by me, and it was revised by Matthew and me.
3. “Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation”, published at CHI 2025, together with my co-authors Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas and Matthew Smith. This chapter contains text from the paper’s recommendation section. I wrote the initial draft of the section, and it was iteratively revised by Matthew and me.
4. “A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen’s d Effect Sizes”, published at CHI 2025 with my co-authors, Julia Angelika Grohs, Simon Lenau, and Matthew Smith. This chapter contains text from the paper’s recommendation section. I wrote the first version of that section, and Matthew, Simon and I iteratively revised it.
5. “Meta-Science in Usable Security and Privacy and HCI”, presented at Meta-HCI: First Workshop on Meta-Research in HCI at CHI’25, on which I collaborated with Florin Martius. This chapter uses text from this publications’ recommendation section. I wrote an

initial draft of the recommendations which were based on my own work, which is the part included in this thesis. Florin wrote the part about the recommendations based on his first-authored paper [321]. We revised the section together.

Work presented in this thesis explored research methods in qualitative and quantitative research, reporting and replications. Based on our work, we present recommendations for research practice in USP, which can also be transferred to an extent to HCI more generally. We separately describe recommendations for qualitative (Section 8.1) and quantitative research (Section 8.2). Many of the recommendations from both these areas apply to replication studies. Transparent reporting of methods and results in both qualitative and quantitative research is especially important when trying to build upon prior work, e.g. by replicating it. As such, we do not make separate recommendations concerning replications.

8.1 Recommendations for Qualitative Research Practice

Based on the work described in Chapter 3, our empirical evaluation of the qualitative coding process and the survey of PC members, we recommend the following to authors aiming to publish qualitative work in USP:

Adapt coding practices to the complexity of research questions and data. When data is simple, research questions are well-defined, and analysis is straightforward, use only a single coder to code the data. Multiple coders are still helpful during the process of establishing a codebook. Use multiple coders when research questions and data are open-ended and complex. In this case, different coders can have bring different perspectives to the analysis and this can lead to identifying additional interesting topics in the data. If multiple coders are participating in the analysis, discussions and interaction between coders are a vital part of the analysis process. Since our studies were designed to be representative of USP studies and included two types of data of a complexity typical for this field, this recommendation applies to USP studies. We are aware that there are many levels of data complexity which our study could not cover. Researchers can use the examples of simple and complex data which we specify in Section 3.2.1 and the description of the data collected in our content-level studies to judge whether their analysis outside USP is comparable to this one, and thus whether this recommendation applies to their study context.

Make clear who the analyzing researchers are and what level of expertise they have. This applies both regarding the topic and the analysis methods used. Since the differences between coders of differing levels of experience were not related to the specific content of our studies but rather to the degree of abstractness and generalization achieved, we believe that this recommendation is also applicable to the broader HCI community, but note that making positionality explicit may be more wide-spread already outside USP.

Which-ever method of analysis you use, describe it in a way that your process can be repeated and explain and justify important methodological decisions. This applies even when it is a well-established method. The sampling frame for our PC member survey was the SOUPS PC of the last five years. As such, this recommendation only represents the expectations of PC members at that conference. Some SOUPS PC members also serve on PCs at other

conferences, including CHI, but since we did not collect data on this aspect, we do not claim that these findings generalize. Given that researcher decisions also lead to different results in quantitative analysis [229], we believe that explicating these decisions is beneficial not only when using qualitative methods but also for quantitative work.

8.2 Recommendations for Quantitative Research Practice

Based on the work described in Chapters 4, 5, and 6, we make recommendations on reporting quantitative analyses, within the NHST framework.

Conduct power analyses. A priori power analysis can be used to determine the sample size needed to detect an effect of a certain size with a specific statistical power and at a specified alpha error level. It is a helpful practice for researchers to think about which size of effect would be practically relevant for a planned study, prior to conducting the study, as it prompts them to consider the possible impact of their research. Based on the resources available to them, researchers can also use the results of the power analysis to decide whether to carry out the study as planned or not. If the goal of a study is to confirm or reject hypotheses, an under-powered study might not reduce the uncertainty enough to reach this goal. Where researchers are also interested in generating hypotheses, or in getting initial insights about a field, even an under-powered study can be worth doing. Chapter 4 discusses power analyses for some tests in more detail. For more complex analyses, even conducting power analyses for comparisons of individual groups in the analysis can help researchers make decision about their study, especially if these sub-group comparisons will be used to answer research questions. Alternatively, simulations can be used to estimate power [296].

Use reporting guidelines. In fields such as psychology and medicine, multiple guidelines exist on how to report statistical results [25, 15, 456]. Using guidelines, such as CONSORT [456], has been shown to improve reporting of study procedures [394]. While some aspects of these guidelines, e.g. statistical genetics [15], 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” [25] (p. 81) should be reported, although exactly what this data encompasses is different across guidelines, and test-specific [25, 15]. Both discuss effect sizes and variability but differ on how they should be reported [25, 15]. Our summary of what is relevant from these guidelines in the context of this thesis 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.

In the following, we specify this general recommendation in more detail.

Specify whether an analysis is confirmatory or exploratory. State whether pre-defined hypotheses are being tested or whether an exploratory analysis is being conducted [516]. This allows others to judge the stability of your claims, while at the same time not impeding new discovery and hypothesis generation.

Report all effect sizes, even those belonging to non-significant tests. To gain a full understanding of the research area and reduce publication bias, effects which do not represent

statistically significant effects are also important to report. To avoid overloading the results section with numbers, consider reporting the full statistics in an appendix.

Report both unitless/standardized and non-standardized effect sizes. Wherever possible, report standardized effect sizes in addition to p-values. For example, Ellis provides an overview of effect sizes in Table 1.1 [136], which can help researchers choose appropriate standardized effect sizes for their analyses. Since standardized effect sizes can be unintuitive and the studies described in Chapter 6 showed that researchers had trouble interpreting standardized effect sizes, also report descriptive statistics showing the effect size, e.g., there was a 1-point difference on a 7-point scale, so that interested researchers can calculate effect sizes themselves. Take into account the following considerations for these descriptive statistics:

- **Report frequencies** within all groups when comparing nominal variables, as effect size calculations commonly use these.
- **Report descriptive statistics for each group** of the independent variable(s) when comparing ordinal or interval variables. Report at least means, standard deviations, and group sizes. If this becomes too extensive, move this information to the appendix or supplemental material.
- **Make sure that the descriptives make sense**, e.g., a mean is not a good summary statistic for a bimodal distribution.

Both non-standardized and standardized forms of effect-size have strengths and weaknesses, so offering both seems beneficial to us.

Make fully anonymized data sets available when needed. If presenting all the descriptive statistics and frequencies is too extensive, e.g., for regression analyses with multiple independent variables, ideally, the anonymized data can be made available. This is especially useful for researchers aiming to replicate published work because it enables them to make more direct comparisons, even for subgroups which were not explicitly analyzed in the original work. Take care not to publish data that can be used to identify participants.

Explain effect size measures. In the studies described in Chapter 6, not all participants were familiar with the effect sizes we investigated, Cohen's d and OR, even though we had chosen ones which we believed to be common. 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 the effect sizes measures.

Interpret and discuss effect sizes. 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 Chapter 6, 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 reader¹. We also explicitly recommend not to rely on conventional scales such as those by Cohen [91], since our results suggest that there is too much context-based variation. We make more specific recommendations about the interpretation of effect sizes in the following.

¹For 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.

Interpret effect sizes within your research context. As the work described in Chapter 5 showed, effect sizes can vary between different research areas and thus it makes sense to compare a found effect size with what is common in the area. The tables in that chapter (e.g. Table 5.4), and in Appendix D.2 provide an initial overview for some research areas, including security and privacy, and for different types of effect sizes. These effect size guidelines are meant as a help to authors, reviewers and readers for judging the relative size of an effect. We want to make absolutely clear that they should not be used to arbitrarily accept or reject research work and that small effects can be highly practically relevant and large effects practically irrelevant.

Discuss effects on an appropriate level of detail based on relevance. Differentiate between the important effect sizes which need to be discussed in detail, e.g. the main hypothesis tests, compared to a potentially large number of secondary tests, e.g. pairwise post-hoc tests. Researcher discretion is needed.

Interpret effect sizes qualitatively based on practical relevance. Apart from stating the size of the effect in the context of the research area, for relevant tests, also discuss the practical relevance. Chapter 6 includes possible influencing factors and interpretation approaches that other researchers used to judge effect sizes.

Interpret Power. As we saw, even small effects can be considered important, but many studies are under-powered. 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.

We highlight two additional, potentially controversial recommendations to encourage discussion:

Move test statistics to supplemental material. Test statistics (e.g. t , u , degrees of freedom) were not explicitly used by the interviewed researchers when interpreting effects size (see Chapter 6), but they are relevant to verify statistical analyses. However, when providing complete information on statistical results, results sections may become cumbersome to read. In our view their main benefit is that they make statistical analyses verifiable, enabling e.g. reviewers to check statistics for coherency [354] and making it a little harder for bad actors to create fake data. As a middle ground between completeness of reporting and readability, these values could be reported in supplemental material or in the appendix, 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. [558] for a perspective criticizing p-values and Murtaugh [339] for a proponents' view on p-values. Even though we held p values constant in an attempt to not confound our study of effect size interpretation (Chapter 6), participants mentioned using p-values in their judgment of effect size and importance. P-values are often the focus of interpretation of results [47], even though they are actually more a measure of uncertainty. Using other ways to represent uncertainty, e.g. confidence intervals, could help researchers focus their interpretations more on effect sizes. The meaning of the interval can and should be contained in the authors interpretation.

Chapter 9

Conclusion

This thesis investigates three areas of meta-science in USP: Qualitative research, quantitative research, and replications.

In the first area, two content-level and two meta-studies with a total of 65 students and 7 researchers were conducted to explore the influence of different qualitative coders and different types of data (simple vs. complex) on qualitative analysis results and a survey with 37 SOUPS PC members was carried out to compare the findings from these studies to reviewers' expectations. Results across multiple coders were more similar for simple data, than complex data, and more experienced researchers produced results that were more abstract and focused than students with less research experience. Many reviewers reported that they would accept various qualitative research practices, as long as authors justified their use.

In the second area, quantitative research, this thesis focuses on two topics, power analysis and effect sizes, both of which are considered in the context of NHST. First, a manual analysis of 54 papers in DCUS, a subfield of USP, showed that many of the analyzed studies are underpowered to detect effects, and that often, not enough data is reported, meaning that researchers cannot conduct power analyses or conduct meta-studies based on published work. This analysis was extended to 1692 quantitative papers published at CHI, i.e. HCI in general, using a LLM-assisted approach. In evaluating the performance of this approach, we consolidated issues with reporting and from the analysis of these HCI papers, we identified differences in common sizes of effects between research areas. We explored the interpretation of effect sizes both through an analysis of 67 quantitative CHI publications, as well as surveys and interviews with 63 researchers in HCI and USP. In the literature analysis, we identified five approaches researchers used to interpret effects in their publications: Comparing test-specific values, assigning size labels, using a statistical and methodological reference frame, comparing different observations and interpretation for the big picture. Through the surveys and interviews, we discovered that researchers' judgments regarding size and importance of effects vary, and we identified additional factors contributing to these judgments, such as context of the study or point of view of those affected by an effect. In these studies, we also noticed misconceptions that researchers have about the effect sizes we investigated.

For the third area, replications, this thesis describes the second replication of a study on expert and non-expert security practices and advice. Ten years after the original study, we found relevant change in security behavior of experts and non-experts, establishing this as a time frame within

which researchers can expect to observe such development. Besides these content-level observations, the thesis showcases on-going and evolving challenges regarding study methodology, recruitment and analysis across the three iterations of the study. These considerations can serve as examples for similar considerations necessary in other studies, both for replications and original studies.

While there have been individual contributions to meta-science in USP for a while, with e.g. replications included in the call for papers for SOUPS since 2012, this thesis promotes a more explicit and intentional pursuit of meta-science in the field of USP. It draws together findings from different areas of meta-science, i.e. methods, reporting and reproducibility meta-science, and uses methods of literature analysis and replicated, as well as original empirical studies to generate insights on the state of practice in USP and HCI research.

From the results of our investigations into these three areas of studies, the thesis draws recommendations on how research practice in USP can be improved. Two recommendations pertain to methods. First, we encourage researchers to take into account what data they are analyzing when making decisions about the number of coders they need for a qualitative analysis. Second, we recommend conducting a priori power analysis to get a sense of necessary sample size before conducting a study. The remaining recommendations concern reporting and specify how researchers can make their methodology as well as results more transparent and useful for other researchers, including recommendations on how to report statistical results. Some of the recommendations, e.g. regarding the amount of coders used in qualitative analysis [595, 597, 50] or specifying positionality of researchers [174, 277, 225], have been applied and referenced in USP research since their publication. To make the recommendations from this thesis, as well as recommendations derived from other meta-research in USP, e.g. Tang et al.'s work on reporting statistical results [500], more accessible to authors submitting their work to USP venues, next steps could include formalizing these recommendations into a set of guidelines, provided to authors and reviewers during the submission process at USP venues.

The thesis shows that some of the meta-scientific findings described therein can be applied not only to USP but also to HCI and possibly to other related research areas, e.g. as described for the effect size judgments of USP and HCI researchers in Chapter 6. However, insights from the scale development process of a social desirability scale suggested differences between measuring social desirability bias in general scenarios, and social desirability bias in relation to security and privacy topics, making a security and privacy specific social desirability scale necessary [2]. In addition, as discussed in Chapter 5, typical sizes of effects in Security and Privacy differ from those in general HCI. This shows that there is nevertheless a need for more USP-specific meta-science in the future. Guidance on research practice from other research fields, e.g. the APA publication manual in psychology [25] or the CONSORT checklist in medicine [456] should be evaluated for its fit for the USP community before including it in USP-specific guidelines. This evaluation should take into account that USP is a comparatively younger field than e.g. psychology, and thus more community engagement is likely necessary to agree on guidelines that fit the whole community as well as possible.

List of References

- [1] Jacob Abbott and Sameer Patil. How Mandatory Second Factor Affects the Authentication User Experience. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [2] Laura Marie Abels, Matthew Smith, and Anna-Marie Ortloff. I never reuse passwords! Development and Validation of a Security and Privacy Social Desirability Scale (SP-SDS) for end users without a background in computer science. In *Proceedings of the Twentyfirst Symposium on Usable Privacy and Security (SOUPS)*, Seattle, USA, 2025. USENIX Association.
- [3] Svetlana Abramova, Rainer Böhme, Helmut Elsinger, Helmut Stix, and Martin Summer. What Can Central Bank Digital Currency Designers Learn from Asking Potential Users? In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 151–170, Anaheim, USA, 2023. USENIX Association.
- [4] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. Comparing the Usability of Cryptographic APIs. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 154–171, San Jose, USA, 2017. IEEE.
- [5] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. You Get Where You’re Looking for: The Impact of Information Sources on Code Security. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 289–305, San Jose, CA, 2016. IEEE.
- [6] Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. You are Not Your Developer, Either: A Research Agenda for Usable Security and Privacy Research Beyond End Users. In *Proceedings of the IEEE Cybersecurity Development (SecDev)*, pages 3–8, Boston, USA, 2016. IEEE.
- [7] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L. Mazurek, and Sascha Fahl. Security Developer Studies with GitHub Users: Exploring a Convenience Sample. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 81–95, Santa Clara, USA, 2017. USENIX Association.
- [8] Yasemin Acar, Christian Stransky, Dominik Wermke, Charles Weir, Michelle L. Mazurek, and Sascha Fahl. Developers Need Support, Too: A Survey of Security Advice for Software Developers. In *Proceedings of the IEEE Cybersecurity Development (SecDev)*, pages 22–26, Cambridge, USA, 2017. IEEE.

-
- [9] Anne Adams and Martina Angela Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.
 - [10] Yongsu Ahn, Yu-Ru Lin, Panpan Xu, and Zeng Dai. ESCAPE: Countering Systematic Errors from Machine’s Blind Spots via Interactive Visual Analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
 - [11] Devdatta Akhawe and Adrienne Porter Felt. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *Proceedings of the 22nd USENIX Security Symposium*. USENIX Association, 2013.
 - [12] Mamtaj Akter, Madiha Tabassum, Nazmus Sakib Miazi, Leena Alghamdi, Jess Kropczynski, Pamela J. Wisniewski, and Heather Lipford. Evaluating the Impact of Community Oversight for Managing Mobile Privacy and Security. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 437–456, Anaheim, USA, 2023. USENIX Association.
 - [13] Eman Alhelali, Kopo M. Ramokapane, and Jose Such. Multiuser Privacy and Security Conflicts in the Cloud. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
 - [14] Ahmed Alshehri, Eugin Pahk, Joseph Spielman, Jacob T Parker, Benjamin Gilbert, and Chuan Yue. Exploring the Negotiation Behaviors of Owners and Bystanders over Data Practices of Smart Home Devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
 - [15] 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. Recommendations for Statistical Reporting in Cardiovascular Medicine: A Special Report From the American Heart Association. *Circulation*, 144(4):e70–e91, 2021.
 - [16] Joram Amador, Yiran Ma, Summer Hasama, Eshaan Lumba, Gloria Lee, and Eleanor Birrell. Prospects for Improving Password Selection. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 263–282, Anaheim, USA, 2023. USENIX Association.
 - [17] Sabrina Amft, Sandra Höltervennhoff, Nicolas Huaman, Yasemin Acar, and Sascha Fahl. “Would You Give the Same Priority to the Bank and a Game? I Do Not!” Exploring Credential Management Strategies and Obstacles during Password Manager Setup. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, Anaheim, USA, 2023. USENIX Association.
 - [18] Judd Antin and Aaron Shaw. Social desirability bias and self-reports of motivation: A study of amazon mechanical turk in the US and India. In *Proceedings of the 2012 CHI Conference on Human Factors in Computing Systems*, pages 2925–2934, Austin, USA, 2012. ACM.

- [19] Farid Anvari, Rogier Kievit, Daniël Lakens, Charlotte R Pennington, Andrew K Przybylski, Leo Tiokhin, Brenton M Wiernik, and Amy Orben. Not All Effects Are Indispensable: Psychological Science Requires Verifiable Lines of Reasoning for Whether an Effect Matters. *Perspectives on Psychological Science*, 18(2):503–507, 2023.
- [20] Farid Anvari and Daniël Lakens. The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, 3(3):266–286, 2018.
- [21] Farid Anvari and Daniël Lakens. Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, 96:104159, 2021.
- [22] Afaq Ashraf, Taha, Nida ul Habib Bajwa, Cornelius J. König, Mobin Javed, and Maryam Mustafa. "Stalking is immoral but not illegal": Understanding Security, Cyber Crimes and Threats in Pakistan. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 37–56, Anaheim, USA, 2023. USENIX Association.
- [23] Hala Assal and Sonia Chiasson. Security in the Software Development Lifecycle. In *Proceedings of the Fourteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 281–296, Baltimore, USA, 2018. USENIX Association.
- [24] Hala Assal and Sonia Chiasson. 'Think secure from the beginning': A survey with software developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [25] American Psychological Association, editor. *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, DC, seventh edition, 2020.
- [26] Donald R. Atkinson, Michael J. Furlong, and Bruce E. Wampold. Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology*, 29(2):189–194, 1982.
- [27] Peter C. Austin and Ewout W. Steyerberg. The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, 68(6):627–636, 2015.
- [28] Nathaniel Ayewah and William Pugh. A report on a survey and study of static analysis users. In *Proceedings of the 2008 Workshop on Defects in Large Software Systems*, page 5, Seattle, USA, 2008. ACM.
- [29] Stefan Baack. A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pages 2199–2208, Rio de Janeiro, Brazil, 2024. ACM.
- [30] Peter Bacchetti. Peer review of statistics in medical research: The other problem. *BMJ : British Medical Journal*, 324(7348):1271–1273, 2002.
- [31] Peter Bacchetti. Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine*, 8(1), 2010.
- [32] Thom Baguley. Standardized or simple effect size: What should be reported? *British Journal of Psychology*, 100(3):603–617, 2009.

- [33] Khadija Baig, Elisa Kazan, Kalpana Hundlani, Sana Maqsood, and Sonia Chiasson. Replication: Effects of Media on the Mental Models of Technical Users. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, pages 119–138. USENIX Association, 2021.
- [34] Khadija Baig, Reham Mohamed, Anna-Lena Theus, and Sonia Chiasson. "I'm hoping they're an ethical company that won't do anything that I'll regret": Users Perceptions of At-home DNA Testing Companies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [35] Christopher Bailey, Elaine Pearson, and Voula Gkatzidou. Measuring and comparing the reliability of the structured walkthrough evaluation method with novices and experts. In *Proceedings of the 11th Web for All Conference (W4A)*, Seoul, South Korea, 2014. ACM.
- [36] Arthur Bakker, Jinfa Cai, Lyn English, Gabriele Kaiser, Vilma Mesa, and Wim Van Dooren. Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, 102:1–8, 2019.
- [37] Marjan Bakker, Coosje L. S. Veldkamp, Olmo R. van den Akker, Marcel A. L. M. van Assen, Elise Crompvoets, How Hwee Ong, and Jelte M. Wicherts. Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLOS ONE*, 15(7):e0236079, 2020.
- [38] Nick Ballou, Vivek R. Warriar, and Sebastian Deterding. Are You Open? A Content Analysis of Transparency and Openness Guidelines in HCI Journals. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. ACM.
- [39] Titus Barik, Justin Smith, Kevin Lubick, Elisabeth Holmes, Jing Feng, Emerson Murphy-Hill, and Chris Parnin. Do Developers Read Compiler Error Messages? In *Proceedings of the IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, pages 575–585, Buenos Aires, Argentina, 2017. IEEE.
- [40] Jack J. Baroudi and Wanda J. Orlikowski. The problem of statistical power in MIS research. *MIS Quarterly*, 13(1):87–106, 1989.
- [41] David Barrera, Christopher Bellman, and Paul Van Oorschot. Security Best Practices: A Critical Analysis Using IoT as a Case Study. *ACM Transactions on Privacy and Security*, 26(2):13:1–13:30, 2023.
- [42] Dirk Bassler, Matthias Briel, Victor M. Montori, Melanie Lane, Paul Glasziou, Qi Zhou, Diane Heels-Ansdell, Stephen D. Walter, Gordon H. Guyatt, and the STOPIT-2 Study Group. Stopping Randomized Trials Early for Benefit and Estimation of Treatment Effects: Systematic Review and Meta-regression Analysis. *JAMA : the Journal of the American Medical Association*, 303(12):1180–1187, 2010.
- [43] Kathy Baxter, Catherine Courage, and Kelly Caine. *Understanding Your Users. A Practical Guide to User Research Methods*. Morgan Kaufmann Publications, 2nd edition, 2015.

- [44] Tom Beckmann, Patrick Rein, Stefan Ramson, Joana Bergsiek, and Robert Hirschfeld. Structured Editing for All: Deriving Usable Structured Editors from Grammars. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [45] Donald D Bergh, Barton M Sharp, Herman Aguinis, and Ming Li. Is there a credibility crisis in strategic management research? Evidence on the reproducibility of study findings. *Strategic Organization*, 15(3):423–436, 2017.
- [46] Wicher Bergsma. A bias-correction for Cramér’s V and Tschuprow’s T. *Journal of the Korean Statistical Society*, 42(3):323–328, 2013.
- [47] Lonni Besançon and Pierre Dragicevic. The Continued Prevalence of Dichotomous Inferences at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [48] Scott Bezeau and Roger Graves. Statistical Power and Effect Sizes of Clinical Neuropsychology Research. *Journal of Clinical and Experimental Neuropsychology*, 23(3):399–406, 2001.
- [49] Sruti Bhagavatula, Lujo Bauer, and Apu Kapadia. "Adulthood is trying each of the same six passwords that you use for everything": The Scarcity and Ambiguity of Security Advice on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):264:1–264:27, 2022.
- [50] Divyanshu Bhardwaj, Sumair Izaj Hashmi, Katharina Krombholz, and Maximilian Golla. Understanding How Users Prepare for and React to Smartphone Theft. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security)*, Seattle, USA, 2025. USENIX Association.
- [51] Sarah Bird, Ilana Segall, and Martin Lopatka. Replication: Why We Still Can’t Browse in Peace: On the Uniqueness and Reidentifiability of Web Browsing Histories. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 489–503. USENIX Association, 2020.
- [52] Pernille Bjørn and Steven Drucker. Guide to a Successful Submission. <https://web.archive.org/web/20210614064453/https://chi2021.acm.org/for-authors/presenting/papers/guide-to-a-successful-submission>, 2020.
- [53] Erik Blair. A reflexive exploration of two qualitative data coding techniques. *Journal of Methods and Measurement in the Social Sciences*, 6(1):14–29, 2015.
- [54] Cara Bloom, Joshua Tan, Javed Ramjohn, and Lujo Bauer. Self-driving cars and data collection: Privacy perceptions of networked autonomous vehicles. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 357–375, 2017.
- [55] Andrew Booth, Karin Hannes, Angela Harden, Jane Noyes, Janet Harris, and Allison Tong. COREQ (Consolidated Criteria for Reporting Qualitative Studies). In David Moher, Douglas G. Altman, Kenneth F. Schulz, Iveta Simera, and Elizabeth Wager, editors, *Guidelines for Reporting Health Research: A User’s Manual*, pages 214–226. John Wiley & Sons, Ltd, Oxford, UK, 2014.

- [56] Nele Borgert, Oliver D. Reithmaier, Luisa Jansen, Larina Hillemann, Ian Hussey, and Malte Elson. 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*, Hamburg, Germany, 2023. ACM.
- [57] Emanuela Boros, Nhu Khoa Nguyen, Gaël Lejeune, and Antoine Doucet. Assessing the impact of OCR noise on multilingual event detection over digitised documents. *International Journal on Digital Libraries*, 23(3):241–266, 2022.
- [58] Frank A Bosco, Herman Aguinis, Kulraj Singh, James G Field, and Charles A Pierce. Correlational effect size benchmarks. *Journal of Applied Psychology*, 100(2), 2015.
- [59] Paul Bossauer, Thomas Neifer, Gunnar Stevens, and Christina Pakusch. Trust versus Privacy: Using Connected Car Data in Peer-to-Peer Carsharing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [60] Elijah Robert Bouma-Sims, Megan Li, Yanzi Lin, Adia Sakura-Lemessy, Alexandra Nisenoff, Ellie Young, Eleanor Birrell, Lorrie Faith Cranor, and Hana Habib. A US-UK Usability Evaluation of Consent Management Platform Cookie Consent Interface Design on Desktop and Mobile. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [61] Riccardo Bovo, Daniele Giunchi, Ludwig Sidenmark, Joshua Newn, Hans Gellersen, Enrico Costanza, and Thomas Heinis. Speech-Augmented Cone-of-Vision for Exploratory Data Analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [62] Nicholas David Bowman and Justin Robert Keene. A Layered Framework for Considering Open Science Practices. *Communication Research Reports*, 35(4):363–372, 2018.
- [63] Maia J. Boyd, Jamar L. Sullivan Jr., Marshini Chetty, and Blase Ur. Understanding the Security and Privacy Advice Given to Black Lives Matter Protesters. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. ACM.
- [64] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.
- [65] Virginia Braun and Victoria Clarke. *Successful Qualitative Research: A Practical Guide for Beginners*. SAGE, 2013.
- [66] Larissa Braz, Enrico Fregnan, Gül Çalikli, and Alberto Bacchelli. Why Don’t Developers Detect Improper Input Validation? ’; DROP TABLE Papers; –. In *Proceedings of the IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 499–511, Madrid, Spain, 2021. IEEE.
- [67] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.
- [68] John Brooke. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry*, 189(194):4–10, 1996.

- [69] Margaret E. Brooks, Dev K. Dalal, and Kevin P. Nolan. Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, 99(2):332–340, 2014.
- [70] Emeline Brulé, Brianna J. Tomlinson, Oussama Metatla, Christophe Jouffrais, and Marcos Serrano. Review of quantitative empirical evaluations of technology for people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [71] Marc Brysbaert and Michaël Stevens. Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, 1(1), 2018.
- [72] Matthew K. Burns, Anne F. Zaslowsky, Rebecca Kanive, and David C. Parker. Meta-Analysis of Incremental Rehearsal Using Phi Coefficients to Compare Single-Case and Group Designs. *Journal of Behavioral Education*, 21(3):185–202, 2012.
- [73] Loraine Busetto, Wolfgang Wick, and Christoph Gumbinger. How to use and assess qualitative research methods. *Neurological Research and Practice*, 2(1), 2020.
- [74] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: No one can hack my mind revisiting a study on expert and non-expert security practices and advice. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security (SOUPS)*, Santa Clara, USA, 2019. USENIX Association.
- [75] Danilo Bzdok, Denis Engemann, and Bertrand Thirion. Inference and Prediction Diverge in Biomedicine. *Patterns*, 1(8):100119, 2020.
- [76] Ashley A. Cain, Morgan E. Edwards, and Jeremiah D. Still. An exploratory study of cyber hygiene behaviors and knowledge. *Journal of Information Security and Applications*, 42:36–45, 2018.
- [77] Kelly Caine. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 981–992, San Jose, USA, 2016. ACM.
- [78] Paul Cairns. 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, 2007. British Computer Society.
- [79] Stephane Champely. Pwr: Basic functions for power analysis, 2020.
- [80] Jake Chanenson, Brandon Sloane, Navaneeth Rajan, Amy Morril, Jason Chee, Danny Yuxing Huang, and Marshini Chetty. Uncovering Privacy and Security Challenges In K-12 Schools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [81] Kathy Charmaz. *Constructing Grounded Theory*. SAGE, 2nd edition, 2014.
- [82] Yiqun T. Chen, Angela D. R. Smith, Katharina Reinecke, and Alexandra To. Why, when, and from whom: Considerations for collecting and reporting race and ethnicity data in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.

- [83] Eunyong Cheon, Jun Ho Huh, and Ian Oakley. GestureMeter: Design and Evaluation of a Gesture Password Strength Meter. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [84] Geumhwan Cho, Sungsu Kwag, Jun Ho Huh, Bedeuro Kim, Choong-Hoon Lee, and Hyounghick Kim. Towards Usable and Secure Location-based Smartphone Authentication. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security*. USENIX Association, 2021.
- [85] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. "Nobody speaks that fast!" an empirical study of speech rate in conversational agents for people with vision impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [86] Evgenia Christoforou, Gianluca Demartini, and Jahna Otterbacher. Generative AI in Crowdwork for Web and Social Media Research: A Survey of Workers at Three Platforms. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media*, volume 18, pages 2097–2103, Buffalo, USA, 2024.
- [87] Christopher Clarke, Jingnan Xu, Ye Zhu, Karan Dharamshi, Harry McGill, Stephen Black, and Christof Lutteroth. FakeForward: Using Deepfake Technology for Feedforward Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [88] Nicola J. Clarke, Martin E. H. Willis, Jemima S. Barnes, Nick Caddick, John Cromby, Hilary McDermott, and Gareth Wiltshire. Analytical Pluralism in Qualitative Research: A Meta-Study. *Qualitative Research in Psychology*, 12(2):182–201, 2015.
- [89] Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8):70–79, 2020.
- [90] Andy Cockburn, Carl Gutwin, and Alan Dix. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, Canada, 2018. ACM.
- [91] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, Hillsdale, N.J, 2nd edition, 1988.
- [92] Jacob Cohen. A power primer. In A. E. Kazdin, editor, *Methodological Issues and Strategies in Clinical Research*, pages 279–284. American Psychological Association, Washington, DC, US, 2016.
- [93] Nissim Cohen and Tamar Arieli. Field research in conflict environments: Methodological challenges and snowball sampling. *Journal of Peace Research*, 48(4):423–435, 2011.
- [94] Shaanan Cohny, Ross Teixeira, Anne Kohlbrenner, Arvind Narayanan, Mihir Kshirsagar, Yan Shvartzshnaider, and Madelyn Sanfilippo. Virtual Classrooms and Real Harms: Remote Learning at {U.S}. Universities. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, pages 653–674. USENIX Association, 2021.

- [95] David A. Cook and Rose Hatala. Got power? A systematic review of sample size adequacy in health professions education research. *Advances in Health Sciences Education*, 20(1):73–83, 2015.
- [96] Kovila Coopamootoo and Thomas Gross. A Systematic Evaluation of Evidence-Based Methods in Cyber Security User Studies. Technical Report CS_TR-1518, Newcastle University School of Computing, July 2019.
- [97] Emmanuel Courtoux, Caroline Appert, and Olivier Chapuis. SurfAirs: Surface + Mid-air Input for Large Vertical Displays. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [98] Harald Cramér. *Mathematical Methods of Statistics*. Princeton University Press, 1946.
- [99] Lorrie Faith Cranor. A Framework for Reasoning About the Human in the Loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security*, Berkeley, USA, 2008. USENIX Association.
- [100] Lorrie Faith Cranor. Symposium on Usable Privacy and Security. Call for Papers, Poster, And Proposals. <https://cups.cs.cmu.edu/soups/2012/cfp.html>, 2012.
- [101] Noelle M. Crooks, Anna N. Bartel, and Martha W. Alibali. Conceptual Knowledge of Confidence Intervals in Psychology Undergraduate and Graduate Students. *Statistics Education Research Journal*, 18(1):46–62, 2019.
- [102] Sophia Crüwell, Johnny van Doorn, Alexander Etz, Matthew C. Makel, Hannah Moshontz, Jesse C. Niebaum, Amy Orben, Sam Parsons, and Michael Schulte-Mecklenbeck. Seven Easy Steps to Open Science. *Zeitschrift für Psychologie*, 227(4):237–248, 2019.
- [103] Wenzhe Cui, Rui Liu, Zhi Li, Yifan Wang, Andrew Wang, Xia Zhao, Sina Rashidian, Furqan Baig, IV Ramakrishnan, Fusheng Wang, and Xiaojun Bi. GlanceWriter: Writing Text by Glancing Over Letters with Gaze. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [104] Hai Dang, Sven Goller, Florian Lehmann, and Daniel Buschek. Choice Over Control: How Users Write with Large Language Models using Diegetic and Non-Diegetic Prompting. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [105] Anastasia Danilova, Stefan Horstmann, Matthew Smith, and Alena Naiakshina. Testing time limits in screener questions for online surveys with programmers. In *Proceedings of the 44th International Conference on Software Engineering*, pages 2080–2090, Pittsburgh, USA, 2022. ACM.
- [106] Anastasia Danilova, Alena Naiakshina, Johanna Deuter, and Matthew Smith. Replication: On the Ecological Validity of Online Security Developer Studies: Exploring Deception in a {Password-Storage} Study with Freelancers. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 165–183. USENIX Association, 2020.

- [107] Anastasia Danilova, Alena Naiakshina, Stefan Horstmann, and Matthew Smith. Do you Really Code? Designing and Evaluating Screening Questions for Online Surveys with Programmers. In *Proceedings of the IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 537–548, Madrid, Spain, 2021. IEEE.
- [108] Anastasia Danilova, Alena Naiakshina, Anna Rasgauski, and Matthew Smith. Code Reviewing as Methodology for Online Security Studies with Developers - a Case Study with Freelancers on Password Storage. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, pages 397–416. USENIX Association, 2021.
- [109] Anastasia Danilova, Alena Naiakshina, and Matthew Smith. One size does not fit all: A grounded theory and online survey study of developer preferences for security warning types. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, pages 136–148, Seoul, South Korea, 2020. ACM.
- [110] 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. The widespread misuse of effect sizes. *Journal of Science and Medicine in Sport*, 20(5):446–450, 2017.
- [111] Andrew Gary Darwin Holmes. Researcher Positionality - A Consideration of Its Influence and Place in Qualitative Research - A New Researcher Guide. *Shanlax International Journal of Education*, 8(4):1–10, 2020.
- [112] Sanchari Das, Andrew Kim, Shrirang Mare, Joshua Streiff, and L. Jean Camp. Security Mandates are Pervasive: An Inter-School Study on Analyzing User Authentication Behavior. In *Proceedings of the IEEE 5th International Conference on Collaboration and Internet Computing*, pages 306–313, Los Angeles, USA, 2019. IEEE.
- [113] Elmira Deldari, Diana Freed, Julio Poveda, and Yaxing Yao. An Investigation of Teenager Experiences in Social Virtual Reality from Teenagers’, Parents’, and Bystanders’ Perspectives. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, Anaheim, USA, 2023. USENIX Association.
- [114] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. "Yours is better!": Participant response bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1321–1330, Austin, USA, 2012. ACM.
- [115] Yuntian Deng, David Rosenberg, and Gideon Mann. Challenges in End-to-End Neural Scientific Table Recognition. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 894–901, Sydney, Australia, 2019. IEEE.
- [116] Erik Derr, Sven Bugiel, Sascha Fahl, Yasemin Acar, and Michael Backes. Keep me Updated: An Empirical Study of Third-Party Library Updatability on Android. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 2187–2200, Dallas Texas USA, 2017. ACM.
- [117] Jayati Dev, Bahman Rashidi, and Vaibhav Garg. Models of Applied Privacy (MAP): A Persona Based Approach to Threat Modeling. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.

- [118] Julian di Stephano. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. *Functional Ecology*, 17(5):707–709, 2003.
- [119] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. Investigating System Operators’ Perspective on Security Misconfigurations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1272–1289, Toronto, Canada, 2018. ACM.
- [120] Leonie Disch, Angela Fessl, and Viktoria Pammer-Schindler. Designing for Knowledge Construction to Facilitate the Uptake of Open Science: Laying out the Design Space. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New Orleans, USA, 2022. ACM.
- [121] Verena Distler. The Influence of Context on Response to Spear-Phishing Attacks: An In-Situ Deception Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [122] Verena Distler, Matthias Fassl, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Lorrie Faith Cranor, and Vincent Koenig. A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research. *ACM Transactions on Computer-Human Interaction*, 28(6):43:1–43:50, 2021.
- [123] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2623–2634, San Jose, USA, 2016. ACM.
- [124] Pierre Dragicevic. My Technique is 20% Faster: Problems with Reports of Speed Improvements in HCI. Report, INRIA Saclay Ile de France, Paris, France, 2016.
- [125] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [126] Thorsten Dresing and Thorsten Pehl. *Praxisbuch Interview, Transkription & Analyse: Anleitungen und Regelsysteme für qualitativ Forschende*. Eigenverlag, Marburg, 7th edition, 2017.
- [127] Ellen A Drost. Validity and Reliability in Social Science Research. *Education Research and Perspectives*, 38(1):105–123, 2011.
- [128] Isabelle F. Dufour and Marie-Claude Richard. Theorizing from secondary qualitative data: A comparison of two data analysis methods. *Cogent Education*, 6(1), 2019.
- [129] William P. Dunlap. Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin*, 116(3):509–511, 1994.
- [130] Tessa Eagle, Leya Breanna Baltaxe-Admony, Lee Taber, and Kathryn E. Ringland. "[CHI2023] Decisions on paper 42 for CHI2023": A Playful Twist on the Peer Review Process and Methodological Gaslighting. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg Germany, 2023. ACM.

- [131] Florian Echtler and Maximilian Häußler. Open Source, Open Science, and the Replication Crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, Canada, 2018. ACM.
- [132] W. Keith Edwards, Erika Shehan Poole, and Jennifer Stoll. Security Automation Considered Harmful? In *Proceedings of the 2007 Workshop on New Security Paradigms*, pages 33–42, New Hampshire, USA, 2008. ACM.
- [133] Alexander Eiselmayer. Supporting the Design and Analysis of HCI Experiments. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [134] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. Touchstone2: An interactive environment for exploring trade-offs in HCI experiment design. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [135] Passant Elagroudy, Mohamed Khamis, Florian Mathis, Diana Irmischer, Ekta Sood, Andreas Bulling, and Albrecht Schmidt. Impact of Privacy Protection Methods of Lifelogs on Remembered Memories. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [136] Paul D. Ellis. *The Essential Guide to Effect Sizes. Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, 2010.
- [137] Pardis Emami-Naeini, Joseph Breda, Wei Dai, Tadayoshi Kohno, Kim Laine, Shwetak Patel, and Franziska Roesner. Understanding People’s Concerns and Attitudes Toward Smart Cities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [138] David Evans and Giovanni Vigna. IEEE Symposium on Security and Privacy 2010 - Call For Papers. <https://www.ieee-security.org/TC/SP2010/archived/cfp.html>, 2009.
- [139] Stefanos Evripidou, Uchenna D. Ani, Stephen Hailes, and Jeremy D. McK Watson. Exploring the Security Culture of Operational Technology (OT) Organisations: The Role of External Consultancy in Overcoming Organisational Barriers. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 113–129, Anaheim, USA, 2023. USENIX Association.
- [140] Michael Fagan and Mohammad Maifi Hasan Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Proceedings of the Twelfth Symposium on Usable Privacy and Security (SOUPS)*, pages 59–75, Denver, USA, 2016. USENIX Association.
- [141] Michael Fagan, Mohammad Maifi Hasan Khan, and Ross Buck. A study of users’ experiences and beliefs about software update messages. *Computers in Human Behavior*, 51:504–519, 2015.
- [142] Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS)*, Newcastle, UK, 2013. ACM.

- [143] Sascha Fahl, Marian Harbach, Henning Perl, Markus Koetter, and Matthew Smith. Rethinking SSL development in an appified world. In *Proceedings of the ACM SIGSAC Conference on Computer & Communications Security (CCS)*, pages 49–60, Berlin, Germany, 2013. ACM.
- [144] Sascha Fahl, Florian Schaub, Jan H. Klemmer, Juliane Schmüser, Byron M Lowens, Jan-Ulrich Holtgrave, and Jacques Suray. Transparency @ SOUPS. Workshop on Developing Research Transparency Guidelines in Human-centered Privacy and Security. <https://transparency-at-soups.teamusec.de/>, 2025.
- [145] Norman Fairclough. Critical discourse analysis. In *The Routledge Handbook of Discourse Analysis*, pages 9–34. Routledge, 2013.
- [146] Shariff AM Faleel, Yishuo Liu, Roya A Cody, Bradley Rey, Linghao Du, Jiangyue Yu, Da-Yuan Huang, Pourang Irani, and Wei Li. T-Force: Exploring the Use of Typing Force for Three State Virtual Keyboards. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [147] Jacqui Fashimpaur, Amy Karlson, Tanya R. Jonker, Hrvoje Benko, and Aakar Gupta. Investigating Wrist Deflection Scrolling Techniques for Extended Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [148] Matthias Fassl and Katharina Krombholz. Why I Can’t Authenticate — Understanding the Low Adoption of Authentication Ceremonies with Autoethnography. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [149] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, 2007.
- [150] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Paweł W. Woźniak, and Albrecht Schmidt. The Role of HCI in Reproducible Science: Understanding, Supporting and Motivating Core Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [151] Christopher J Ferguson. An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40(5):532–538, 2009.
- [152] Piyum Fernando and Stacey Kuznetsov. OSch in the Wild: Dissemination of Open Science Hardware and Implications for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [153] Andy Field, Jeremy Miles, and Zoe Field. *Discovering Statistics Using R*. SAGE Publications Ltd, London ; Thousand Oaks, Calif, 1st edition, 2012.
- [154] Andy P. Field and Raphael Gillett. How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63(3):665–694, 2010.

-
- [155] Paul D. S. Fink, Velin Dimitrov, Hiroshi Yasuda, Tiffany L. Chen, Richard R. Corey, Nicholas A. Giudice, and Emily S. Sumner. Autonomous is Not Enough: Designing Multisensory Mid-Air Gestures for Vehicle Interactions Among People with Visual Impairments. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [156] Felix Fischer, Yannick Stachelscheid, and Jens Grossklags. The Effect of Google Search on Software Security: Unobtrusive Security Interventions via Content Re-ranking. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3070–3084, Seoul, South Korea, 2021. ACM.
- [157] Felix Fischer, Huang Xiao, Ching-Yu Kao, Yannick Stachelscheid, Benjamin Johnson, Danial Razar, Paul Fawkesley, Nat Buckley, Konstantin Bottinger, Paul Muntean, and Jens Grossklags. Stack Overflow Considered Helpful! Deep Learning Security Nudges Towards Stronger Cryptography. In *Proceedings of the 28th USENIX Security Symposium (USENIX Security)*, pages 339–356, Santa Clara, USA, 2019. USENIX Association.
- [158] Diana Freed, Natalie N. Bazarova, Sunny Consolvo, Eunice J Han, Patrick Gage Kelley, Kurt Thomas, and Dan Cosley. Understanding Digital-Safety Experiences of Youth in the U.S. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [159] Boris Freidlin and Edward L Korn. Stopping clinical trials early for benefit: Impact on estimation. *Clinical Trials*, 6(2):119–125, 2009.
- [160] Luanne Freund and Barbara M. Wildemuth. Documenting and studying the use of assigned search tasks: RepAST. *Proceedings of the American Society for Information Science and Technology*, 51(1):1–4, 2014.
- [161] Catherine O. Fritz, Peter E. Morris, and Jennifer J. Richler. Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, 141(1):2–18, 2012.
- [162] Nollaig Frost. ‘Do you know what I mean?’: The use of a pluralistic narrative analysis approach in the interpretation of an interview. *Qualitative Research*, 9(1):9–29, 2009.
- [163] Nollaig Frost, Sevasti Melissa Nolas, Belinda Brooks-Gordon, Cigdem Esin, Amanda Holt, Leila Mehdizadeh, and Pnina Shinebourne. Pluralism in qualitative research: The impact of different researchers and qualitative approaches on the analysis of qualitative data. *Qualitative Research*, 10(4):441–460, 2010.
- [164] Kelsey R Fulton, Anna Chan, Daniel Votipka, Michael Hicks, and Michelle L Mazurek. Benefits and Drawbacks of Adopting a Secure Programming Language: Rust as a Case Study. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*. USENIX Association, 2021.
- [165] Kelsey R. Fulton, Joseph Lewis, Nathan Malkin, and Michelle L. Mazurek. Write, Read, or Fix? Exploring Alternative Methods for Secure Development Studies. In *Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS)*, pages 81–100, Philadelphia, USA, 2024. USENIX Association.

- [166] David C. Funder and Daniel J. Ozer. Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2):156–168, 2019.
- [167] Dominic Furniss, Ann Blandford, and Paul Curzon. Confessions from a grounded theory PhD: Experiences and lessons learnt. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 113–122, Vancouver, Canada, 2011. ACM.
- [168] Maria Gäde, Marijn Koolen, Mark Hall, Toine Bogers, and Vivien Petras. A Manifesto on Resource Re-Use in Interactive Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 141–149, Canberra, Australia, 2021. ACM.
- [169] Anirudh Ganesh, Chinenye Ndulue, and Rita Orji. Tailoring a Persuasive Game to Promote Secure Smartphone Behaviour. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [170] Simson Garfinkel and Heather Richter Lipford. Usable Security: History, Themes, and Challenges. In Simson Garfinkel and Heather Richter Lipford, editors, *Usable Security: History, Themes, and Challenges*. Springer International Publishing, Cham, 2014.
- [171] Christine Geeng, Mike Harris, Elissa Redmiles, and Franziska Roesner. "Like Lesbians Walking the Perimeter": Experiences of {U.S}. {LGBTQ+} Folks With Online Security, Safety, and Privacy Advice. In *Proceedings of the 31st USENIX Security Symposium (USENIX Security)*, pages 305–322, Boston, USA, 2022. USENIX Association.
- [172] Lisa Geierhaas, Florin Martius, Arthi Arumugam, and Matthew Smith. "Not the Right Question?" A Study on Attitudes Toward Client-Side Scanning with Security and Privacy Researchers and a U.S. Population Sample. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 2246–2263, San Francisco, USA, 2025. IEEE.
- [173] Lisa Geierhaas, Anna-Marie Ortloff, Matthew Smith, and Alena Naiakshina. Let's Hash: Helping Developers with Password Security. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 503–522, Boston, USA, 2022. USENIX Association.
- [174] Dañiel Gerhardt, Matthias Fassel, Carolyn Guthoff, Adrian Dabrowski, and Katharina Krombholz. AirTag-Facilitated Stalking Protection: Evaluating Unwanted Tracking Notifications and Tracker Locating Features. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security)*, Seattle, USA, 2025. USENIX Association.
- [175] Eva Gerlitz, Maximilian Häring, Charlotte Theresa Mädler, Matthew Smith, and Christian Tiefenau. Adventures in Recovery Land: Testing the Account Recovery of Popular Websites When the Second Factor is Lost. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 227–243, Anaheim, USA, 2023. USENIX Association.
- [176] Eva Gerlitz, Maximilian Häring, and Matthew Smith. Please do not use !?_ or your License Plate Number: Analyzing Password Policies in German Companies. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, pages 17–36. USENIX Association, 2021.

- [177] Eva Gerlitz, Maximilian Häring, Matthew Smith, and Christian Tiefenau. Evolution of Password Expiry in Companies: Measuring the Adoption of Recommendations by the German Federal Office for Information Security. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 191–210, Anaheim, USA, 2023. USENIX Association.
- [178] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. Is FIDO2 the Kingslayer of User Authentication? A Comparative Usability Study of FIDO2 Passwordless Authentication. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 268–285, San Francisco, USA, 2020. IEEE.
- [179] Magdalena Glas, Manfred Vielberth, and Guenther Pernul. Train as you Fight: Evaluating Authentic Cybersecurity Training in Cyber Ranges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [180] Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. What is Sensitive About (Sensitive) Data? Characterizing Sensitivity and Intimacy with Google Assistant Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [181] Steven Goodman. A dirty dozen: Twelve P-value misconceptions. *Seminars in Hematology*, 45(3):135–140, 2008.
- [182] Steven N. Goodman and Jesse A. Berlin. The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting Results. *Annals of Internal Medicine*, 121(3):200–206, 1994.
- [183] Peter Leo Gorski, Yasemin Acar, Luigi Lo Iacono, and Sascha Fahl. Listen to Developers! A Participatory Design Study on Security Warnings for Cryptographic APIs. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [184] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Moeller, Yasemin Acar, and Sascha Fahl. Developers Deserve Security Warnings, Too. In *Proceedings of the Fourteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 265–281, Baltimore, USA, 2018. USENIX Association.
- [185] Matthew Green and Matthew Smith. Developers are Not the Enemy!: The Need for Usable Security APIs. *IEEE Security & Privacy*, 14(5):40–46, 2016.
- [186] Samuel B. Green. How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behavioral Research*, 26(3):499–510, 1991.
- [187] Saul Greenberg and Bill Buxton. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 111–120, Florence, Italy, 2008. ACM.
- [188] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350, 2016.

- [189] Thomas Groß. Fidelity of Statistical Reporting in 10 Years of Cyber Security User Studies. In Thomas Groß and Theo Tryfonas, editors, *Socio-Technical Aspects in Security and Trust*, Lecture Notes in Computer Science, pages 3–26, Cham, 2021. Springer International Publishing.
- [190] Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang 'Anthony' Chen. Augmenting Pathologists with NaviPath: Design and Evaluation of a Human-AI Collaborative Navigation System. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [191] Egon Guba. ERIC/ECTJ Annual Review Paper: Criteria for Assessing the Trustworthiness of Naturalistic Inquiries. *Educational Communication and Technology*, 29(2):75–91, 1981.
- [192] Wentao Guo, Paige Pepitone, Adam J Aviv, and Michelle L Mazurek. How Researchers De-Identify Data in Practice. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security)*, Seattle, USA, 2025. USENIX Association.
- [193] C. Keith Haddock, David Rindskopf, and William R. Shadish. Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods*, 3(3):339–353, 1998.
- [194] Heiko Haller and Stefan Krauss. Misinterpretations of Significance: A Problem Students Share with Their Teachers? *Methods of Psychological Research Online*, 7(1), 2002.
- [195] Joseph Hallett, Nikhil Patnaik, Benjamin Shreeve, and Awais Rashid. “Do this! Do that!, and Nothing will Happen” Do Specifications Lead to Securely Stored Passwords? In *Proceedings of the IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 486–498, Madrid, Spain, 2021. IEEE.
- [196] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic HCI research data: A case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [197] Peter Hamm, David Harborth, and Sebastian Pape. A Systematic Analysis of User Evaluations in Security Research. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, Canterbury, UK, 2019. ACM.
- [198] David Hammer and Leema K. Berland. Confusing Claims for Data: A Critique of Common Practices for Presenting Qualitative Research on Learning. *Journal of the Learning Sciences*, 23(1):37–46, 2014.
- [199] Sangyoon Han, Jaejun Park, and Seungmoon Choi. Generating Haptic Motion Effects for Multiple Articulated Bodies for Improved 4D Experiences: A Camera Space Approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. Association for Computing Machinery.
- [200] Paul HP Hanel and David MA Mehler. Beyond reporting statistical significance: Identifying informative effect sizes to improve scientific communication. *Public Understanding of Science*, 28(4):468–485, 2019.

- [201] Julie M Haney, Mary F Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. "We make it a big deal in the company": Security Mindsets in Organizations that Develop Cryptographic Products. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 357–373, Baltimore, USA, 2018. USENIX Association.
- [202] Norman Hänsch, Andrea Schankin, and Mykolai Protsenko. Programming Experience Might Not Help in Comprehending Obfuscated Source Code Efficiently. In *Proceedings of the Fourteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 341–356, Baltimore, USA, 2018. USENIX Association.
- [203] Marian Harbach, Markus Hettig, Susanne Weber, and Matthew Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2647–2656, Toronto, Canada, 2014. ACM.
- [204] Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception. In *Proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS)*, pages 213–230, Menlo Park, USA, 2014. USENIX Association.
- [205] Maximilian Häring, Eva Gerlitz, Matthew Smith, and Christian Tiefenau. Less About Privacy: Revisiting a Survey about the German COVID-19 Contact Tracing App. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [206] Rakibul Hasan, Rebecca Weil, Rudolf Siegel, and Katharina Krombholz. A Psychometric Scale to Measure Individuals' Value of Other People's Privacy (VOPP). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [207] Ayako A. Hasegawa, Daisuke Inoue, and Mitsuaki Akiyama. How WEIRD is Usable Privacy and Security Research? In *Proceedings of the 33rd USENIX Security Symposium (USENIX Security)*, pages 3241–3258, Philadelphia, USA, 2024. USENIX Association.
- [208] Ayako A. Hasegawa, Naomi Yamashita, Tatsuya Mori, Daisuke Inoue, and Mitsuaki Akiyama. Understanding {Non-Experts'} Security- and Privacy-Related Questions on a {Q&A} Site. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 39–56, Boston, USA, 2022. USENIX Association.
- [209] Jordan Hayes, Smirity Kaushik, Charlotte Emily Price, and Yang Wang. Cooperative Privacy and Security: Learning from People with Visual Impairments and Their Allies. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security (SOUPS)*, Santa Clara, USA, 2019. USENIX Association.
- [210] Hussein Hazazi and Mohamed Shehab. Exploring the Usability, Security, and Privacy of Smart Locks from the Perspective of the End User. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 559–577, Anaheim, USA, 2023. USENIX Association.

- [211] Ziyao He, Yunpeng Song, Shurui Zhou, and Zhongmin Cai. Interaction of Thoughts: Towards Mediating Task Assignment in Human-AI Cooperation with a Capability-Aware Shared Mental Model. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [212] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, 2010.
- [213] Moonseong Heo, Singh R. Nair, Judith Wylie-Rosett, Myles S. Faith, Angelo Pietrobelli, Nancy R. Glassman, Sarah N. Martin, Stephanie Dickinson, and David B. Allison. Trial Characteristics and Appropriateness of Statistical Methods Applied for Design and Analysis of Randomized School-Based Studies Addressing Weight-Related Issues: A Literature Review. *Journal of Obesity*, 2018:e8767315, 2018.
- [214] Franziska Herbert, Steffen Becker, Leonie Schaewitz, Jonas Hielscher, Marvin Kowalewski, Angela Sasse, Yasemin Acar, and Markus Dürmuth. A World Full of Privacy and Security (Mis)conceptions? Findings of a Representative Survey in 12 Countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [215] Cormac Herley. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, pages 133–144, Oxford, UK, 2009. ACM.
- [216] Morten Hertzum and Niels Ebbe Jacobsen. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction*, 13(4):421–443, 2001.
- [217] Jonas Hielscher, Markus Schöps, Uta Menges, Marco Gutfleisch, Mirko Helbling, and M. Angela Sasse. Lacking the Tools and Support to Fix Friction: Results from an Interview Study with Security Managers. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 131–150, Anaheim, USA, 2023. USENIX Association.
- [218] Rink Hoekstra, Richard D. Morey, Jeffrey N. Rouder, and Eric-Jan Wagenmakers. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5):1157–1164, 2014.
- [219] John M Hoenig and Dennis M Heisey. The Abuse of Power. *The American Statistician*, 55(1):19–24, 2001.
- [220] Jake M. Hofman, Daniel G. Goldstein, and Jessica Hullman. 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*, Honolulu, USA, 2020. ACM.
- [221] Jan Karem Höhne, Joshua Claassen, Saijal Shahania, and David Broneske. Bots in web survey interviews: A showcase. *International Journal of Market Research*, 67(1):3–12, 2025.
- [222] William L Holcomb, Tinnakorn Chaiworapongsa, Douglas A Luke, and Kevin D Burgdorf. An odd measure of risk: Use and misuse of the odds ratio. *Obstetrics & Gynecology*, 98(4):685–688, 2001.

- [223] Kasper Hornbæk and Effie Lai-Chong Law. Meta-analysis of correlations among usability measures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 617–626, San Jose, USA, 2007. ACM.
- [224] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. Is once enough? on the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3523–3532, Toronto, Canada, 2014. ACM.
- [225] Stefan Albert Horstmann, Sandy Hong, David Klein, Raphael Serafini, Martin Degeling, Martin Johns, Veelasha Moonsamy, and Alena Naiakshina. “Sorry for Bugging you so much.” Exploring Developers’ Behavior Towards Privacy-Compliant Implementation. In *Proceedings of the 2025 IEEE Symposium on Security and Privacy*, pages 1215–1233, San Francisco, USA, 2024. IEEE.
- [226] Hsiao-Ying Huang, Soteris Demetriou, Muhammad Hassan, Güliz Seray Tuncay, Carl A. Gunter, and Masooda Bashir. Evaluating User Behavior in Smartphone Security: A Psychometric Perspective. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 509–524, Anaheim, USA, 2023. USENIX Association.
- [227] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang ‘Anthony’ Chen, Young-Ho Kim, and Amy Pavel. AVscript: Accessible Video Editing with Audio-Visual Scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [228] Beat Hulliger. *Modi: Multivariate Outlier Detection and Imputation for Incomplete Survey Data*, 2023.
- [229] Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R Bloem, Pralhad Burli, Naibin Chen, Paul Greico, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. The Influence of Hidden Researcher Decisions in Applied Microeconomics. *Economic Inquiry*, 59(3):944–960, 2021.
- [230] Hannah J Hutton and David A Ellis. Exploring User Motivations Behind iOS App Tracking Transparency Decisions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [231] Zainab Iftikhar, Yumeng Ma, and Jeff Huang. “Together but not together”: Evaluating Typing Indicators for Interaction-Rich Communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [232] Jane Im, Ruiyi Wang, Weikun Lyu, Nick Cook, Hana Habib, Lorrie Faith Cranor, Nikola Banovic, and Florian Schaub. Less is Not More: Improving Findability and Actionability of Privacy Controls for Online Behavioral Advertising. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [233] John P. A. Ioannidis. Why Most Published Research Findings Are False. *PLOS Medicine*, 2(8):e124, 2005.

- [234] John P. A. Ioannidis. Why Most Discovered True Associations Are Inflated. *Epidemiology*, 19(5):640–648, 2008.
- [235] John P. A. Ioannidis, Daniele Fanelli, Debbie Drake Dunne, and Steven N. Goodman. Meta-research: Evaluation and Improvement of Research Methods and Practices. *PLOS Biology*, 13(10):e1002264, 2015.
- [236] Iulia Ion, Rob Reeder, and Sunny Consolvo. “...No One Can Hack My Mind”: Comparing Expert and Non-Expert Security Practices. In *Proceedings of the Eleventh Symposium on Usable Privacy and Security (SOUPS)*, pages 327–346, Ottawa, Canada, 2015. USENIX Association.
- [237] ISO Central Secretary. Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts. Standard ISO 9241-11:2018, International Organization for Standardization, Geneva, Switzerland, 2018.
- [238] Andreas Ivarsson, Mark B. Andersen, Urban Johnson, and Magnus Lindwall. To adjust or not adjust: Nonparametric effect sizes, confidence intervals, and real-world meaning. *Psychology of Sport and Exercise*, 14(1):97–102, 2013.
- [239] Rune Møberg Jacobsen, Kasper Fangel Skov, Stine S Johansen, Mikael B. Skov, and Jesper Kjeldskov. Living with Sound Zones: A Long-Term Field Study of Dynamic Sound Zones in a Domestic Context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [240] Alfusainey Jallow and Sven Bugiel. Stack Overflow Meets Replication: Security Research Amid Evolving Code Snippets. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security)*, Seattle, USA, 2025. USENIX Association.
- [241] Bernard J. Jansen, Soon-gyo Jung, and Joni Salminen. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020, 2023.
- [242] Yvonne Jansen, Kasper Hornbæk, and Pierre Dragicevic. What Did Authors Value in the CHI’16 Reviews They Received? In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 596–608, San Jose, USA, 2016. ACM.
- [243] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):248:1–248:38, 2023.
- [244] Crescent Jicol, Christopher Clarke, Emilia Tor, Rebecca M Dakin, Tom Charlie Lancaster, Sze Tung Chang, Karin Petrini, Eamonn O’Neill, Michael J Proulx, and Christof Lutteroth. Realism and Field of View Affect Presence in VR but Not the Way You Think. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [245] Hye-Young Jo, Laurenz Seidel, Michel Pahud, Mike Sinclair, and Andrea Bianchi. FlowAR: How Different Augmented Reality Visualizations of Online Fitness Videos Support Flow for At-Home Yoga Exercises. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.

- [246] Brittany Johnson, Yoonki Song, Emerson Murphy-Hill, and Robert Bowdidge. Why don't software developers use static analysis tools to find bugs? In *Proceedings of the 35th International Conference on Software Engineering (ICSE)*, pages 672–681, San Francisco, USA, 2013. IEEE.
- [247] Rob Johnson, Anthony Watkinson, and Michael Mabe. The STM report. An overview of scientific and scholarly publishing. Technical Report 5, International Association of Scientific, Technical and Medical Publishers, 2018.
- [248] AHA / ASA Journals. AHA / ASA - Statistical Reporting Recommendations. <https://www.ahajournals.org/statistical-recommendations>, December 2023.
- [249] Alex Kale, Matthew Kay, and Jessica Hullman. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [250] Alex Kale, Matthew Kay, and Jessica Hullman. Visual Reasoning Strategies for Effect Size Judgments and Decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):272–282, 2021.
- [251] Alex Kale, Sarah Lee, Terrance Goan, Elizabeth Tipton, and Jessica Hullman. MetaExplorer : Facilitating Reasoning with Epistemic Uncertainty in Meta-analysis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [252] Annika Kaltenhauser, Gian-Luca Savino, Nick von Felten, and Johannes Schöning. CHI's Greatest Hits: Analyzing the 100 Most-Cited Papers in 43 Years of Research at ACM CHI. *Interactions*, 32(1):28–33, 2025.
- [253] Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag I.K. Sjøberg. A systematic review of effect size in software engineering experiments. *Information and Software Technology*, 49(11):1073–1086, 2007.
- [254] Fumihiro Kanei, Ayako A. Hasegawa, Eitaro Shioji, and Mitsuaki Akiyama. 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*, Hamburg, Germany, 2023. ACM.
- [255] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy Attitudes of Mechanical Turk Workers and the {U.S}. Public. In *Proceedings of the Tenth Symposium on Usable Privacy and Security (SOUPS)*, pages 37–49, Menlo Park, USA, 2014. USENIX Association.
- [256] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Proceedings of the Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 39–52, Ottawa, 2015. USENIX Association.

- [257] Maurits Kaptein and Judy Robertson. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1105–1114, Austin, USA, 2012. ACM.
- [258] Samantha Katcher, Liana Wang, Caroline Yang, Chloé Messdaghi, Michelle L. Mazurek, Marshini Chetty, Kelsey R. Fulton, and Daniel Votipka. A Survey of Cybersecurity Professionals’ Perceptions and Experiences of Safety and Belonging in the Community. In *Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS)*, Philadelphia, USA, 2024. USENIX Association.
- [259] Keiko Katsuragawa, Qi Shu, and Edward Lank. PledgeWork: Online Volunteering through Crowdwork. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [260] Harjot Kaur, Sabrina Amft, Daniel Votipka, Yasemin Acar, and Sascha Fahl. Where to Recruit for Security Development Studies: Comparing Six Software Developer Samples. In *Proceedings of the 31st USENIX Security Symposium (USENIX Security)*, Boston, USA, 2022. USENIX Association.
- [261] Smirity Kaushik, Natã M. Barbosa, Yaman Yu, Tanusree Sharma, Zachary Kilhoffer, JooYoung Seo, Sauvik Das, and Yang Wang. GuardLens: Supporting Safer Online Browsing for People with Visual Impairments. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 361–380, Anaheim, USA, 2023. USENIX Association.
- [262] Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. Special Interest Group on Transparent Statistics in HCI. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1081–1084, San Jose, USA, 2016. ACM.
- [263] Matthew Kay, Steve Haroz, Shion Guha, Pierre Dragicevic, and Chat Wacharamanotham. Moving Transparent Statistics Forward at CHI. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 534–541, Denver, USA, 2017. ACM.
- [264] Matthew Kay, Gregory L. Nelson, and Eric B. Hekler. Researcher-Centered Design of Statistics: Why Bayesian Statistics Better Fit the Culture and Incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4521–4532, San Jose, USA, 2016. ACM.
- [265] Ken Kelley and Kristopher J. Preacher. On effect size. *Psychological Methods*, 17(2):137–152, 2012.
- [266] Patrick Gage Kelley. Conducting Usable Privacy & Security Studies with Amazon’s Mechanical Turk. In *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS)*, Redmond, USA, 2010. ACM.
- [267] Patrick Gage Kelley, Celestina Cornejo, Lisa Hayes, Ellie Shuo Jin, Aaron Sedley, Kurt Thomas, Yongwei Yang, and Allison Woodruff. “There will be less privacy, of course”: How and why people in 10 countries expect AI will affect privacy in the future. In *Proceedings of*

- the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, Anaheim, USA, 2023. USENIX Association.
- [268] Riko Kelter. 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, 2020.
- [269] Maurice George Kendall. *Rank Correlation Methods*. Griffin, Oxford, England, 4th edition, 1970.
- [270] Leon Kersten, Tom Mulders, Emmanuele Zambon, Chris Snijders, and Luca Allodi. 'Give Me Structure': Synthesis and Evaluation of a (Network) Threat Analysis Process Supporting Tier 1 Investigations in a Security Operation Center. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 97–111, Anaheim, USA, 2023. USENIX Association.
- [271] Martin Kessner, Jo Wood, Richard F. Dillon, and Robert L. West. On the reliability of usability testing. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, pages 97–98, Seattle, USA, 2001. ACM.
- [272] Yea-Seul Kim, Jake M Hofman, and Daniel G Goldstein. 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*, New Orleans, USA, 2022. ACM.
- [273] Nigel King, Linda Finlay, Peter Ashworth, Jonathan A. Smith, Darren Langdridge, and Trevor Butt. "Can't Really Trust That, So What Can I Trust?": A Polyvocal, Qualitative Analysis of the Psychology of Mistrust. *Qualitative Research in Psychology*, 5(2):80–102, 2008.
- [274] Roger E. Kirk. Practical Significance: A Concept Whose Time Has Come. *Educational and Psychological Measurement*, 56(5):746–759, 1996.
- [275] Roger E. Kirk. The Importance of Effect Magnitude. In *Handbook of Research Methods in Experimental Psychology*, pages 83–105. Blackwell Publishing, 2003.
- [276] Jan H. Klemmer, Julianne Schmöser, Fabian Fischer, Jacques Suray, Jan-Ulrich Holtgrave, Simon Lenau, Byron M. Lowens, Florian Schaub, and Sascha Fahl. How Transparent is Usable Privacy and Security Research? A Meta-Study on Current Research Transparency Practices. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security)*, Seattle, USA, 2025. USENIX Association.
- [277] Jan H Klemmer, Julianne Schmöser, Byron M Lowens, Fabian Fischer, Lea Schmöser, Florian Schaub, and Sascha Fahl. Transparency in Usable Privacy and Security Research: Scholars' Perspectives, Practices, and Recommendations. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 2658–2677, San Francisco, USA, 2025. IEEE.
- [278] Sabrina Klivan, Sandra Höltervennhoff, Nicolas Huaman, Yasemin Acar, and Sascha Fahl. "Would You Give the Same Priority to the Bank and a Game? I Do Not!" Exploring Credential Management Strategies and Obstacles during Password Manager Setup. In

- Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 171–190, Anaheim, USA, 2023. USENIX Association.
- [279] Brandy Klug. An Overview of the System Usability Scale in Library Website and System Usability Testing. *Weave: Journal of Library User Experience*, 1(6), 2017.
- [280] Max Korbmacher, Flavio Azevedo, Charlotte R. Pennington, Helena Hartmann, Madeleine Pownall, Kathleen Schmidt, Mahmoud Elsherif, Nate Breznau, Olly Robertson, Tamara Kalandadze, Shijun Yu, Bradley J. Baker, Aoife O’Mahony, Jørgen Ø. S. Olsnes, John J. Shaw, Biljana Gjoneska, Yuki Yamada, Jan P. Röer, Jennifer Murphy, Shilaan Alzahawi, Sandra Grinschgl, Catia M. Oliveira, Tobias Wingen, Siu Kit Yeung, Meng Liu, Laura M. König, Nihan Albayrak-Aydemir, Oscar Lecuona, Leticia Micheli, and Thomas Evans. The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology*, 1(3), 2023.
- [281] Yubo Kou, Jian Zhao, and Sadeghian Shadan. Alt.chi. <https://chi2025.acm.org/for-authors/alt-chi/>, 2024.
- [282] Monica Kowalczyk, Johanna T. Gunawan, David Choffnes, Daniel J Dubois, Woodrow Hartzog, and Christo Wilson. Understanding Dark Patterns in Home IoT Devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [283] Jeanine Krath, Maximilian Altmeyer, Gustavo F. Tondello, and Lennart E. Nacke. Hexad-12: Developing and Validating a Short Version of the Gamification User Types Hexad Scale. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [284] Laura Krefting. Rigor in Qualitative Research: The Assessment of Trustworthiness. *The American Journal of Occupational Therapy*, 45(3), 1991.
- [285] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. "If HTTPS Were Secure, I Wouldn’t Need 2FA" - End User and Administrator Mental Models of HTTPS. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 246–263, San Francisco, USA, 2019. IEEE.
- [286] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. “I Have No Idea What I’m Doing” – On the Usability of Deploying HTTPS. In *Proceedings of the 26th USENIX Security Symposium (USENIX Security)*, pages 1339–1356, Vancouver, Canada, 2017. USENIX Association.
- [287] Anton Kühberger, Astrid Fritz, and Thomas Scherndl. Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE*, 9(9):e105825, 2014.
- [288] Patrick Kühtreiber, Viktoriya Pak, and Delphine Reinhardt. Replication: The Effect of Differential Privacy Communication on German Users’ Comprehension and Data Sharing Attitudes. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 117–134, Boston, USA, 2022. USENIX Association.

- [289] Radha Kumaran, You-Jin Kim, Anne E Milner, Tom Bullock, Barry Giesbrecht, and Tobias Höllerer. The Impact of Navigation Aids on Search Performance and Object Recall in Wide-Area Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [290] Lin Kyi, Sushil Ammanaghatta Shivakumar, Cristiana Teixeira Santos, Franziska Roesner, Frederike Zufall, and Asia J. Biega. Investigating Deceptive Design in GDPR’s Legitimate Interest. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [291] Daniel Lakens. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 2013.
- [292] Daniël Lakens. Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7):701–710, 2014.
- [293] Daniël Lakens. Sample Size Justification. *Collabra: Psychology*, 8(1), 2022.
- [294] Daniël Lakens and Aaron R. Caldwell. Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920951503, 2021.
- [295] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.
- [296] Sean P. Lane and Erin P. Hennes. Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, 35(1):7–31, 2018.
- [297] Florian Lang, Verena Pues, Albrecht Schmidt, and Tonja-Katrin Machulla. BrailleBuddy: A Tangible User Interface to Support Children with Visual Impairment in Learning Braille. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [298] LangGraph. LangGraph.
<https://web.archive.org/web/20250822122351/https://www.langchain.com/langgraph>.
- [299] Victor Le Pochat and Wouter Joosen. Analyzing Cyber Security Research Practices through a Meta-Research Framework. In *Proceedings of the 16th Cyber Security Experimentation and Test Workshop*, pages 64–74, Marina del Rey, USA, 2023. ACM.
- [300] Matthew L Lee, Scott Carter, Rumen Iliev, Nayeli Suseth Bravo, Monica P Van, Laurent Denoue, Everlyne Kimani, Alexandre L. S. Filipowicz, David A. Shamma, Kate A Sieck, Candice Hogan, and Charlene C. Wu. Understanding People’s Perception and Usage of Plug-in Electric Hybrids. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [301] Lawrence Leung. Validity, reliability, and generalizability in qualitative research. *Journal of family medicine and primary care*, 4(3):324–327, 2015.

- [302] Brittany Lewis, Tina-Marie Ranalli, Alexandra Gourley, Piriyanakan Kirupaharan, and Krishna Venkatasubramanian. "I... caught a person casing my house... and scared him off:" The Use of Security-Focused Smart Home Devices by People with Disabilities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2023. ACM.
- [303] Brenna Li, Tetyana Skoropad, Puneet Seth, Mohit Jain, Khai Truong, and Alex Mariakakis. Constraints and Workarounds to Support Clinical Consultations in Synchronous Text-based Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [304] Zhi Li, Yu-Jung Ko, Aini Putkonen, Shirin Feiz, Vikas Ashok, Iv Ramakrishnan, Antti Oulasvirta, and Xiaojun Bi. Modeling Touch-based Menu Selection Performance of Blind Users via Reinforcement Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [305] David Lie and Véronique Cortier. ACM CCS 2025. Call for Papers. <https://www.sig-sac.org/ccs/CCS2025/call-for-papers/>, 2024.
- [306] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. How WEIRD is CHI? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. ACM.
- [307] Mark W. Lipsey and David B. Wilson. The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation From Meta-Analysis. *American Psychologist*, 48(12):1181–1209, 1993.
- [308] Mark W. Lipsey and David B. Wilson. *Practical Meta-Analysis*. Sage Publications, Inc, Thousand Oaks, CA, US, 2001.
- [309] Maxim Lisnic, Cole Polychronis, Alexander Lex, and Marina Kogan. Misleading Beyond Visual Tricks: How People Actually Lie with Charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [310] Chengzhong Liu, Shixu Zhou, Dingdong Liu, Junze Li, Zeyu Huang, and Xiaojuan Ma. CoArgue : Fostering Lurkers' Contribution to Collective Arguments in Community-based QA Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [311] Enze Liu, Lu Sun, Alex Bellon, Grant Ho, Geoffrey M. Voelker, Stefan Savage, and Imani N. S. Munyaka. Understanding the Viability of Gmail's Origin Indicator for Identifying the Sender. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 77–95, Anaheim, USA, 2023. USENIX Association.
- [312] Jiazhou Liu, Barrett Ens, Arnaud Prouzeau, Jim Smiley, Isobel Kara Nixon, Sarah Goodwin, and Tim Dwyer. DataDancing: An Exploration of the Design Space For Visualisation View Management for 3D Surfaces and Spaces. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. Association for Computing Machinery.

-
- [313] Shengmei Liu, Atsuo Kuwahara, James J Scovell, and Mark Claypool. The Effects of Frame Rate Variation on Game Player Quality of Experience. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [314] Xingyu "Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang "Anthony" Chen, and Ruofei Du. Visual Captions: Augmenting Verbal Communication with On-the-fly Visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [315] Yang Liu, Tim Althoff, and Jeffrey Heer. Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [316] Alexandra Mai, Katharina Pfeffer, Matthias Gusenbauer, Edgar Weippl, and Katharina Krombholz. User Mental Models of Cryptocurrency Systems - A Grounded Theory Approach. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 341–358, 2020.
- [317] Dominique Makowski, Mattan S. Ben-Shachar, S. H. Annabel Chen, and Daniel Lüdecke. Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, 10:2767, 2019.
- [318] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research*, 15(4):336–355, 2004.
- [319] Megh Marathe and Kentaro Toyama. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, Canada, 2018. ACM.
- [320] Ioana Andreea Marin, Pavlo Burda, Nicola Zannone, and Luca Allodi. The Influence of Human Factors on the Intention to Report Phishing Emails. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. Association for Computing Machinery.
- [321] Florin Martius, Luisa Jansen, Lukas Struck, Arthi Arumugam, Lisa Geierhaas, Anna-Marie Ortloff, Matthew Smith, and Christian Tiefenau. Out of sight, out of mind? Exploring data protection practices for personal data in usable security & privacy studies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2025. ACM.
- [322] Arunesh Mathur and Marshini Chetty. Impact of User Characteristics on Attitudes Towards Automatic Mobile Application Updates. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 175–193. USENIX Association, 2017.
- [323] MaxQDA. MAXQDA Manual.
<https://web.archive.org/web/20250822123004/https://www.maxqda.com/de/hilfe-mx22/willkommen>.

- [324] Scott E. Maxwell, Ken Kelley, and Joseph R. Rausch. Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology*, 59(1):537–563, 2008.
- [325] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. Measuring password guessability for an entire university. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pages 173–186, Berlin, Germany, 2013. ACM.
- [326] McKenna McCall, Eric Zeng, Faysal Hossain Shezan, Mitchell Yang, Lujo Bauer, Abhishek Bichhawat, Camille Cobb, Limin Jia, and Yuan Tian. Towards Usable Security Analysis Tools for Trigger-Action Programming. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 301–320, Anaheim, USA, 2023. USENIX Association.
- [327] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.
- [328] Kenneth O. McGraw and S. P. Wong. A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365, 1992.
- [329] Tamir Mendel and Eran Toch. Social Support for Mobile Security: Comparing Close Connections and Community Volunteers in a Field Experiment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [330] Abraham H Mhaidli, Yixin Zou, and Florian Schaub. "We Can't Live Without Them!" App Developers' Adoption of Ad Networks and Their Considerations of Consumer Risks. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 225–244, Santa Clara, USA, 2019. USENIX Association.
- [331] David Moher, Kenneth F. Schulz, and Douglas G. Altman. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, 357(9263):1191–1194, 2001.
- [332] Maria D. Molina, Emily S Zhan, Devanshi Agnihotri, Saeed Abdullah, and Pallav Deka. Motivation to Use Fitness Application for Improving Physical Activity Among Hispanic Users: The Pivotal Role of Interactivity and Relatedness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [333] Raydel Montesino and Stefan Fenz. Information Security Automation: How Far Can We Go? In *Proceedings of the Sixth International Conference on Availability, Reliability and Security*, pages 280–285, Vienna, Austria, 2011. IEEE.
- [334] Victor M. Montori, P. J. Devereaux, Neill K. J. Adhikari, Karen E. A. Burns, Christoph H. Eggert, Matthias Briel, Christina Lacchetti, Teresa W. Leung, Elizabeth Darling, Dianne M. Bryant, Heiner C. Bucher, Holger J. Schünemann, Maureen O. Meade, Deborah J. Cook, Patricia J. Erwin, Amit Sood, Richa Sood, Benjamin Lo, Carly A.

- Thompson, Qi Zhou, Edward Mills, and Gordon H. Guyatt. Randomized Trials Stopped Early for Benefit: A Systematic Review. *JAMA : the journal of the American Medical Association*, 294(17):2203–2209, 2005.
- [335] Benjamin Morrison, Lynne Coventry, and Pam Briggs. How do Older Adults feel about engaging with Cyber-Security? *Human Behavior and Emerging Technologies*, 3(5):1033–1049, 2021.
- [336] Florian Müller, Arantxa Ye, Dominik Schön, and Julian Rasch. UndoPort: Exploring the Influence of Undo-Actions for Locomotion in Virtual Reality on the Efficiency, Spatial Understanding and User Experience. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [337] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [338] Collins W. Munyendo, Yasemin Acar, and Adam J. Aviv. "In Eighty Percent of the Cases, I Select the Password for Them": Security and Privacy Challenges, Advice, and Opportunities at Cybercafes in Kenya. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 570–587, San Francisco, USA, 2023. IEEE.
- [339] Paul A. Murtaugh. In defense of P values. *Ecology*, 95(3):611–617, 2014.
- [340] Sarah Nadi, Stefan Krüger, Mira Mezini, and Eric Bodden. Jumping through hoops: Why do Java developers struggle with cryptography APIs? In *Proceedings of the 38th International Conference on Software Engineering*, pages 935–946, Austin, USA, 2016. ACM.
- [341] Ruth G Nagassa, Matthew Butler, Leona Holloway, Cagatay Goncu, and Kim Marriott. 3D Building Plans: Supporting Navigation by People who are Blind or have Low Vision in Multi-Storey Buildings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [342] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. On Conducting Security Developer Studies with CS Students: Examining a Password-Storage Study with CS Students, Freelancers, and Company Developers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [343] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zeischwitz, and Matthew Smith. "If you want, I can store the encrypted password": A Password-Storage Field Study with Freelance Developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [344] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why Do Developers Get Password Storage Wrong?: A Qualitative Usability Study. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 311–328, Dallas Texas USA, 2017. ACM.

- [345] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. Deception Task Design in Developer Password Studies: Exploring a Student Sample. In *Proceedings of the Fourteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 297–313, Baltimore, USA, 2018. USENIX Association.
- [346] Daniela Napoli, Khadija Baig, Sana Maqsood, and Sonia Chiasson. "I'm Literally Just Hoping This Will Work": Obstacles Blocking the Online Security and Privacy of Users with Visual Disabilities. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security*, pages 263–280. USENIX Association, 2021.
- [347] Lorenzo Neil, Harshini Sri Ramulu, Yasemin Acar, and Bradley Reaves. Who Comes Up with this Stuff? Interviewing Authors to Understand How They Produce Security Advice. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 283–299, Anaheim, USA, 2023. USENIX Association.
- [348] EQUATOR Network. Reporting checklists for medical researchers.
<https://www.goodreports.org/>.
- [349] EQUATOR Network. Reporting guidelines.
<https://www.equator-network.org/reporting-guidelines/>.
- [350] Binh Vinh Duc Nguyen, Jihae Han, Maarten Houben, Yssmin Bayoumi, and Andrew Vande Moere. Engaging Passers-by with Rhythm: Applying Feedforward Learning to a Xylophonic Media Architecture Facade. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [351] Duc Cuong Nguyen, Dominik Wermke, Yasemin Acar, Michael Backes, Charles Weir, and Sascha Fahl. A Stitch in Time: Supporting Android Developers in Writing Secure Code. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1065–1077, Dallas Texas USA, 2017. ACM.
- [352] Jakob Nielsen and Jonathan Levy. Measuring usability: Preference vs. performance. *Communications of the ACM*, 37(4):66–75, 1994.
- [353] Tim Nosco, Jared Ziegler, and Zechariah Clark. The Industrial Age of Hacking. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security)*, pages 1129–1146. USENIX Association, 2020.
- [354] Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226, 2016.
- [355] Natalia Obukhova. A Meta-Analysis of Effect Sizes of CHI Typing Experiments. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. ACM.
- [356] Cliodhna O'Connor and Helene Joffe. Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods*, 19, 2020.

- [357] Sangeun Oh, Hyuck Yoo, Dae R. Jeong, Duc Hoang Bui, and Insik Shin. Mobile Plus: Multi-device Mobile Platform for Cross-device Functionality Sharing. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 332–344, Niagra Falls, USA, 2017. ACM.
- [358] Daniel J O’Keefe. Misunderstandings of effect sizes in message effects research. *Communication Methods and Measures*, 11(3):210–219, 2017.
- [359] Stephen Olejnik and James Algina. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychological Methods*, 8(4):434–447, 2003.
- [360] Daniela Seabra Oliveira, Tian Lin, Muhammad Sajidur Rahman, Rad Akefirad, Donovan Ellis, Eliany Perez, and Rahul Bobhate. API Blindspots: Why Experienced Developers Write Vulnerable Code. In *Proceedings of the Fourteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 315–328, Baltimore, MD, 2018. USENIX Association.
- [361] Marten Oltrogge, Yasemin Acar, Sergej Dechand, Matthew Smith, and Sascha Fahl. To Pin or Not to Pin Helping App Developers Bullet Proof Their TLS Connections. In *Proceedings of the 24th USENIX Security Symposium (USENIX Security)*, pages 239–254, Washington, D.C., USA, 2015. USENIX Association.
- [362] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.
- [363] OpenAI. GPT-4 Technical Report. <https://cdn.openai.com/papers/gpt-4.pdf>, 2023.
- [364] Jonas Oppenlaender. Past, Present, and Future of Citation Practices in HCI. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2025. ACM.
- [365] Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of an Interview and Survey Study on Security Advice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [366] Anna-Marie Ortloff, Julia Angelika Grohs, Simon Lenau, and Matthew Smith. 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 *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2025. ACM.
- [367] Anna-Marie Ortloff, Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas, and Matthew Smith. Small, medium, large? A meta-study of effect sizes at CHI to aid interpretation of effect sizes and power calculation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2025. ACM.
- [368] Anna-Marie Ortloff, Jenny Tang, Arthi Arumugam, Daniel Huschina, Lisa Geierhaas, Florin Martius, Luisa Jansen, Kolja von der Twer, Lilly Jungbluth, and Matthew Smith.

- Replication: "No One Can Hack My Mind" - 10 Years Later: An Update and Outlook on Experts' and Non-Experts' Security Practices and Advice. In *Proceedings of the Twentyfirst Symposium on Usable Privacy and Security (SOUPS)*, pages 435–454, Seattle, USA, 2025. USENIX Association.
- [369] Anna-Marie Ortloff, Christian Tiefenau, and Matthew Smith. 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 *Proceedings of the Nineteenth Symposium on Usable Privacy and Security*, pages 341–359, Anaheim, USA, 2023. USENIX Association.
- [370] Antti Oulasvirta and Kasper Hornbæk. HCI Research as Problem-Solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4956–4967, San Jose, USA, 2016. ACM.
- [371] Our World in Data. Data Page: Annual articles published in scientific and technical journals. <https://archive.ourworldindata.org/20250704-091220/grapher/scientific-and-technical-journal-articles.html>, 2025.
- [372] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, and Joseph Chee Chang. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [373] Hernan Palombo, Armin Ziaie Tabari, Daniel Lende, Jay Ligatti, and Xinming Ou. An Ethnographic Understanding of Software (In)Security and a Co-Creation Model to Improve Secure Software Development. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS)*. USENIX Association, 2020.
- [374] Shanta Pandey and Charlotte Lyn Bright. What Are Degrees of Freedom? *Social Work Research*, 32(2):119–128, 2008.
- [375] Aswati Panicker, Novia Nurain, Zaidat Ibrahim, Chun-Han (Ariel) Wang, Seung Wan Ha, Yuxing Wu, Kay Connelly, Katie A. Siek, and Chia-Fang Chung. Understanding fraudulence in online qualitative studies: From the researcher's perspective. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2024. ACM.
- [376] Rebecca Panskus, Max Ninow, Sascha Fahl, and Karola Marky. Privacy Mental Models of Electronic Health Records: A German Case Study. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 525–542, Anaheim, USA, 2023. USENIX Association.
- [377] Jaewoo Park, Eunji Lee, Yoonsik Kim, Isaac Kang, Hyung Il Koo, and Nam Ik Cho. Multi-Lingual Optical Character Recognition System Using the Reinforcement Learning of Character Segmenter. *IEEE Access*, 8:174437–174448, 2020.
- [378] Sunyup Park, Anna Lenhart, Michael Zimmer, and Jessica Vitak. "Nobody's Happy": Design Insights from Privacy-Conscious Smart Home Power Users on Enhancing Data Transparency, Visibility, and Control. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, Anaheim, USA, 2023. USENIX Association.

- [379] Simon Parkin, Kat Krol, Ingolf Becker, and M. Angela Sasse. Applying Cognitive Control Modes to Identify Security Fatigue Hotspots. In *Proceedings of the Twelfth Symposium on Usable Privacy and Security (SOUPS)*, Security Fatigue Workshop'16. USENIX Association, 2016.
- [380] Viral Parmar, Harshal A. Sanghvi, Riki H Patel, and Abhijit S. Pandya. A Comprehensive Study on Passwordless Authentication. In *Proceedings of the International Conference on Sustainable Computing and Data Communication Systems*, pages 1266–1275, Erode, India, 2022. IEEE.
- [381] Ivan Pashchenko, Duc-Ly Vu, and Fabio Massacci. A Qualitative Study of Dependency Management and Its Security Implications. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 1513–1531. ACM, 2020.
- [382] Irene V. Pasquetto, Ashley E. Sands, Peter T. Darch, and Christine L. Borgman. Open Data in Scientific Settings: From Policy to Practice. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1585–1596, San Jose, USA, 2016. ACM.
- [383] Jessica Pater, Amanda Coupe, Rachel Pfafman, Chanda Phelan, Tammy Toscos, and Maia Jacobs. Standardizing Reporting of Participant Compensation in HCI: A Systematic Literature Review and Recommendations for the Field. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. ACM.
- [384] Nikhil Patnaik, Joseph Hallett, and Awais Rashid. Usability Smells: An Analysis of Developers' Struggle With Crypto Libraries. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 245–257, Santa Clara, USA, 2019. USENIX Association.
- [385] Alexander Persoskie and Rebecca A. Ferrer. A Most Odd Ratio:: Interpreting and Describing Odds Ratios. *American Journal of Preventive Medicine*, 52(2):224–228, 2017.
- [386] Marco Perugini, Marcello Gallucci, and Giulio Costantini. A Practical Primer To Power Analysis for Simple Experimental Designs. *Internation Review of Social Psychology*, 31(1), 2018.
- [387] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [388] Anne E Pezalla, Jonathan Pettigrew, and Michelle Miller-Day. Researching the researcher-as-instrument: An exercise in interviewer self-reflexivity. *Qualitative Research*, 12(2):165–185, 2012.
- [389] Katharina Pfeffer, Alexandra Mai, Adrian Dabrowski, Matthias Gusenbauer, Philipp Schindler, Edgar Weippl, Michael Franz, and Katharina Krombholz. On the Usability of Authenticity Checks for Hardware Security Tokens. In *Proceedings of the 30th USENIX Security Symposium (USENIX Security)*, pages 37–54. USENIX Association, 2021.

- [390] Katharina Pfeffer, Alexandra Mai, Edgar Weippl, Emilee Rader, and Katharina Krombholz. Replication: Stories as informal lessons about security. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS)*, Boston, USA, 2022. USENIX Association.
- [391] Chanda Phelan, Jessica Hullman, Matthew Kay, and Paul Resnick. Some prior(s) experience necessary: Templates for getting started with bayesian analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [392] Rachel Phinnemore, Mohi Reza, Blaine Lewis, Karthik Mahadevan, Bryan Wang, Michelle Annett, and Daniel Wigdor. Creepy Assistant: Development and Validation of a Scale to Measure the Perceived Creepiness of Voice Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [393] Andreas Plesner, Tobias Vontobel, and Roger Wattenhofer. Breaking reCAPTCHA v2. In *Proceedings of the IEEE 48th Annual Computers, Software, and Applications Conference*, pages 1047–1056, Osaka, Japan, 2024. IEEE.
- [394] Amy C Plint, David Moher, Andra Morrison, Kenneth Schulz, Douglas G Altman, Catherine Hill, and Isabelle Gaboury. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical Journal of Australia*, 185(5):263–267, 2006.
- [395] Stephan Plöger, Mischa Meier, and Matthew Smith. A Qualitative Usability Evaluation of the Clang Static Analyzer and {libFuzzer} with {CS} Students and {CTF} Players. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, pages 553–572. USENIX Association, 2021.
- [396] Stephan Plöger, Mischa Meier, and Matthew Smith. A Usability Evaluation of AFL and libFuzzer with CS Students. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [397] Luke Plonsky and Frederick L Oswald. How big is “big”? Interpreting effect sizes in L2 research. *Language learning*, 64(4):878–912, 2014.
- [398] Henning Pohl and Aske Mottelson. How we Guide, Write, and Cite at CHI. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [399] Leo Poom and Anders af Wåhlberg. Accuracy of conversion formula for effect sizes: A Monte Carlo simulation. *Research Synthesis Methods*, 13(4):508–519, 2022.
- [400] Sarah Prange, Pascal Knierim, Gabriel Knoll, Felix Dietz, Alexander De Luca, and Florian Alt. “I do (not) need that Feature!” – Understanding Users’ Awareness and Control of Privacy Permissions on Android Smartphones. In *Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS)*, pages 453–472, Philadelphia, USA, 2024. USENIX Association.

-
- [401] Lutz Prechelt. Plat _Forms: A Web Development Platform Comparison by an Exploratory Experiment Searching for Emergent Platform Properties. *IEEE Transactions on Software Engineering*, 37(1):95–108, 2011.
- [402] Viva Sarah Press and Hadas Erel. Humorous Robotic Behavior as a New Approach to Mitigating Social Awkwardness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [403] Prolific. Approvals, rejections & returns.
<https://web.archive.org/web/20250710075422/https://researcher-help.prolific.com/en/article/3405a6>, 2025.
- [404] Mirjana Prpa, Sarah Fdili-Alaoui, Thecla Schiphorst, and Philippe Pasquier. Articulating Experience: Reflections from Experts Applying Micro-Phenomenology to Design Research in HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [405] Xiaoying Pu, Licheng Zhu, Matthew Kay, and Frederick Conrad. Designing for Preregistration: A User-Centered Perspective. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland, UK, 2019. ACM.
- [406] Yue Qin, Chun Yu, Wentao Yao, Jiachen Yao, Chen Liang, Yueting Weng, Yukang Yan, and Yuanchun Shi. Selecting Real-World Objects via User-Perspective Phone Occlusion. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [407] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2022.
- [408] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2024.
- [409] Emilee Rader. Data Privacy and Pluralistic Ignorance. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 457–471, Anaheim, USA, 2023. USENIX Association.
- [410] Deepthi Raghunandan, Aayushi Roy, Shenzhi Shi, Niklas Elmqvist, and Leilani Battle. Code Code Evolution: Understanding How People Change Data Science Notebooks Over Time. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [411] Shwetha Rajaram, Chen Chen, Franziska Roesner, and Michael Nebeling. Eliciting Security & Privacy-Informed Sharing Techniques for Multi-User Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [412] Gulnar Rakhmetulla and Ahmed Sabbir Arif. Crownboard: A One-Finger Crown-Based Smartwatch Keyboard for Users with Limited Dexterity. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.

- [413] Jack Ratcliffe, Francesco Soave, Nick Bryan-Kinns, Laurissa Tokarchuk, and Ildar Farkhatdinov. Extended reality (XR) remote research: A survey of drawbacks and opportunities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. ACM.
- [414] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How I Learned to be Secure: A Census-Representative Survey of Security Advice Sources and Behavior. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 666–677, Vienna, Austria, 2016. ACM.
- [415] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 1326–1343, San Francisco, USA, 2019. IEEE.
- [416] Elissa M. Redmiles, Amelia R. Malone, and Michelle L. Mazurek. I Think They’re Trying to Tell Me Something: Advice Sources and Selection for Digital Security. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 272–288, San Jose, CA, 2016. IEEE.
- [417] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security)*, pages 89–108. USENIX Association, 2020.
- [418] Elissa M. Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L. Mazurek. Asking for a Friend: Evaluating Response Biases in Security User Studies. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1238–1255, Toronto, Canada, 2018. ACM.
- [419] Rob Reeder. If you could tell a user three things to do to stay safe online, what would they be?, March 2014.
- [420] Robert W. Reeder, Iulia Ion, and Sunny Consolvo. 152 Simple Steps to Stay Safe Online: Security Advice for Non-Tech-Savvy Users. *IEEE Security & Privacy*, 15(5):55–64, 2017.
- [421] Stuart Reeves. Human-computer interaction as science. In *Proceedings of the Fifth Decennial Aarhus Conference on Critical Alternatives*, pages 73–84, Aarhus N, 2015. Aarhus University Press.
- [422] Jun Rekimoto. WESPER: Zero-shot and Realtime Whisper to Normal Voice Conversion for Whisper-based Speech Interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [423] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. ACM.
- [424] John T. E. Richardson. Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, 6(2):135–147, 2011.

-
- [425] Zak Risha, Deniz Sonmez Unal, and Erin Walker. Soliloquy: Fostering Poetry Comprehension Using an Interactive Think-aloud Visualization. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [426] Raquel Breejon Robinson, Alberto Alvarez, and Elisa D. Mekler. How to write a CHI paper (asking for a friend). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2024. ACM.
- [427] Filipa Rocha, Filipa Correia, Isabel Neto, Ana Cristina Pires, João Guerreiro, Tiago Guerreiro, and Hugo Nicolau. Coding Together: On Co-located and Remote Collaboration between Children with Mixed-Visual Abilities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [428] Kat Roemmich, Florian Schaub, and Nazanin Andalibi. Emotion AI at Work: Implications for Workplace Surveillance, Emotional Labor, and Emotional Privacy. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [429] Katja Rogers, Teresa Hirzle, Sukran Karaosmanoglu, Paula Toledo Palomino, Ekaterina Durmanova, Seiji Isotani, and Lennart E. Nacke. An Umbrella Review of Reporting Quality in CHI Systematic Reviews: Guiding Questions and Best Practices for HCI. *ACM Transactions on Computer-Human Interaction*, 31(5):1–55, 2024.
- [430] Katja Rogers and Katie Seaborn. 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*, Hamburg, Germany, 2023. ACM.
- [431] Wade Rose and Inder Jit Singh Mann. The Variability of Pseudo R²s in Logistic Regression Models. *The IUP Journal of Computational Mathematics*, IV(1):7–18, 2011.
- [432] Robert Rosenthal. Parametric Measures of Effect Size. In Harris Cooper and Larry Hedges, editors, *The Handbook of Research Synthesis*, pages 231–244. Russel Sage Foundation, New York, 1994.
- [433] Robert Rosenthal and Donald B Rubin. A Simple, General Purpose Display of Magnitude of Experimental Effect. *Journal of Educational Psychology*, 74(2):166–169, 1982.
- [434] Ralph L Rosnow and Robert Rosenthal. Statistical Procedures and the Justification of Knowledge in Psychological Science. *American Psychologist*, 44(10):1276–1284, 1989.
- [435] Joseph Rossi. Statistical Power of Psychological Research: What Have We Gained in 20 Years? *Journal of Consulting and Clinical Psychology*, 58(5):646–656, 1990.
- [436] Sebastian Roth, Lea Gröber, Michael Backes, Katharina Krombholz, and Ben Stock. 12 Angry Developers - A Qualitative Study on Developers’ Struggles with CSP. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3085–3103, Seoul, South Korea, 2021. ACM.

- [437] Anna Lena Rotthaler, Harshini Sri Ramulu, Lucy Simko, Sascha Fahl, and Yasemin Acar. “It’s time. Time for digital security.”: An End User Study on Actionable Security and Privacy Advice. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2024.
- [438] Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L. Mazurek, and Piotr Mardziel. Build It, Break It, Fix It: Contesting Secure Development. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 690–703, Vienna Austria, 2016. ACM.
- [439] Melvin T. Rupinski and William P. Dunlap. Approximating Pearson Product-Moment Correlations from Kendall’s Tau and Spearman’s Rho. *Educational and Psychological Measurement*, 56(3):419–429, 1996.
- [440] David Saffo, Caglar Yildirim, Sara Di Bartolomeo, and Cody Dunne. Crowdsourcing Virtual Reality Experiments using VRChat. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [441] Anca Salagean, Eleanor Crellin, Martin Parsons, Darren Cosker, and Danaë Stanton Fraser. Meeting Your Virtual Twin: Effects of Photorealism and Personalization on Embodiment, Self-Identification and Perception of Self-Avatars in Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [442] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanotham, and Mauro Cherubini. 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*, Hamburg, Germany, 2023. ACM.
- [443] Carrie B. Sanders and Carl J. Cuneo. Social Reliability in Qualitative Team Research. *Sociology*, 44(2):325–343, 2010.
- [444] Shruti Sannon and Andrea Forte. 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):455:1–455:33, 2022.
- [445] Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, and Salil Vadhan. Don’t Look at the Data! How Differential Privacy Reconfigures the Practices of Data Science. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [446] Eve Schade, Gian-Luca Savino, Jasmin Niess, and Johannes Schöning. MapUncover: Fostering Spatial Exploration through Gamification in Mobile Map Apps. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [447] Thomas Schäfer. On the Use and Misuse of Standardized Effect Sizes in Psychological Research, 2023.

- [448] Thomas Schäfer and Marcus A. Schwarz. The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, 10, 2019.
- [449] Stuart Schechter. Common Pitfalls in Writing about Security and Privacy Human Subjects Experiments, and How to Avoid Them, 2010.
- [450] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The Emperor’s New Security Indicators. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 51–65, Berkeley, USA, 2007. IEEE.
- [451] Anne M. Scheel, Mitchell R. M. J. Schijen, and Daniël Lakens. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, 4(2), 2021.
- [452] Roberta W Scherer, Joerg J Meerpohl, Nadine Pfeifer, Christine Schmucker, Guido Schwarzer, and Erik von Elm. Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews*, 2018(11), 2018.
- [453] Ruben Schlagowski, Dariia Nazarenko, Yekta Can, Kunal Gupta, Silvan Mertes, Mark Billingham, and Elisabeth André. Wish You Were Here: Mental and Physiological Effects of Remote Music Collaboration in Mixed Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [454] Xenia Schmalz, José Biurrun Manresa, and Lei Zhang. What is a Bayes factor? *Psychological Methods*, 28(3):705–718, 2023.
- [455] I. Manjula Schou and Ian C. Marschner. Meta-analysis of clinical trials with early stopping: An investigation of potential bias. *Statistics in Medicine*, 32(28):4859–4874, 2013.
- [456] Kenneth F. Schulz, Douglas G. Altman, and David Moher. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and Pharmacotherapeutics*, 1(2):100–107, 2010.
- [457] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.
- [458] Peter Sedlmeier and Gerd Gigerenzer. Do studies of statistical power have an effect on the power of studies? In *Methodological Issues & Strategies in Clinical Research*, pages 389–406. American Psychological Association, Washington, DC, US, 1992.
- [459] Raphael Serafini, Marco Gutfleisch, Stefan Albert Horstmann, and Alena Naiakshina. On the Recruitment of Company Developers for Security Studies: Results from a Qualitative Interview Study. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security*, pages 321–340, Anaheim, USA, 2023. USENIX Association.
- [460] Raphael Serafini, Stefan Albert Horstmann, and Alena Naiakshina. Engaging Company Developers in Security Research Studies: A Comprehensive Literature Review and Quantitative Survey. In *Proceedings of the 33rd USENIX Security Symposium (USENIX Security)*, pages 3277–3294, Philadelphia, USA, 2024. USENIX Association.

- [461] Raphael Serafini, Clemens Otto, Stefan Albert Horstmann, and Alena Naiakshina. ChatGPT-Resistant Screening Instrument for Identifying Non-Programmers. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, Lisbon, Portugal, 2024. ACM.
- [462] William Seymour, Mark Cote, and Jose Such. Legal Obligation and Ethical Best Practice: Towards Meaningful Verbal Consent for Voice Assistants. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [463] Filipo Sharevski and Benjamin Kessell. Fight Fire with Fire: Hacktivists’ Take on Social Media Misinformation. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 19–36, Anaheim, USA, 2023. USENIX Association.
- [464] Tanusree Sharma, Smirity Kaushik, Yaman Yu, Syed Ishtiaque Ahmed, and Yang Wang. User Perceptions and Experiences of Targeted Ads on Social Media Platforms: Learning from Bangladesh and India. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [465] James P. Shaver. What Statistical Significance Testing Is, and What It Is Not. *The Journal of Experimental Education*, 61(4):293–316, 1993.
- [466] Ximing Shen, Yun Suen Pai, Dai Kiuchi, Kehan Bao, Tomomi Aoki, Hikari Meguro, Kanoko Oishi, Ziyue Wang, Sohei Wakisaka, and Kouta Minamizawa. Dementia Eyes: Co-Design and Evaluation of a Dementia Education Augmented Reality Experience for Medical Workers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [467] Swapneel Sheth, Gail Kaiser, and Walid Maalej. Us and them: A study of privacy requirements across north america, asia, and europe. In *Proceedings of the 36th International Conference on Software Engineering*, pages 859–870, Hyderabad, India, 2014. ACM.
- [468] Patrick E. Shrout and Joseph L. Rodgers. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annual Review of Psychology*, 69(1):487–510, 2018.
- [469] Erica Shusas, Patrick Skeba, Eric P. S. Baumer, and Andrea Forte. Accounting for Privacy Pluralism: Lessons and Strategies from Community-Based Privacy Groups. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [470] Amit Kumar Sikder, Leonardo Babun, Z. Berkay Celik, Hidayet Aksu, Patrick McDaniel, Engin Kirda, and A. Selcuk Uluagac. Who’s Controlling My Device? Multi-User Multi-Device-Aware Access Control System for Shared Smart Home Environment. *ACM Transactions on Internet of Things*, 3(4):27:1–27:39, 2022.
- [471] Joseph Simmons, Leif Nelson, and Uri Simonsohn. Pre-registration: Why and How. *Journal of Consumer Psychology*, 31(1):151–162, 2021.

- [472] Adrian Simpson. On the misinterpretation of effect size. *Educational Studies in Mathematics*, 103:125–133, 2020.
- [473] Aneesha Singh, Martin Johannes Dechant, Dilisha Patel, Ewan Soubutts, Giulia Barbareschi, Amid Ayobi, and Nikki Newhouse. Exploring Positionality in HCI: Perspectives, Trends, and Challenges. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2025. ACM.
- [474] Slack. ACM SIGIR slack. <https://acmsigir.slack.com/>, n.d.
- [475] Thomas J Smith and Cornelius M McKenna. A Comparison of Logistic Regression Pseudo R2 Indices. *Multiple Linear Regression Viewpoints*, 39(2):17–26, 2013.
- [476] Ananta Soneji, Faris Bugra Kokulu, Carlos Rubio-Medrano, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, and Adam Doupé. “Flawed, but like democracy we don’t have a better system”: The Experts’ Insights on the Peer Review Process of Evaluating Security Papers. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 1845–1862, San Francisco, USA, 2022. IEEE.
- [477] Jean Y. Song, Sangwook Lee, Jisoo Lee, Mina Kim, and Juho Kim. ModSandbox: Facilitating Online Community Moderation Through Error Prediction and Improvement of Automated Rules. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [478] Sicheng Song, Juntong Chen, Chenhui Li, and Changbo Wang. GVQA: Learning to Answer Questions about Graphs with Visualizations via Knowledge Base. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [479] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. “I did it because I trusted you”: Challenges with the Study Environment Biasing Participant Behaviours. In *SOUPS Usable Security Experiment Reports (USER) Workshop*. ACM, 2010.
- [480] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. How to do better with gender on surveys: A guide for HCI researchers. *Interactions*, 26(4):62–65, 2019.
- [481] Jackson Stokes, Tal August, Robert A Marver, Alexei Czeskis, Franziska Roesner, Tadayoshi Kohno, and Katharina Reinecke. How Language Formality in Security and Privacy Interfaces Impacts Intended Compliance. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [482] Andie Storozuk, Marilyn Ashley, Véronic Delage, and Erin A. Maloney. Got Bots? Practical Recommendations to Protect Online Survey Data from Bot Attacks. *The Quantitative Methods for Psychology*, 16(5):472–481, 2020.
- [483] Christian Stransky, Yasemin Acar, Duc Cuong Nguyen, Dominik Wermke, Elissa M Redmiles, Doowon Kim, Michael Backes, Simson Garfinkel, Michelle L Mazurek, and Sascha Fahl. Lessons Learned from Using an Online Platform to Conduct Large-Scale, Online Controlled Security Experiments with Software Developers. In *10th USENIX Workshop on Cyber Security Experimentation and Test (CSET 17)*, Vancouver, Canada, 2017. USENIX Association.

- [484] Anselm Strauss and Juliet M Corbin. *Grounded Theory in Practice*. Sage, London, 1997.
- [485] David L. Streiner and Geoffrey R. Norman. Correction for Multiple Testing: Is There a Resolution? *CHEST*, 140(1):16–18, 2011.
- [486] Shuyan Sun, Wei Pan, and Lihshing Leigh Wang. A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, 102(4):989–1004, 2010.
- [487] Sathya Chandran Sundaramurthy, Alexandru G. Bardas, Jacob Case, Xinming Ou, Michael Wesch, John McHugh, and S. Raj Rajagopalan. A Human Capital Model for Mitigating Security Analyst Burnout. In *Proceedings of the Eleventh Symposium On Usable Privacy and Security (SOUPS)*, pages 347–359, Ottawa, Canada, 2015. USENIX Association.
- [488] Joshua Sunshine, Serge Egelman, Hazim Almuhiemedi, Neha Atri, and Lorrie Faith Cranor. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *Proceedings of the 18th USENIX Security Symposium (USENIX Security)*, Montréal, Canada, 2009. USENIX Association.
- [489] Jacques Suray, Jan H. Klemmer, Juliane Schmöser, and Sascha Fahl. How the Future Works at SOUPS: Analyzing Future Work Statements and Their Impact on Usable Security and Privacy Research, May 2024.
- [490] Jacques Suray, Jan H. Klemmer, Juliane Schmöser, and Sascha Fahl. Poster: Future Work Statements at SOUPS. In *Proceedings of the Twentieth Symposium on Usable Privacy and Security (SOUPS)*. USENIX Association, 2024.
- [491] Alistair Sutcliffe. Assessing the reliability of heuristic evaluation for Web site attractiveness and usability. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, pages 1838–1847, Big Island, USA, 2002. IEEE.
- [492] Jennifer L. Tackett, Cassandra M. Brandes, Kevin M. King, and Kristian E. Markon. Psychology’s Replication Crisis and Clinical Psychological Science. *Annual Review of Psychology*, 15:579–604, 2019.
- [493] Mohammad Tahaei, Ruba Abu-Salma, and Awais Rashid. Stuck in the Permissions With You: Developer & End-User Perspectives on App Permissions & Their Privacy Ramifications. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [494] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Deciding on Personalized Ads: Nudging Developers About User Privacy. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, pages 573–596. USENIX Association, 2021.
- [495] Mohammad Tahaei and Kami Vaniea. A Survey on Developer-Centred Security. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 129–138, Stockholm, Sweden, 2019. IEEE.
- [496] Mohammad Tahaei and Kami Vaniea. Recruiting Participants With Programming Skills: A Comparison of Four Crowdsourcing Platforms and a CS Student Mailing List. In

- Proceedings of the CHI Conference on Human Factors in Computing Systems*, New Orleans, USA, 2022. ACM.
- [497] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. Understanding Privacy-Related Questions on Stack Overflow. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [498] Anne Clara Tally, Jacob Abbott, Ashley M Bochner, Sanchari Das, and Christena Nippert-Eng. Tips, Tricks, and Training: Supporting Anti-Phishing Awareness among Mid-Career Office Workers Based on Employees' Current Practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [499] Chun Wah Michael Tam, Abeer Hasan Khan, Andrew Knight, Joel Rhee, Karen Price, and Katrina McLean. How Doctors Conceptualise P values: A Mixed Methods Study. *Australian Journal of General Practice*, 47(10):705–710, 2018.
- [500] Jenny Tang, Lujo Bauer, and Nicolas Christin. Misuse, Misreporting, Misinterpretation of Statistical Methods in Usable Privacy and Security Papers. In *Proceedings of the Twentyfirst Symposium on Usable Privacy and Security (SOUPS)*, pages 475–493, Seattle, USA, 2025. USENIX Association.
- [501] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 367–385, Boston, USA, 2022. USENIX Association.
- [502] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. Scripted Vicarious Dialogues: Educational Video Augmentation Method for Increasing Isolated Students' Engagement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [503] Ye Tao, Shuhong Wang, Junzhe Ji, Linlin Cai, Hongmei Xia, Zhiqi Wang, Jinghai He, Yitao Fan, Shengzhang Pan, Jinghua Xu, Cheng Yang, Lingyun Sun, and Guanyun Wang. 4Doodle: 4D Printing Artifacts Without 3D Printers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [504] Beck Taylor, Catherine Henshall, Sara Kenyon, Ian Litchfield, and Sheila Greenfield. Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? A mixed methods study comparing rapid and thematic analysis. *BMJ Open*, 8(10), 2018.
- [505] Jennyfer Lawrence Taylor, Alessandro Soro, Paul Roe, Anita Lee Hong, and Margot Brereton. "Debrief O'Clock": Planning, Recording, and Making Sense of a Day in the Field in Design Research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, Canada, 2018. ACM.
- [506] Arnout Terpstra, Alwin De Rooij, and Alexander Schouten. Online Proctoring: Privacy Invasion or Study Alleviation? Discovering Acceptability Using Contextual Integrity. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.

- [507] David R. Thomas. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2):237–246, 2006.
- [508] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences, May 2024.
- [509] Christian Tiefenau, Maximilian Häring, and Katharina Krombholz. Security, Availability, and Multiple Information Sources: Exploring Update Behavior of System Administrators. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 239–258. USENIX Association, 2020.
- [510] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. A Usability Evaluation of Let’s Encrypt and Certbot: Usable Security Done Right. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1971–1988, London United Kingdom, 2019. ACM.
- [511] Xin Tong, Matthew Louis Mauriello, Marco Antonio Mora-Mendoza, Nina Prabhu, Jane Paik Kim, and Pablo E Paredes Castro. Just Do Something: Comparing Self-proposed and Machine-recommended Stress Interventions among Online Workers with Home Sweet Office. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [512] Phoebe Toups-Dugas, Bongshin Lee, and Marshini Chetty. Guide to a Successful Submission – CHI 2025. <https://chi2025.acm.org/guide-to-a-successful-submission/>, 2024.
- [513] Peter Khoa Duc Tran, Purna Valli Anusha Gadepalli, Jaeyeon Lee, and Aditya Shekhar Nittala. Augmenting On-Body Touch Input with Tactile Feedback Through Fingernail Haptics. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [514] Transparent Statistics in Human–Computer Interaction Working Group. Transparent Statistics Guidelines. <https://transparentstats.github.io/guidelines>, June 2019.
- [515] Transparent Statistics in Human–Computer Interaction Working Group. Transparent Statistics in HCI. <https://transparentstatistics.org/>, April 2024.
- [516] Andrew T. Tredennick, Giles Hooker, Stephen P. Ellner, and Peter B. Adler. A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102(6), 2021.
- [517] Patrizio E. Tressoldi and David Giofré. The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology*, 6, 2015.
- [518] Patrizio E. Tressoldi, David Giofré, Francesco Sella, and Geoff Cumming. High Impact = High Statistical Standards? Not Necessarily So. *PLOS ONE*, 8(2), 2013.
- [519] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Sciences. Addison-Wesley Pub. Co., Reading, Mass., 1977.

-
- [520] Anwesh Tuladhar, Daniel Lende, Jay Ligatti, and Xinming Ou. An Analysis of the Role of Situated Learning in Starting a Security Culture in a Software Company. In *Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS)*, pages 617–632. USENIX Association, 2021.
- [521] Kieron Ivy Turk and Alice Hutchings. Click Here to Exit: An Evaluation of Quick Exit Buttons. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [522] Lucy Turner, Larissa Shamseer, Douglas G. Altman, Laura Weeks, Jodi Peters, Thilo Kober, Sofia Dias, Kenneth F. Schulz, Amy C. Plint, and David Moher. Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database of Systematic Reviews*, 2013(11), 2012.
- [523] Mark Turner, Martin Schmitz, Morgan Masichi Bierey, Mohamed Khamis, and Karola Marky. Tangible 2FA – An In-the-Wild Investigation of User-Defined Tangibles for Two-Factor Authentication. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 245–261, Anaheim, USA, 2023. USENIX Association.
- [524] Sarah Turner, Jason Nurse, and Shujun Li. When Googling It Doesn’t Work: The Challenge of Finding Security Advice for Smart Home Devices. In Steven Furnell and Nathan Clarke, editors, *Human Aspects of Information Security and Assurance*, volume 613, pages 115–126, Cham, 2021. Springer International Publishing.
- [525] Jakob Carl Uhl, Helmut Schrom-Feiertag, Georg Regal, Katja Gallhuber, and Manfred Tscheligi. Tangible Immersive Trauma Simulation: Is Mixed Reality the next level of medical skills training? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [526] Ehsan Ul Haque, Mohammad Maifi Hasan Khan, and Md Abdullah Al Fahim. The Nuanced Nature of Trust and Privacy Control Adoption in the Context of Google. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [527] USENIX. SOUPS 2017 Call for Papers.
<https://www.usenix.org/conference/soups2017/call-for-papers>, November 2016.
- [528] USENIX. SOUPS 2019 Call for Papers.
<https://www.usenix.org/conference/soups2019/call-for-papers>, 2018.
- [529] USENIX. USENIX Security ’24 Call for Papers.
<https://www.usenix.org/conference/usenixsecurity24/call-for-papers>, 2023.
- [530] USENIX. USENIX Security ’25 Call for Papers | USENIX.
<https://www.usenix.org/conference/usenixsecurity25/call-for-papers>, 2024.
- [531] USENIX. SOUPS 2025 Call for Papers.
<https://www.usenix.org/conference/soups2025/call-for-papers>, January 2025.

- [532] Warda Usman, Jackie Hu, McKynlee Wilson, and Daniel Zappala. Distrust of big tech and a desire for privacy: Understanding the motivations of people who have voluntarily adopted secure email. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 473–490, Anaheim, USA, 2023. USENIX Association.
- [533] Tammi Vacha-Haase and Bruce Thompson. How to Estimate and Interpret Various Effect Sizes. *Journal of Counseling Psychology*, 51(4):473–481, 2004.
- [534] Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G. Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, and Christopher Yau. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1), 2021.
- [535] Dirk van der Linden, Pauline Anthonysamy, Bashar Nuseibeh, Thein Than Tun, Marian Petre, Mark Levine, John Towse, and Awais Rashid. Schrödinger’s Security: Opening the Box on App Developers’ Security Rationale. In *Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering (ICSE)*, pages 149–160, Seoul, South Korea, 2020. ACM.
- [536] Kami E. Vaniea, Emilee Rader, and Rick Wash. Betrayed by updates: How negative experiences affect future security. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2671–2674, Toronto, Canada, 2014. ACM.
- [537] Ivan Vankov, Jeffrey Bowers, and Marcus Munafò. On the persistence of low power in psychological science. *Quarterly journal of experimental psychology*, 67(6):1037–1040, 2014.
- [538] Veniamin Veselovsky, Manoel Horta Ribeiro, Philip J. Cozzolino, Andrew Gordon, David Rothschild, and Robert West. Prevalence and Prevention of Large Language Model Use in Crowd Work. *Communications of the ACM*, 68(3):42–47, 2025.
- [539] Ruben Vicente-Saez and Clara Martinez-Fuentes. Open Science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88:428–436, 2018.
- [540] Dong-Bach Vo, Sylvain Pauchet, Isabelle Jolly, Florine Simon, Anke M. Brock, and Jérémie Garcia. Tactilient: Turbulence resilient tactile icons for pilot feedback. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [541] Jan B. Vornhagen, April Tyack, and Elisa D. Mekler. Statistical Significance Testing at CHI PLAY: Challenges and Opportunities for More Transparency. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pages 4–18, Ottawa, Canada, 2020. ACM.
- [542] Daniel Votipka, Desiree Abrokwa, and Michelle L. Mazurek. Building and Validating a Scale for Secure Software Development Self-Efficacy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
- [543] Daniel Votipka, Kelsey R Fulton, James Parker, Matthew Hou, Michelle L Mazurek, and Michael Hicks. Understanding security mistakes developers make: Qualitative analysis from Build It, Break It, Fix It. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security)*, pages 109–126. USENIX Association, 2020.

- [544] Daniel Votipka, Seth Rabin, Kristopher Micinski, Jeffrey S Foster, and Michelle L Mazurek. An Observational Investigation of Reverse Engineers' Processes. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security)*, pages 1875–1892. USENIX Association, 2020.
- [545] Daniel Votipka, Rock Stevens, Elissa Redmiles, Jeremy Hu, and Michelle Mazurek. Hackers vs. Testers: A Comparison of Software Vulnerability Discovery Processes. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 374–391, San Francisco, USA, 2018. IEEE.
- [546] Chat Wacharamanotham, Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. Special interest group on transparent statistics guidelines. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, Canada, 2018. ACM.
- [547] April Yi Wang, Dakuo Wang, Jaimie Drozdal, Xuye Liu, Soya Park, Steve Oney, and Christopher Brooks. What Makes a Well-Documented Notebook? A Case Study of Data Scientists' Documentation Practices in Kaggle. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, 2021. ACM.
- [548] Leijie Wang, Ruotong Wang, Sterling Williams-Ceci, Sanketh Menda, and Amy X. Zhang. "Is Reporting Worth the Sacrifice of Revealing What I've Sent?": Privacy Considerations When Reporting on End-to-End Encrypted Platforms. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 491–508, Anaheim, USA, 2023. USENIX Association.
- [549] Tony Wang, Haard K Shah, Raj Sanjay Shah, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. Metrics for Peer Counseling: Triangulating Success Outcomes for Online Therapy Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [550] Yiwen Wang, Ziming Li, Pratheep Kumar Chelladurai, Wendy Dannels, Tae Oh, and Roshan L Peiris. Haptic-Captioning: Using Audio-Haptic Interfaces to Enhance Speaker Indication in Real-Time Captions for Deaf and Hard-of-Hearing Viewers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [551] Yuntao Wang, Zirui Cheng, Xin Yi, Yan Kong, Xueyang Wang, Xuhai Xu, Yukang Yan, Chun Yu, Shwetak Patel, and Yuanchun Shi. Modeling the Trade-off of Privacy Preservation and Activity Recognition on Low-Resolution Images. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 15, Hamburg, Germany, 2023. ACM.
- [552] Zijie J. Wang, Jennifer Wortman Vaughan, Rich Caruana, and Duen Horng Chau. GAM Coach: Towards Interactive and User-centered Algorithmic Recourse. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [553] Rick Wash. Folk Models of Home Computer Security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS)*, Redmond, USA, 2010. ACM.

- [554] Rick Wash and Emilee Rader. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In *Proceedings of the Eleventh Symposium On Usable Privacy and Security (SOUPS)*, pages 309–325, Ottawa, Canada, 2015. USENIX Association.
- [555] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. Understanding Password Choices: How Frequently Entered Passwords Are Re-used across Websites. In *Proceedings of the Twelfth Symposium on Usable Privacy and Security (SOUPS)*, pages 175–188, Denver, USA, 2016. USENIX Association.
- [556] Rick Wash, Emilee Rader, Kami Vaniea, and Michelle Rizor. Out of the Loop: How Automated Software Updates Cause Unintended Security Consequences. In *Proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS)*, pages 89–104, Menlo Park, USA, 2014. USENIX Association.
- [557] Renita Washburn, Tangila Islam Tanni, Yan Solihin, Apu Kapadia, and Mary Jean Amon. Bottom-up psychosocial interventions for interdependent privacy: Effectiveness based on individual and content differences. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [558] Ronald L. Wasserstein, Allen L. Schirm, and Nicole A. Lazar. Moving to a World Beyond “ $p < 0.05$ ”. *The American Statistician*, 73(sup1):1–19, 2019.
- [559] Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Franziska Roesner, and Kurt Thomas. “There’s so much responsibility on users right now.” Expert Advice for Staying Safer From Hate and Harassment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [560] Miranda Wei, Jaron Mink, Yael Eiger, Tadayoshi Kohno, Elissa M. Redmiles, and Franziska Roesner. SoK (or SoLK?): On the Quantitative Study of Sociodemographic Factors and Computer Security Behaviors. In *Proceedings of the 33rd USENIX Security Symposium (USENIX Security)*, pages 7011–7030, Philadelphia, USA, 2024. USENIX Association.
- [561] Miranda Wei, Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. Anti-Privacy and Anti-Security advice on TikTok: Case studies of Technology-Enabled surveillance and control in intimate partner and Parent-Child relationships. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 447–462, Boston, USA, 2022. USENIX Association.
- [562] Charles Weir, Ben Hermann, and Sascha Fahl. From Needs to Actions to Secure Apps? The Effect of Requirements and Developer Practices on App Security. In *Proceedings of the 29th USENIX Security Symposium (USENIX Security)*, pages 289–305. USENIX Association, 2020.
- [563] Charles Weir, Awais Rashid, and James Noble. How to Improve the Security Skills of Mobile App Developers? Comparing and Contrasting Expert Views. In *Proceedings of the Twelfth Symposium on Usable Privacy and Security (SOUPS)*, Denver, USA, 2016. USENIX Association.

- [564] Frederick J. Wertz, Kathy Charmaz, Linda M. McMullen, Ruthellen Josselson, Rosemarie Anderson, and Emalinda McSpadden. *Five Ways of Doing Qualitative Analysis: Phenomenological Psychology, Grounded Theory, Discourse Analysis, Narrative Research, and Intuitive Inquiry*. The Guilford Press, New York, 1st edition, 2011.
- [565] Gareth R. White, Pejman Mirza-babaei, Graham McAllister, and Judith Good. Weak inter-rater reliability in heuristic evaluation of video games. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, pages 1441–1446, Vancouver, Canada, 2011. ACM.
- [566] Alma Whitten and J. D. Tygar. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *Proceedings of the 8th USENIX Security Symposium (USENIX Security)*, pages 169–184, Washington, D.C., USA, 1999. USENIX Association.
- [567] Corrine M. Wickens. The investigation of power in written texts through the use of multiple textual analytic frames. *International Journal of Qualitative Studies in Education*, 24(2):151–164, 2011.
- [568] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemond, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kokske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.
- [569] Olivia Williams, Yee-Yin Choong, and Kerriane Buchanan. Youth Understandings of Online Privacy and Security: A Dyadic Study of Children and Their Parents. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 399–416, Anaheim, USA, 2023. USENIX Association.
- [570] Max Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, and Jeffrey Nichols. RepliCHI SIG: From a panel to a new submission venue for replication. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, pages 1185–1188, Austin, USA, 2012. ACM.
- [571] Max L. Wilson, Ed H. Chi, Stuart Reeves, and David Coyle. RepliCHI: The workshop II. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, pages 33–36, San Jose, USA, 2014. ACM.
- [572] Max L. Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, Dan Russell, and Harold Thimbleby. RepliCHI - CHI should be replicating and validating results more: Discuss. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, pages 463–466, Vancouver, Canada, 2011. ACM.
- [573] Max L. L. Wilson, Paul Resnick, David Coyle, and Ed H. Chi. RepliCHI: The workshop. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, pages 3159–3162, Paris, France, 2013. ACM.
- [574] Donna M Windish, Stephen J Huot, and Michael L Green. Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature. *JAMA*, 298(9):1010–1022, 2007.

- [575] Maximiliane Windl, Anna Scheidle, Ceenu George, and Sven Mayer. Investigating Security Indicators for Hyperlinking Within the Metaverse. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, Anaheim, USA, 2023. USENIX Association.
- [576] Maximiliane Windl, Albrecht Schmidt, and Sebastian S. Feger. Investigating Tangible Privacy-Preserving Mechanisms for Future Smart Homes. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [577] Maximiliane Windl, Verena Winterhalter, Albrecht Schmidt, and Sven Mayer. Understanding and Mitigating Technology-Facilitated Privacy Violations in the Physical World. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [578] Leon Würsching, Florentin Putz, Steffen Haesler, and Matthias Hollick. FIDO2 the Rescue? Platform vs. Roaming Authentication on Smartphones. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [579] Ziang Xiao, Tiffany Wenting Li, Karrie Karahalios, and Hari Sundaram. Inform the Uninformed: Improving Online Informed Consent Reading with an AI-Powered Chatbot. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [580] Chengyuan Xu, Kuo-Chin Lien, and Tobias Höllerer. Comparing Zealous and Restrained AI Recommendations in a Real-World Human-AI Collaboration Task. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [581] Khaled Yakdan, Sergej Dechand, Elmar Gerhards-Padilla, and Matthew Smith. Helping Johnny to Analyze Malware: A Usability-Optimized Decompiler and Malware Analysis User Study. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 158–177, San Jose, USA, 2016. IEEE.
- [582] Xin Yi, Shuning Zhang, Ziqi Pan, Louisa Shi, Fengyan Han, Yan Kong, Hewu Li, and Yuanchun Shi. Squeeze’In: Private Authentication on Smartphones based on Squeezing Gestures. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2023. ACM.
- [583] Bereket A. Yilma and Luis A. Leiva. The Elements of Visual Art Recommendation: Learning Latent Semantic Representations of Paintings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [584] Miuyin Yong Wong, Matthew Landen, Manos Antonakakis, Douglas M. Blough, Elissa M. Redmiles, and Mustaque Ahamad. An Inside Look into the Practice of Malware Analysis. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3053–3069, Seoul, South Korea, 2021. ACM.
- [585] Alyson L. Young, Wayne G. Lutters, Nicholas R. Magliocca, and Erle C. Ellis. Designing a system for land change science meta-study. In *CHI ’13 Extended Abstracts on Human Factors in Computing Systems*, pages 1473–1478, Paris, France, 2013. ACM.

- [586] Koen Yskout, Riccardo Scandariato, and Wouter Joosen. Does organizing security patterns focus architectural choices? In *2012 34th International Conference on Software Engineering (ICSE)*, pages 617–627, Zürich, Switzerland, 2012. IEEE.
- [587] Koen Yskout, Riccardo Scandariato, and Wouter Joosen. Do security patterns really help designers? In *Proceedings of the 37th International Conference on Software Engineering*, pages 292–302, Florence, Italy, 2015. IEEE.
- [588] Oleg Zendel, J. Shane Culpepper, Falk Scholer, and Paul Thomas. Enhancing Human Annotation: Leveraging Large Language Models and Efficient Batch Processing. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 340–345, Sheffield, UK, 2024. ACM.
- [589] Yan Zhang, Ziang Li, Haole Guo, Luyao Wang, Qihe Chen, Wenjie Jiang, Mingming Fan, Guyue Zhou, and Jiangtao Gong. "I am the follower, also the boss": Exploring Different Levels of Autonomy and Machine Forms of Guiding Robots for the Visually Impaired. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, 2023. ACM.
- [590] Zhuohao (Jerry) Zhang, Smirity Kaushik, JooYoung Seo, Haolin Yuan, Sauvik Das, Leah Findlater, Danna Gurari, Abigale Stangl, and Yang Wang. ImageAlly: A Human-AI Hybrid Approach to Support Blind People in Detecting and Redacting Private Image Content. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 417–436, Anaheim, USA, 2023. USENIX Association.
- [591] Leah Zhang-Kennedy, Saira Aziz, Oluwafunminitemi (Temi) Oluwadare, Lyndon Pan, Zeyu Wu, Sydney E.C. Lamorea, Soda Li, Michael Sun, and Ville Mäkelä. Passenger Perceptions, Information Preferences, and Usability of Crowding Visualizations on Public Displays in Transit Stations and Vehicles. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 15, Hamburg, Germany, 2023. ACM.
- [592] Sarah Y. Zheng and Ingolf Becker. Checking, nudging or scoring? Evaluating e-mail user security tools. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 57–76, Anaheim, USA, 2023. USENIX Association.
- [593] Zhixuan Zhou, Tanusree Sharma, Luke Emano, Sauvik Das, and Yang Wang. Iterative Design of An Accessible Crypto Wallet for Blind Users. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 381–398, Anaheim, USA, 2023. USENIX Association.
- [594] Samira Zibaei, Amirali Salehi-Abari, and Julie Thorpe. Dissecting Nudges in Password Managers: Simple Defaults are Powerful. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS)*, pages 211–225, Anaheim, USA, 2023. USENIX Association.
- [595] Verena Zimmermann, Stina Schäfer, Markus Dürmuth, and Karola Marky. Authenticate as You Go: From Exploring Smart Home Authentication with Daily Objects to Authenticating with Primary Tasks. *ACM Transactions on Computer-Human Interaction*, 32(2):13:1–13:69, 2025.

-
- [596] Yixin Zou, Kevin Roundy, Acar Tamersoy, Saurabh Shintre, Johann Roturier, and Florian Schaub. Examining the Adoption and Abandonment of Security, Privacy, and Identity Theft Protection Practices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, USA, 2020. ACM.
 - [597] Noé Zufferey, Sarah Abdelwahab Gaballah, Karola Marky, and Verena Zimmermann. “AI is from the devil.” Behaviors and Concerns Toward Personal Data Sharing with LLM-based Conversational Agents. *Proceedings on Privacy Enhancing Technologies*, 2025(3):5–28, 2025.

Appendix A

Background

This appendix contains additional material related to the work presented in Chapter 2. It contains lists of papers which we used to investigate whether effect sizes are reported in USP publications at CHI and SOUPS 2023, in Sections A.1 and A.2 respectively. Data, including links, excerpts and analysis scripts used to analyze the data on calls for papers in USP is on OSF: <https://osf.io/sthj8/>.

A.1 USP Papers Evaluated for Effect Size Reporting at CHI'23

Alhelali et al. [13], Alshehri et al. [14], Borgert et al. [56], Bouma-Sims et al. [60], Chanenson et al. [80], Cheon et al. [83], Dav et al. [117], Distler [121], Elagroudy et al. [135], Emami-Naeini et al. [137], Fassl et al. [148], Freed et al. [158], Ganesh et al. [169], Glas et al. [179], Gomez Ortega et al. [180], Hasan et al. [206], Herbert et al. [214], Häring et al. [205], Hutton et al. [230], Im et al. [232], Kanei et al. [254], Kowalczyk et al. [282], Kyi et al. [290], Lewis et al. [302], Marin et al. [320], Mendel et al. [329], Ortloff et al. [365], Plöger et al. [396], Rajaram et al. [411], Roemmich et al. [428], Tally et al. [498], Terpstra et al. [506], Sarathy et al. [445], Seymour et al. [462], Sharma et al. [464], Shusas et al. [469], Stokes et al. [481], Tahaei et al. [493], Turk et al. [521], Ul Haque et al. [526], Wang et al. [551], Washburn et al. [557], Wei et al. [559], Windl et al. [576], Windl et al. [577], Würsching et al. [578], Yi et al. [582]

A.2 USP Papers Evaluated for Effect Size Reporting at SOUPS'23

Abramova et al. [3], Akter et al. [12], Amador et al. [16], Klivan et al. [278], Ashraf et al. [22], Deldari et al. [113], Evripidou et al. [139], Gerlitz et al. [175], Gerlitz et al. [177], Hazazi et al. [210], Hielscher et al. [217], Huang et al. [226], Kaushik et al. [261], Gage Kelley et al. [267], Kersten et al. [270], Liu et al. [311], McCall et al. [326], Neil et al. [347], Ortloff et al. [369], Panskus et al. [376], Park et al. [378], Rader et al. [409], Serafini et al. [459], Sharevski et al. [463], Turner et al. [523], Usman et al. [532], Wang et al. [548], Williams et al. [569], Windl et al. [575], Zhang et al. [590], Zheng et al. [592], Zhou et al. [593], Zibaei et al. [594]

Appendix B

Influence Factors on Results from Qualitative Analysis

This appendix contains additional material related to the work presented in Chapter 3. First we provide those data collection instruments unique to that publication: The interview guideline which the course participants used to gain insight into the interviewee's reasons for trusting security advice, and their perceptions of realism and effectiveness of security advice is in Section B.1. The data collection instrument for the survey data were questions from prior work, and we refer interested parties to Busse et al. for the full questionnaire [74]. We also include the instructions given to the course participants and researcher groups coding this interview data in Section B.1.1 and the instructions for coding the survey data in Section B.1.2. Note that these instructions were translated to English from the original language by the authors.

Second, we provide an overview of our coding for the meta-level analysis of the coding process, as well as the survey of PC members. We show the codebooks used to compare the content of interview outcomes (Section B.2.1) and survey outcomes (Section B.2.2). We also provide both the survey questions for the survey of SOUPS PC members, which is in Section B.3, as well as the codebook resulting from the analysis of this data (Section B.4).

B.1 Data Collection Instruments

Interview Guideline

B.1.0.1 Preparation

The day before the interview

- ☐ Conduct a test recording using Zoom and find out where the generated files from this recording are saved

- ☐ Make sure that there are about 200 - 300 MB of disk space available for the Zoom recordings (numbers determined from a pre-test interview)
- ☐ Conduct a test recording with a backup recording device (probably your smartphone with a recording application) and try out where to position it, so that both the interviewer and the interviewee are clearly intelligible.

Directly before the interview

- ☐ Prepare a computer with a webcam in a quiet room
- ☐ Have paper and pen ready for taking notes
- ☐ Have this interview guideline available (either on a separate device / second screen or as a print-out)
- ☐ Keep the backup recording device ready, and make sure the battery charge status is sufficient
- ☐ open the video conferencing system (Zoom) and send your participant the invitation link to the Zoom room
- ☐ Link to the Informed Consent: [\[Link\]](#)
- ☐ Note down the time when interaction with your participant starts

Greeting

Hello, my name is (name) and I will lead you through this interview today.

[Depending on situation, some small talk may be appropriate to gain rapport with the participant / relax the situation.]

Before we start, I have some information for you, which you should read thoroughly. I have already sent them to you before this meeting, but I will send you a link to a survey, which contains the same information about the goals and the procedure of this study. At the end of the survey, we ask you for your consent to participate in the the interview and that we can record this conversation. Unfortunately, Zoom does not allow only recording audio, but we will only use the audio file of this conversation for analysis and will delete the additional video file directly after our meeting is over. We additionally ask for your name in the consent survey, since we otherwise cannot establish that you have consented to participate in this interview. We will not connect this information with the interview itself and will save it separately from the interview data.

You can find more detailed information at this link: [\[Link\]](#)

☐ **Send informed consent link to participant in Zoom-Chat: [Link]**

Before participant begins to read:

You can either give consent by filling out the survey, by signing the PDF you have previously received digitally, or by printing the PDF, signing it physically and scanning it. Which do you prefer?

If survey

Please send me a screenshot in the chat of your consent (name does not need to be visible)

If PDF

Please send me the signed PDF in the chat.

If you have any questions about the consent form, you can ask them at any time.

Wait until you get confirmation of consent.

At this point, do you have any more questions?

If that is okay for you, I will start the recording now.

Wait for interviewee affirmation / consent.

- ☐ Save screenshot / PDF
- ☐ Start recording in Zoom
- ☐ Start recording on the backup recording device and position it, so that the sound input is good for interviewer and interviewee
- ☐ Write down the time for the beginning of the actual interview

Interview

- ☐ Note down technical terms which interviewees use during the interview
- ☐ If not enough is said on a topic during the interview, you can try to dig deeper by asking additional questions: Could you please expand on this a bit? / Can you explain this in more detail? / How exactly ...? / Why exactly is this ...?

At the beginning, I have some general questions about you. I want to point out again that you can decline to answer any question, if you feel uncomfortable about answering.

- How old are you?
- As which gender do you identify?
- What is the highest level of education you have attained?
- What is your main occupation? *Job/university*
 - *If student*
 - What subjects do you study?
 - *If job:*
 - In which field of business do you work?
 - What is your job title?
 - *If an IT-related area was named for studied subjects or job:*
 - How many years of experience do you have in the area of IT?
 - How many years of experience do you have in the area of IT security?
- How do you judge your knowledge about IT security?
 - *Let the participants answer in words first.*
 - On a scale from 1 (very low) to 7 (very high), how do you judge your knowledge?
 - Why? / Could you explain this a bit more?
- How much technical affinity do you think you have?
 - *Let the participants answer in words first.*
 - On a scale from 1 (no technical affinity at all) bis 7 (very high technical affinity), where would you position yourself?
 - Why? / Could you explain this a bit more?

This interview is about IT security advice.

- What are the top three best pieces of advice which you would give to a non-tech-savvy user, to protect their IT security?

☐ Note down the advice in bullet points to have them available for the rest of the interview.

Note for interviewers: If the participant does not refer back to each of the pieces of advice in the following questions, explicitly inquire about the respective advice.

- Why did you choose these three pieces of advice?
- In how far do you follow this advice in your private life?
 - Why? / Why not?
- In how far do you follow this advice in your work life / your university life?
 - Is this different than for your private behavior?
 - Why? / Why not?
- How did you hear of these pieces of advice?
- Why do you trust these pieces of advice?

Let's look at each of your advice in turn.

Note for interviewer: The following questions will be repeated for each of the pieces of advice.

ideas for phrasing: You named advice [x] before. / You had [y] on your list of top 3 advice. / You had also mentioned [Z] as an important piece of advice before.

- How likely do you think it is that a non-tech-savvy user (e.g. an older person) would actually use [advice]?
 - Why?
- How useful do you think [advice] is to protect the IT security of the users from threats?
 - Why? / What make the advice useful for this purpose in your eyes?

We are also interested in your opinion on two specific examples of IT security advice.

- ☐ In this part of the interview, we want to ask about advice which the participant has not named before, i.e. which weren't in their top three.
- ☐ If one of the pieces of advice has already been named by the participant, you can try to dig deeper with some additional questions, or any questions which you have already asked, can be skipped, depending on how detailed the participants answers were before. Questions which were not explicitly asked before:

- What do you think about [advice]?
 - How does [advice] improve your IT security?
- We have alternatives for the first piece of advice: If the first advice (2FA) has already been named, use the next one (use a unique password everywhere), if this advice has also already been named, then use “Be suspicious about links”.

Let's talk about the advice [to use two-factor authentication /
to use a unique password everywhere /
to be suspicious about links] first.

Note for interviewers: This advice is highly effective, but not realistic, and is popular with experts.

- Have you heard of this advice before?
 - If no, explanation:
 - **Two-factor authentication** is the proof of identity of user through a combination of two different and independent components. An example is using a password and a transaction number in online banking.
 - **Using a unique password everywhere** means that you do not re-use passwords, but have a different password for every application.
 - **Being suspicious about links** means that you are careful when interacting with link, e.g. online or in e-mails, and that you do not just click on them, but investigate them more closely, e.g. by hovering over them, where applicable.
- What do you think about this advice?
- How often do you yourself [use two-factor authentication /
use different passwords /
are you yourself suspicious about links.]
 - Why?
- How realistic do you think it is, that a non-tech-savvy user would actually apply this advice?
 - Why?
- How does [two-factor authentication /
the use of different passwords /
suspicion about links] improve your IT security?
- How useful do you think this advice is for improving IT security?
 - Why? What makes the advice (not) effective/useful in your eyes?

Let's talk about antivirus software now.

Note for interviewers: This advice has lower effectiveness, but is more realistic, and is not very popular with experts.

- Have you heard of this advice before?
 - If no: explanation
 - **Antivirus software** is software which is supposed to find, block and, if applicable, dispose of malware, like computer viruses.
- What do you think about this advice?
- Do you yourself use antivirus software?
- Why ? / Why not?
- How realistic do you think it is, that a non-tech-savvy user would actually apply this advice?
 - Why?
- How does antivirus software improve your IT security?
- How useful do you think this advice is for improving IT security?
 - Why? What makes the advice (not) effective/useful in your eyes?

- ☐ Check the duration of the interview
- ☐ If the interview has already taken **half an hour** or longer, then skip the block of questions on password managers.
- ☐ If the interview is shorter, you can continue with the questions on password managers.

Next I want to talk about password managers.

Note for interviewers: This advice has relatively high effectiveness and realism ratings, and is popular with experts.

- Have you heard of this advice before?
 - If no: explanation:
 - **A password manager** is an application, with which users can encryptedly save, manage and use access data, such as passwords or pins.

- What do you think about this advice?
- To which extent do you yourself use a password manager?
 - Why?
- How realistic do you think it is, that a non-tech-savvy user would actually apply this advice?
 - Why?
- How does a password manager improve your IT security?
- How useful do you think this advice is for improving IT security?
 - Why? What makes the advice (not) effective/useful in your eyes?

☐ If the questions on password managers were skipped, then continue from here.

Since we have talked about several specific pieces of advice now, I would like to ask you for a short summary:

- What makes IT security advice realistic to apply for you?
- Which aspects influence your judgment of how effective a piece of IT security advice is in protecting users from threats?

And to conclude, I have some questions on your attitude towards IT security advice.

- How do you decide whether to follow an advice or not? *Possibly dig deeper for adoption reasons.*
- Which characteristics make an advice trustworthy for you?
- What role does the source of an advice play for you in this context?

Thank you, that was all from my side. Do you have any questions, or anything else you want to say about this topic?

Give participant some time to think about it, wait for their answer.

Okay, then I will stop the recording now.

- ☐ Stop Zoom recording. (The generation of the files may take a while depending on the duration of the interview)
- ☐ Stop the recording on the backup device
- ☐ **Thank the participant for their participation**

Debriefing

Do you have any further questions on the topic or the procedure of the interview?

Possible topics regarding the background of the study/the interview:

- Research questions
 - Why is some security advice rated effective, but not realistic?
 - What factors influence how much security advice is trusted?
- Information on specific pieces of security advice and their choice (see notes for interviews on this topic)

Follow-up work

After the interview

- ☐ Once Zoom has finished saving: Delete the video recording and chat protocol.
- ☐ Save audio recording using the naming format yyyy-mm-dd.m4a, and replace **yyyy-mm-dd** with the date of the interview
- ☐ check whether the audio recording worked
- ☐ If the audio recording from Zoom did not work, check the audio recording of your backup device and name it using the scheme presented above
- ☐ Go through your notes on technical terms and digitize them
- ☐ Hand in: Audio recording (only one working recording, either from Zoom or the backup device), your notes on technical terms and the consent form
- ☐ Possibly delete the superfluous recording of your backup recording device

B.1.1 Interview Analysis Task Description

Admin

You have access to the interviews assigned to you through [E-learning system]. They are in a folder named *GR-[your group id]-[course]-2*. We have already prepared a MaxQDA-Project for each of you and loaded the interviews to be analyzed. Please use two interviews (marked with initial) to establish a codebook and afterwards code the remaining four interviews with this codebook. In this group-folder there is also a folder *hand-in*, where you should save hand-ins concerning the coding.

Please make regular backup copies of your work, especially before doing irreversible things, like merging or changing your codebook radically. To do this, close MaxQDA and copy the mx20-project file. Name the backup copies as follows:

- Append the version number to the end of the previous file name.
- Version number in format [submissionnumber.versionnumber]
- Iterate up from version v0-0 (for the file provided by us)
 - Submission number 1 is the file in which you alone code the two interviews and develop your own codebook
 - Submission number 2 is the file in which you use the merged codebook to recode the two interviews.
 - Submission number 3 is the file in which you will code four more interviews using your merged codebook
- *Your submission of your first two independently coded interviews would then have e.g. the version number v1-3 (depending on how often you made backup copies).*

In any case, please do not modify the transcripts yourself, but only add codes and/or memos to them.

Reminder: The research questions we are interested in are:

- Why is some IT security advice rated effective, but not realistic?
- What factors influence how much IT security advice is trusted?

Please analyze and code with respect to these research questions

Preparation

Install MaxQDA. Instructions can be found on [E-learning system] along with this exercise assignment. In case you have any problems, please contact us.

Discuss with your team member roughly how large the text segments you want to code should be. Possibilities are e.g.:

- Line-wise (if you choose this, make sure your and your partner's lines are of equal length)
- Sentence-wise
- Partial sentence-wise (Separation e.g. at subordinate clauses etc.)

Coding of the first two interviews

Start with reading the first two interviews carefully. You may take notes, but don't add any codes yet.

Once you are familiar with the content of the first two interviews, start coding. To do this, you mark text fragments and assign codes to them. These codes can be in-vivo-codes, i.e. their phrasing comes directly from the text, or the codes can be abstracted and reformulated by you. A text fragment can also be assigned several codes.

Work through the initial interviews several times and check whether codes you have created later may also be relevant for other parts of the interviews. Work alone at this stage and don't discuss with your partner yet.

Consider your codes/your codebook. Which of them belong together or deal with related topics? Group your codes and create categories (if you have not already done this in the last step), so that you build a hierarchical code structure. If it seems to be reasonable, you can also create several levels of categories. It may be useful for more abstract categories (high-level) to orientate towards the research questions.

Now save this version of your codebook and the current state of your coded interviews to the submission subfolder on [E-learning system].

- To export your codebook, on the codes tab, click on Export Code System and then select the .rtf format. Name the file as your project.
- To save the state of your coding, make a backup of your MaxQDA project (file format .mx20)

You should be at a submission number of 1 in your versioning. It is best to upload these files to the delivery folder on [E-learning system].

Now unify your codebooks with your partner. Compare both your codebooks and how you assigned codes.

- Which of your codes and categories are similar? Agree on a wording for the code tag.
- Have you assigned similar codes to similar text fragments? Why?/Why not?
- Which of your codes and categories are different? Do you want to adopt several of them or discard some codes or categories?

If you want to see the codebooks together in one project, you should make a backup first. To do this, you have to select the option Teamwork under the Start tab (more info [here](#) under the heading *Transfer elements – Teamwork export and import* and [here](#)).

- One of you has to do Teamwork Export.
- In the following dialogue, select all of your documents and all of your codes.
- The result is a .mex file.
- The other one of you selects Teamwork Import
- Select all of your codes and files again in the following dialogue. As conflict resolving process, select *Use outer segment boundaries of the code assignments*.

Now you should have both codebooks available in one project and be able to edit them. It is not possible for two people to work together on the same MaxQDA project, so for this step it is a good idea for one team partner to activate screen sharing and work on the project while the other partner provides advice. Agree on one codebook (which can/should contain elements of both of your individual codebooks).

In order to work in a consistent manner, you should create definitions for your codes and categories and describe for which cases a certain code should be assigned. You should save these definitions in form of code memos in MaxQDA.

Save this unified version of your codebook and use the export function again, to get the codebook in the .rtf format. In your versioning, your submission number should now be 2. It is best to upload the file directly into the submission subfolder. Even if both group members now use the same codebook, please upload the unified codebook for each group member.

Using the unified codebook, work through both initial interviews and code them again. Again, it may be useful to go through both interviews several times. For this step, work alone again without consulting your partner.

Then save the current status of your coded initial interviews by creating a backup copy of the MaxQDA project. In the versioning you should be at submission number 2. It is best to upload these files to the submission folder of your group right away. Again, each group member should submit a separate project file.

Coding of the remaining four interviews

Next, you will code the remaining four interviews using your unified codebook. To do this, proceed in a similar way as you did for the initial interviews and work alone in this phase. Mark relevant text fragments with the codes from your unified codebook and work through the four remaining interviews this way.

During the coding process, aspects may come up in the four additional interviews that were not present in the initial interviews and therefore are not covered in your codebook. In such cases, you can revise your codebook and add new codes or rename/extend old codes and adjust their definitions. Discuss the changes with your partner and agree on a procedure so that you continue using the same codebook. You don't have to call a meeting every time you change a code, you can also tell your team partner when you add a new code and only meet if they don't find it useful.

Finally, you should go again through all six interviews with your final codebook and check that the codes you assigned are still appropriate and correspond to the final codebook.

Now save this version of your codebook and your coded six interviews. For the codebook, export another .rtf file and name it as your project. You should now be at submission number 3.

Communicating results

Discuss the result of your analysis with your partner: the final codebooks and the developed codes. Depending on how reduced the top level of your codebook already is, you should make further groupings and sortings so that you have only 3-8 categories on the top level. Also consider how these are related and what the link between the categories is. Using the "MaxMaps" function of MaxQDA (start the help on this topic [here](#)) or another mindmapping tool, design a mind map that represents your analysis results. MaxMaps works similarly to the *creative coding* interface you've already encountered for organizing and unifying your codebooks. The main difference is that in Creative Coding, changes you make to the code system (renaming, adding or deleting codes, etc.) are then reflected in your codebook. This is not the case for MaxMaps, but here you have the possibility to add further elements, e.g. line labels or similar. This mind map does not have to contain all your codes and categories, you can limit yourself to the ones you find important.

Save and upload this mind map as .png file (function Export MaxMap) right away to the submission folder in the [E-learning system] folder shared with your partner.

Submission

Please submit the following files in the [E-learning system] folder:

- The first version of your codebook (that you have developed individually), as .rtf file (versioning 1-x)
- The first version of the interviews coded by you (initial interviews) – i.e. the .mx20 file of your project (versioning 1-x)

- The second version of your codebook (that you have unified together with your partner), as .rtf file (versioning 2-x)
- The initial interviews recoded with this unified codebook – again, the .mx20 file of your project (versioning 2-x)
- The final version of your codebook (after you coded all interviews using this codebook, as .rtf file (versioning 3-x)
- All six interviews you coded with this codebook, i.e. the .mx20 file of the final state of your project (versioning 3-x)
- Your mindmap to your research results, as .png file

B.1.2 Survey Analysis Task Description

This assignment was handed out in form of slides due to practical reasons.

Developing a codebook:

- (Hierarchical) list of codes
- Multi-level hierarchies possible
- Memos to explain codes and note down a common definition
- Process:
 - Start by reading over the advices assigned to you with your team partner
 - Brainstorm codes for the type of advice (can be in-vivo or abstract codes)
 - Group the codes and assign names to the group (= categories)

Coding in Excel (or other spreadsheets)

- Last year: We used MaxQDA (CAQDA-Software)
- This year: No more Microsoft Azure Dev Tools for Teaching available
- MaxQDA is only available for Windows/MacOS
- → Switch to Spreadsheet software

You will receive a Spreadsheet where each row corresponds to one advice from from the survey answers.

- There are separate columns for
 - the participant id

- From which study the advice came (usage – replication, advice – new version)
- participant utterance (in this case, the advice)
- the assigned code (This is your task, Excel makes a suggestion based on previous entries)
- Copy a row of data if you want to assign more than one code
- When adding a new code, also add it to your codebook

Hand-in

- Each group should hand in the following [E-learning system] folder: *Link*
 - our shared codebook (spreadsheet filetype, e.g. .ods or .xlsx), named your_id1 __[your_id2]_codebook
 - e.g. for us: [naming]_codebook.ods
- After you hand in your codebook, we give you access to the template for your advice data
 - We check for new hand-ins once per day on work days
- Code advices independently
 - If necessary, adjust codebook (If you encounter new themes when looking at details)
- Calculate IRR between you and your team member for your coding

Hand-in

- Each of you individually should hand in the following [E-learning system] folder: *Link*
 - A Zip-file named [your_id1]_coding.zip
 - e.g. for us: [naming]_coding.zip
 - Containing the following:
 - Coded advice: the filled out spreadsheet (subsample__[x].csv)
 - Your codebook file
 - IRR value in a txt file

B.2 Meta-level analysis outcomes

B.2.1 Interview Content Codebook

- Other ideas

- adoption process
- Relationships
 - Realism - trust
 - Effectiveness - realism
 - Effectiveness - trust
- Effectiveness
 - Plausibility
 - Protection
 - Threatening of protection
 - No protection
 - Not best possible protection / limited protection
 - Protection use case dependent
 - Protection improved
 - Protection (advice protects)
 - Technical
 - Effort
 - Cost-benefit considerations
 - Usability
 - User characteristics
- Realism
 - Feelings of security
 - Future-proof advice
 - Source
 - Protection improved
 - Popularity /coverage / media
 - Fear of consequences
 - Expert knowledge
 - Plausibility
 - Force / compulsion
 - Use case dependent
 - Raising Awareness
 - Habit
 - Cost-benefit considerations
 - User characteristics
 - Effort

- Usability
- Additional utility
- Trust
 - Technology
 - Relevance
 - Plausibility
 - Popularity / coverage / media
 - Social context
 - Expert knowledge
 - User characteristics
 - Source

B.2.2 Survey Content Codebook

When developing our codebook, we started with the codebooks of Busse et al. [74] and Ion et al. [236] and extended it with new codes. Original codes come in standard font while newly created codes come in *italic font*. Codes that were included in the codebooks of the prior work but have not been assigned in our study are listed below our codebook.

- New in student work
 - *Developer specific measures*
 - *Protect companies / critical infrastructure*
 - *Junk*
- Account security
 - *Store passwords safely*
 - *Don't use auto-password login*
 - *Terminate unused accounts / software*
 - *Use patterns*
 - *Use PINs*
 - *Use passwords*
 - *Use biometrics*
 - *Generate passwords*
 - *Don't share authentication*
 - *Always log out*
 - Change password
 - Don't write down passwords

- use a password manager
 - Use strong passwords
 - Use unique passwords
 - Use two-factor authentication
 - Write down passwords
- Mindfulness
 - *Keep devices safe*
 - *Don't share devices*
 - *Ask for help*
 - *Avoid phishing / scam*
 - *Hide data*
 - *Protect personal information*
 - *Use privacy settings*
 - *Clear browsing history*
 - *Close browser after use*
 - *Use throw-away accounts*
 - *Use fake identities*
 - *Selective sharing of information*
 - *Cookie specific precautions*
 - *Inform yourself*
 - *Be suspicious of files*
 - *Be suspicious of e-mail*
 - *Be suspicious of downloads*
 - Don't open email attachments
 - Don't click links from unknown people
 - Be careful when online
 - Be skeptical of everything
 - Be suspicious of links
 - Check if HTTPS
 - Clear cookies
 - Don't share info
 - Look at the URL bar
 - Visit only known or trusted websites
- Security software
 - *Use cloud storage*

- *Backups*
 - *Use different search engines*
 - Use a VPN
 - Use antivirus
 - Use a firewall
 - Use incognito browsing
 - Use script and/or ad blockers
 - Use security software in general
- Updates
 - *Update device*
 - Install latest OS updates
 - Update applications
 - Update system
- Technical
 - *Check logs / search for anomalies*
 - *Defragmentize hard drive*
 - *Use only your own system*
 - *Use encryption*
 - *Use only secure/not public WiFi*
 - Use verified software

Codes that were included in the codebooks of prior work [74, 236] but have not been assigned in our study:

- Save passwords in a file
- Don't enter passwords on links in e-mail
- Turn on automatic updates
- Compartmentalize systems for different tasks or levels of security
- Employ virtual machines
- Use linux

B.3 PC Member Survey

Demographics

- Do you consider yourself mostly a
 - qualitative researcher
 - quantitative researcher
 - both qualitative and quantitative researcher
- Approximately how often have you been a PC member for SOUPS?
 - 1-2
 - 3-4
 - 5+
- Do you currently work in ...
 - Academia
 - Industry
 - Other – please specify *Text-input field*

Quality Criteria for Qualitative Work

In general, how important are the following quality criteria for you when you are reviewing qualitative SOUPS papers?

- More than one researcher is involved in coding the data.
 - Very important
 - Somewhat important
 - Somewhat unimportant
 - Very unimportant
 - I don't know
 - It depends *Text-input field*
 - Not applicable
- The method used for coding and data analysis is identified clearly - by describing in detail how the coding and data analysis was done.
 - Very important
 - Somewhat important
 - Somewhat unimportant
 - Very unimportant

- I don't know
 - It depends *Text-input field*
 - Not applicable
- A method of reaching agreement about coding is described in the paper.
 - Very important
 - Somewhat important
 - Somewhat unimportant
 - Very unimportant
 - I don't know
 - It depends *Text-input field*
 - Not applicable
- Full agreement is reached by the end of the analysis process, i.e. all differences in coding are resolved.
 - Very important
 - Somewhat important
 - Somewhat unimportant
 - Very unimportant
 - I don't know
 - It depends *Text-input field*
 - Not applicable
- A numerical measure of interrater reliability (e.g. Cohen's Kappa) is reported.
 - Very important
 - Somewhat important
 - Somewhat unimportant
 - Very unimportant
 - I don't know
 - It depends *Text-input field*
 - Not applicable
- When reviewing a study where short, textual survey responses are coded, such as the security advice in the Ion et al. (2015) paper "No one Can Hack My Mind", which coding practices do you consider acceptable? Please select all that apply.
 - Two coders coding separately
 - Two coders coding jointly
 - Two coders coding a subset jointly and the rest on their own

- I don't mind
 - Other – (please specify) *Text-input field*
- When reviewing a study that analyzes complex answers such as from interviews, which coding practices do you consider acceptable? Please select all that apply.
 - Two coders coding separately
 - Two coders coding jointly
 - Two coders coding a subset jointly and the rest on their own
 - I don't mind
 - Other – (please specify) *Text-input field*
- What is the minimum final level of agreement that is acceptable to you? Categories are according to Landis & Koch (1977)
 - Poor (e.g. Kappa < 0.0)
 - Slight (e.g. Kappa 0.0 - 0.2)
 - Fair (e.g. Kappa 0.21 - 0.4)
 - Moderate (e.g. Kappa 0.41 - 0.6)
 - Substantial (e.g. Kappa 0.61 - 0.8)
 - Almost Perfect (e.g. Kappa > 0.8)
 - I don't require a specific agreement level - only that one is reported
 - I don't require any reporting on agreement
 - It depends - please specify *Text-input field*
- For which phases in the coding process should a numerical form of interrater reliability be reported? Please select all that apply.
 - After codebook establishment
 - After each separate step in the analysis process where data is coded independently
 - After coding is completed
 - I don't think a numerical form of interrater reliability needs to be reported
 - Other – please specify *Text-input field*
- Analysis of qualitative content is a complex process. What else is important to you when you are reviewing qualitative SOUPS submissions? (optional) *Text-input field*

B.4 PC Member Survey Codebook

- Coding phases
 - Iterative calculation

- After coding is completed
 - Importance of coding phases
 - It depends
 - When codebook is stable
- IRR concrete values
 - Moderate
 - Wants Kappa and specific value
 - I don't require a specific agreement level - only that one is r
 - I don't require any reporting on agreement
 - Substantial
 - Substantial / almost perfect
- Additional quality criteria
 - Researcher background
 - Exploration of alternative hypotheses
 - Quantified statements, pro and con
 - Depth of analysis / interpretation
 - Study design
 - Impact of results
 - Justification
 - Research goal
 - Examples for coded text snippets
 - Detailed description into how
 - Additional material
- Quality criteria reasoning
 - Method (qualitative analysis framework)
 - Not necessary, but not harmful either
 - Novelty of method
 - Not applicable
 - Generally good
 - Methods need to be justified
 - Exceptions possible
 - Trust in methods text is overrated
 - Type of study
 - Amount of data

- Complexity
- Codebook size / detail
- Data sources
- How coding results are used
- Type of analysis
 - Reproducibility
- Claims (results)

Appendix C

Power Analysis and Effect Sizes in Developer-Centered Usable Security

This appendix contains additional material related to the work presented in Chapter 4. A list of papers included in the database is in Section C.1 and a companion website to this paper is published at powerdb.info. The website includes a tutorial providing an overview of the data necessary to conduct power analysis for basic hypothesis tests, where to find this data in our database, and how to use it to conduct power analysis using G*Power or R. The database of analyzed papers, and analysis scripts used e.g. to generate figures in the paper are on OSF at osf.io/me3hv.

C.1 Included Papers

Acar et al. [5], Acar et al. [4], Acar et al. [7], Assal et al. [23], Baig et al. [33], Braz et al. [66], Cohnen et al. [94], Danilova et al. [109], Danilova et al. [106], Danilova et al. [108], Derr et al. [116], Dietrich et al. [119], Fahl et al. [143], Fischer et al. [157], Fischer et al. [156], Fulton et al. [164], Gerlitz et al. [176], Gorski et al. [184], Gorski et al. [183], Hänsch et al. [202], Hallett et al. [195], Haney et al. [201], Krombholz et al. [286], Krombholz et al. [285], van der Linden et al. [535], Mhaidli et al. [330], Nadi et al. [340], Naiakshina et al. [344], Naiakshina et al. [345], Naiakshina et al. [343], Naiakshina et al. [342], Nguyen et al. [351], Nosco et al. [353], Oliveira et al. [360], Oltrogge et al. [361], Palombo et al. [373], Pashchenko et al. [381], Plöger et al. [395], Roth et al. [436], Ruef et al. [438], Sheth et al. [467], Tahaei et al. [494], Tiefenau et al. [510], Tiefenau et al. [509], Tuladhar et al. [520], Votipka et al. [545], Votipka et al. [544], Votipka et al. [543], Weir et al. [563], Weir et al. [562], Yakdan et al. [581], Yong Wong et al. [584], Yskout et al. [586], Yskout et al. [587]

Appendix D

Power and Effect Size at CHI

This appendix contains additional material related to the work presented in Chapter 5. This includes the prompt used to extract statistics (see Section D.1) and Tables D.1 to D.9, which contain effect size guidelines for the remaining effect size measures used in the meta-study, η^2 , Cohen's d , OR, CLES, Kendall's τ , Spearman's ρ , Cramer's V , Cohen's f , and f^2 . We also provide these, as well as the guidelines presented in the main paper in CSV-format for increased ease of processing on Github:

<https://github.com/Behavioral-Security/A-meta-study-of-effect-sizes-at-CHI> and for the sake of ensuring availability as a backup version on OSF: <https://osf.io/7u643/>. This repository, and the backup also include data and analysis scripts to enable replication of the analyses we conducted.

D.1 Prompt

Our final prompt consists of the following parts:

System prompt incl. persona You are a statistics expert and want to conduct a meta-analysis based on the results of scientific papers. To do this, you need to extract important statistics from the tests reported in the scientific paper.

Carefully heed the user's instructions. Respond using Markdown.

Task You need to extract the number of participants in the study, i.e. the sample size. For tests in the null hypothesis statistical testing paradigm, you need the following information for each individual test:

- the name of the hypothesis test
- the p-value, which can take on numeric values between 0 and 1 and is often denoted as being smaller $<$ or larger $>$ than a given value
- the effect size, which is a different measure depending on the hypothesis test used, and consists of a measure, sometimes denoted by a Greek letter and the actual numeric effect size value

- confidence intervals around the effect size, which consist of two numeric values, one smaller than the effect size and one larger than the effect size. Confidence intervals can be different types, with 95% confidence intervals being the most common. They are frequently abbreviated CI
- the number of participants whose data was used in this test. This will often be same as the sample size in the study, but not always. E.g. in post-hoc tests for independent samples, only the participants in the conditions compared in the post-hoc tests are relevant for those post-hoc tests. Or when only participants who fulfill a certain condition are considered for an analysis.

Depending on the type of test, you need to extract additional information. Some tests will report multiple different effect sizes. You need to extract them all and if available, their confidence intervals. Some tests, like regression analyses, will report multiple p-values and corresponding effect sizes and confidence intervals, for each factor involved in the analysis. You need to collect the information for each factor separately, but it should still be identifiable to which test the factor belongs. Some tests, like analyses of variance (ANOVAs) will report results (i.e. p-values, corresponding effect sizes and confidence intervals) separately for one or more main effects and one or more interaction effects. You need to collect the information for each effect separately, but it should still be identifiable to which test the factor belongs.

A combination of these is also possible, so that a test can have a p-value, effect sizes and confidence intervals associated with the test in general, also called omnibus test, and additional p-values, effect sizes and confidence intervals associated with each individual factor and effect in the test.

Any of the numeric values, such as p-values, effect sizes or confidence intervals can be missing. Sometimes these values are not stated explicitly for each test, e.g. tests with the same result may be summarized e.g. as: "All other comparisons were not significant". In that case, it is necessary to identify how many tests were conducted in total, e.g. by comparing how many conditions there are in total and how many are not yet accounted for by the reporting. Then list each non-significant test separately. If the numeric values cannot be extrapolated from the context, they should be considered missing and you need to state explicitly that the value is missing.

Other values (like names of hypothesis tests) may also be missing, in which case determine the value from the context as best as you can. Test statistics or descriptions in the method section, regarding the used tests in the data analysis or the study set-up (e.g. to determine whether repeated measures analyses were likely used, or which types of variables were measured) can be helpful.

Some values are only reported in tables. If the table's title or a reference in the text suggests relevant information in the table request it via the tool!

If the paper reports tests in a different statistical paradigm, e.g. using Bayes statistics, you also need information for each individual test:

- the name of the hypothesis test
- all other statistics associated with this hypothesis test, including names and values of the statistics

Paper information The paper you have to analyze has the title [title]. Its abstract is the following:

“ [abstract] “

It has the following sections:

“ [sections] “

And the following tables:

“ [tables] “

Analyze the paper using your tools and make sure to report every test with its parameters. Remember to also report effect sizes that are not explicitly stated including but not limited to:

- correlation coefficients.
- Odds Ratio (OR)
- Risk Ratio (RR)
- Related Measures

After extracting the information for a test (test name, effect size, amount participants, ...), **use the ‘report’ tool immediately** to submit the test. Once the report is submitted, confirm with 'Report submitted' before proceeding to the next test. Do not move forward until you have completed the report.

After reviewing the sections, have a look at the relevant tables.

D.2 Effect Size Guidelines

We provide the effect size guidelines for the remaining effect sizes measures used in the meta-study: η^2 , Cohen's d, OR, CLES, Kendall's τ , Spearman's ρ , Cramer's V, Cohen's f and f^2 .

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	0.02	0.07	0.37	0.01	0.10	0.52
applied computing	0.02	0.11	0.46	0.02	0.19	0.56
collaborative and social computing	0.00	0.07	0.37	0.00	0.07	0.38
computing methodologies	0.00	0.04	0.49	0.00	0.10	0.93
empirical studies in hci	0.02	0.07	0.34	0.01	0.08	0.36
hci theory, concepts and models	0.02	0.10	0.52	0.01	0.36	0.97
human computer interaction (hci)	0.01	0.09	0.42	0.03	0.11	0.41
human-centered computing	0.02	0.07	0.37	0.01	0.10	0.55
interaction design	0.02	0.06	0.08	0.01	0.02	0.98
mixed / augmented reality	0.08	0.15	0.37	0.06	0.23	0.60
security and privacy	0.02	0.05	0.16	0.00	0.03	0.20
social and professional topics	0.02	0.04	0.42	0.04	0.21	0.64
virtual reality	0.07	0.10	0.25	0.07	0.28	0.53

Table D.1 – Guidelines for separate research areas, using η^2 , R^2 , ω^2 , η_p^2 , η_G^2 or related effect size measures

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	0.22	0.44	1.09	0.19	0.52	1.38
applied computing	0.20	0.54	1.26	0.25	0.74	1.46
collaborative and social computing	0.11	0.44	1.09	0.08	0.42	1.12
computing methodologies	0.11	0.32	1.32	0.07	0.52	2.33
empirical studies in hci	0.22	0.43	1.04	0.17	0.46	1.07
hci theory, concepts and models	0.22	0.51	1.38	0.18	1.07	2.46
human computer interaction (hci)	0.19	0.49	1.18	0.27	0.54	1.17
human-centered computing	0.22	0.44	1.09	0.19	0.52	1.44
interaction design	0.23	0.41	0.46	0.13	0.20	2.50
mixed / augmented reality	0.46	0.64	1.08	0.40	0.81	1.55
security and privacy	0.23	0.36	0.65	0.09	0.27	0.75
social and professional topics	0.21	0.34	1.18	0.34	0.77	1.61
virtual reality	0.42	0.52	0.86	0.42	0.92	1.40

Table D.2 – Guidelines for separate research areas, using Cohen's d

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	1.48	2.13	5.90	1.39	2.44	8.72
applied computing	1.42	2.50	7.46	1.54	3.47	9.63
collaborative and social computing	1.21	2.12	5.93	1.16	2.06	6.12
computing methodologies	1.21	1.74	8.10	1.13	2.44	23.93
empirical studies in hci	1.48	2.11	5.46	1.35	2.19	5.71
hci theory, concepts and models	1.46	2.38	8.68	1.37	5.71	26.48
human computer interaction (hci)	1.39	2.30	6.67	1.59	2.52	6.57
human-centered computing	1.46	2.13	5.93	1.39	2.44	9.39
interaction design	1.50	2.02	2.21	1.25	1.43	27.38
mixed / augmented reality	2.22	2.99	5.84	1.99	3.89	10.76
security and privacy	1.50	1.86	3.03	1.17	1.59	3.53
social and professional topics	1.43	1.79	6.67	1.78	3.67	11.63
virtual reality	2.07	2.46	4.17	2.06	4.55	8.94

Table D.3 – Guidelines for separate research areas, using OR

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	0.54	0.59	0.71	0.54	0.60	0.76
applied computing	0.54	0.61	0.74	0.55	0.64	0.77
collaborative and social computing	0.52	0.59	0.71	0.52	0.58	0.71
computing methodologies	0.52	0.56	0.75	0.51	0.60	0.92
empirical studies in hci	0.54	0.59	0.70	0.54	0.59	0.70
hci theory, concepts and models	0.54	0.60	0.76	0.54	0.70	0.94
human computer interaction (hci)	0.54	0.60	0.72	0.55	0.61	0.72
human-centered computing	0.54	0.59	0.71	0.54	0.60	0.77
interaction design	0.55	0.58	0.59	0.53	0.54	0.96
mixed / augmented reality	0.59	0.63	0.71	0.58	0.66	0.78
security and privacy	0.55	0.57	0.63	0.52	0.55	0.65
social and professional topics	0.54	0.57	0.72	0.57	0.65	0.79
virtual reality	0.58	0.60	0.67	0.58	0.68	0.76

Table D.4 – Guidelines for separate research areas, using CLES

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	0.09	0.17	0.41	0.08	0.20	0.51
applied computing	0.08	0.21	0.47	0.10	0.29	0.53
collaborative and social computing	0.04	0.17	0.41	0.03	0.17	0.42
computing methodologies	0.04	0.13	0.49	0.03	0.20	0.81
empirical studies in hci	0.09	0.17	0.39	0.07	0.18	0.40
hci theory, concepts and models	0.09	0.20	0.50	0.07	0.40	0.86
human computer interaction (hci)	0.08	0.19	0.44	0.11	0.21	0.44
human-centered computing	0.09	0.17	0.41	0.08	0.20	0.52
interaction design	0.09	0.16	0.18	0.05	0.08	0.88
mixed / augmented reality	0.18	0.25	0.41	0.16	0.31	0.56
security and privacy	0.09	0.14	0.26	0.04	0.11	0.29
social and professional topics	0.08	0.14	0.44	0.13	0.30	0.58
virtual reality	0.17	0.21	0.33	0.17	0.35	0.51

Table D.5 – Guidelines for separate research areas, using Kendall's τ

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	0.13	0.26	0.58	0.11	0.30	0.69
applied computing	0.12	0.31	0.65	0.15	0.41	0.72
collaborative and social computing	0.06	0.26	0.58	0.05	0.24	0.59
computing methodologies	0.06	0.19	0.67	0.04	0.30	0.94
empirical studies in hci	0.13	0.25	0.56	0.10	0.27	0.57
hci theory, concepts and models	0.13	0.29	0.69	0.11	0.57	0.96
human computer interaction (hci)	0.11	0.28	0.61	0.16	0.31	0.61
human-centered computing	0.13	0.26	0.58	0.11	0.30	0.71
interaction design	0.14	0.24	0.27	0.08	0.12	0.97
mixed / augmented reality	0.27	0.37	0.57	0.23	0.45	0.75
security and privacy	0.14	0.21	0.37	0.05	0.16	0.42
social and professional topics	0.12	0.20	0.61	0.20	0.43	0.77
virtual reality	0.25	0.30	0.47	0.25	0.50	0.70

Table D.6 – Guidelines for separate research areas, using Spearman's ρ

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	0.10	0.19	0.42	0.08	0.22	0.50
applied computing	0.09	0.23	0.47	0.11	0.31	0.52
collaborative and social computing	0.05	0.19	0.42	0.04	0.18	0.43
computing methodologies	0.05	0.14	0.49	0.03	0.22	0.67
empirical studies in hci	0.10	0.19	0.41	0.08	0.20	0.42
hci theory, concepts and models	0.10	0.22	0.50	0.08	0.42	0.68
human computer interaction (hci)	0.08	0.21	0.45	0.12	0.23	0.44
human-centered computing	0.10	0.19	0.42	0.08	0.22	0.52
interaction design	0.10	0.18	0.20	0.06	0.09	0.69
mixed / augmented reality	0.20	0.27	0.42	0.17	0.33	0.54
security and privacy	0.10	0.16	0.27	0.04	0.12	0.31
social and professional topics	0.09	0.15	0.45	0.15	0.32	0.55
virtual reality	0.18	0.22	0.35	0.18	0.37	0.51

Table D.7 – Guidelines for separate research areas, using Cramer's V

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	0.11	0.22	0.55	0.10	0.26	0.69
applied computing	0.10	0.27	0.63	0.12	0.37	0.73
collaborative and social computing	0.06	0.22	0.55	0.04	0.21	0.56
computing methodologies	0.06	0.16	0.66	0.04	0.26	1.17
empirical studies in hci	0.11	0.22	0.52	0.09	0.23	0.53
hci theory, concepts and models	0.11	0.25	0.69	0.09	0.53	1.23
human computer interaction (hci)	0.10	0.24	0.59	0.13	0.27	0.58
human-centered computing	0.11	0.22	0.55	0.10	0.26	0.72
interaction design	0.12	0.20	0.23	0.06	0.10	1.25
mixed / augmented reality	0.23	0.32	0.54	0.20	0.41	0.78
security and privacy	0.12	0.18	0.33	0.04	0.13	0.38
social and professional topics	0.10	0.17	0.59	0.17	0.39	0.81
virtual reality	0.21	0.26	0.43	0.21	0.46	0.70

Table D.8 – Guidelines for separate research areas, using Cohen's f

research area	between groups			within groups		
	small	medium	large	small	medium	large
all	0.02	0.08	0.59	0.01	0.11	1.08
applied computing	0.02	0.12	0.85	0.02	0.24	1.27
collaborative and social computing	0.00	0.08	0.59	0.00	0.07	0.62
computing methodologies	0.00	0.04	0.96	0.00	0.11	13.29
empirical studies in hci	0.02	0.08	0.52	0.01	0.09	0.56
hci theory, concepts and models	0.02	0.11	1.08	0.01	0.56	32.33
human computer interaction (hci)	0.01	0.10	0.71	0.03	0.12	0.69
human-centered computing	0.02	0.08	0.59	0.01	0.11	1.22
interaction design	0.02	0.07	0.09	0.01	0.02	57.82
mixed / augmented reality	0.09	0.18	0.58	0.07	0.30	1.53
security and privacy	0.02	0.05	0.18	0.00	0.03	0.25
social and professional topics	0.02	0.05	0.71	0.05	0.27	1.75
virtual reality	0.07	0.12	0.33	0.07	0.39	1.13

Table D.9 – Guidelines for separate research areas, using f^2

Appendix E

Interpretation of Effect Sizes

This appendix contains additional material related to the work presented in Chapter 6. It contains a list of papers which we used to investigate how effect sizes are interpreted in a sample of publications at CHI in Section E.1. It also includes information related to the data collection in our studies with researchers. In particular, we provide the vignettes shown to the participants in the survey and interview studies, including all variations to them, in Section E.2, the interview guidelines for study 1 (S-CHI) and study 2 (S-SOUPS) in Section E.3 and the survey questions used in S-CHI and S-SOUPS in Section E.4. Additional material to enable repetition of the analyses conducted on the data collected in the studies with researchers is on OSF: <https://osf.io/qy2u9/>. This includes analysis scripts and anonymized data.

E.1 CHI'23 Publications Used for Investigation of Effect Size Interpretation Strategies

NHST-quantitative papers

Beckmann et al. [44], Bovo et al. [61], Clarke et al. [87], Courtoux et al. [97], Cui et al. [103], Dang et al. [104], Faleel et al. [146], Fashimpaur et al. [147], Fink et al. [155], Gu et al. [190], Han et al. [199], He et al. [211], Huh et al. [227], Hutton et al. [230], Iftikhar et al. [231], Jicol et al. [244], Jo et al. [245], Kowalczyk et al. [282], Kumaran et al. [289], Lang et al. [297], Lee et al. [300], Li et al. [304], Lisnic et al. [309], Liu et al. [310], Liu et al. [312], Liu et al. [313], Liu et al. [314], Nagassa et al. [341], Nguyen et al. [350], Molina et al. [332], Müller et al. [336], Petridis et al. [387], Press et al. [402], Qin et al. [406], Rakhmetulla et al. [412], Rekimoto et al. [422], Risha et al. [425], Rocha et al. [427], Salagean et al. [441], Schlagowski et al. [453], Shen et al. [466], Song et al. [478], Song et al. [477], Tanprasert et al. [502], Tran et al. [513], Tong et al. [511], Turk et al. [521], Uhl et al. [525], Vo et al. [540], Wang et al. [552], Wang et al. [550], Wang et al. [549], Würsching et al. [578], Xu et al. [580], Yilma et al. [583], Zhang et al. [589], Zhang-Kennedy et al. [591]

Method papers

Ahn et al. [10], Borgert et al. [56], Krath et al. [283], Phinnemore et al. [392], SalehzadehNiksirat et al. [442], Tao et al. [503]

Bayesian papers

Xiao et al. [579]

Primarily Qualitative Papers

Jacobsen et al. [239], Li et al. [303], Schade et al. [446]

E.2 Vignettes

This section contains the different iterations and variations of the vignettes used in the interview and survey studies. We provide one full example of a vignette as the participants would have seen it in Section E.2.1. Otherwise, we adhere to variable and variable names introduced in the paper. We first present the text introducing the **research question and study design** for each different context. For the warning adherence context, we adjusted our explanation and definition from [11]. The text for the **descriptive statistics** and **inferential statistics** was similar across all three contexts, so we only present it once for each effect size measure, and highlight, where there would be differences in phrasing for different contexts. Finally, we include the effect size explanation for OR used in S-CHI.

E.2.1 Vignette from Participants' Perspective

The following vignette is an example of the consistency checker context, presented with a conventionally large odds ratio as participants in S-SOUPS would have seen it (there is no effect size explanation included in the vignette).

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. 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. 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.

E.2.2 Research Question and Study Design

E.2.2.1 Password Manager Context

A study investigated the difference between the adoption of a **password manager**, in a baseline group that received a general introduction to the password manager and an intervention group that additionally was informed that **using a password manager was the top recommendation made by security experts**.

E.2.2.2 Consistency Checker Context

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**.

E.2.2.3 Warning Adherence Context

A study investigated the efficacy of a new browser security warning for self-signed certificates. If a browser connects to a domain using a self-signed certificate, it is very hard for the user to tell whether they are connecting to the correct site or whether an attacker is intercepting their traffic. While the vast majority of users are unlikely to be attacked, political dissidents, whistle-blowers, journalists, or other at-risk groups have been targeted by this kind of attack in the past, with dire consequences.

E.2.3 Descriptive Statistics and Inferential Statistics

Bold text refers to variables, which take on different values depending on the context (prefixed with c-) or the conventional effect size (pre-fixed with e-). Both of these also vary for different effect size measures. Table E.1 presents the different phrasing variations for different contexts, and Table E.2 presents the values for different conventional effect sizes.

E.2.3.1 OR Effect Size Measure

c-study-design

e-amount-intervention out of 1050 (**e-percent-intervention** %) of those who **c-descriptives-intervention**. **e-amount-baseline** out of 1050 (**e-percent-baseline** %) of those who **c-descriptives-control**. Fisher's exact test showed that this difference was statistically significant ($p < 0.001$, odds ratio [OR]=**e-OR-value**, 95% CI=[**e-CI-values**]). The effect size (Odds ratio) is **e-OR-value**.

E.2.3.2 Cohen's d Effect Size Measure

c-study-design

c-descriptives-intervention was **e-mean-intervention**(SD=**e-sd-intervention**). **c-descriptives-control** was **e-mean-control**(SD=**e-sd-control**). A t-test for independent samples showed that this difference was statistically significant, $t(\mathbf{e-df})=\mathbf{e-t}$, $p < 0.001$. The effect size (Cohen's d) was **d=e-d-value**.

c-context	effect size measure	c-study-design	c-descriptives-control	c-descriptives-intervention
password manager	OR	2100 Participants used the c-context during the study. Two weeks after the end of the study, they were asked whether they were still using the c-context or not.	did not see the intervention were still using it	received the intervention were still using it
consistency checker	OR	see above	see above	see above
warning adherence	OR	In the study with 2100 participants, a baseline group of participants was shown the standard browser warning. In the intervention group, participants were shown a newly designed scarier warning with the aim of increasing warning adherence.	saw the baseline warning adhered to the warning	saw the newly designed warning adhered to the warning
password manager	Cohen's d	2100 participants used the c-context during the study. After the study, they filled out a survey, including the System Usability Scale, which provides a subjective measure of usability, and has values between 0 and 100.	The mean SUS score in the control group	The mean SUS score in the experimental group
consistency checker	Cohen's d	see above	see above	see above
warning adherence	Cohen's d	In the study with 2100 participants, a baseline group of participants was shown the standard browser warning. In the intervention group, participants were shown a newly designed scarier version with the aim of increasing warning adherence. We measured how much time the users spent on the warning before making a decision.	The mean amount of time spent on the baseline warning ... seconds	The mean amount of time spent on the new, scarier warning ... seconds

Table E.1 – Phrasing variations depending on context and effect size measure, prefixed with c- in the vignettes

effect size	conventional effect size	e-amount-intervention	e-percent-intervention	e-amount-control	e-percent-control	e-OR-value	e-CI-values
measure:	small	294	28.0	224	21.3	1.43	[1.17, 1.76]
OR	medium	399	38.0	224	21.3	2.26	[1.85, 2.76]
	large	511	48.7	224	21.3	3.50	[2.88, 4.25]

effect size	conventional effect size	e-mean-intervention	e-sd-intervention	e-mean-control	e-sd-control	e-t	e-df	e-d-value
measure:	small	63.7	12.6	60.7	12.3	-5.52	2098	0.24
Cohen's d	medium	67.2	12.5	60.6	12.3	-12.2	2098	0.53
	large	71.5	12.7	60.8	12.4	-19.6	2098	0.85

Table E.2 – Numerical variations depending on the effect size measure and conventional effect size, prefixed with e- in the vignettes

E.2.4 Effect Size Explanation

This explanation was only included in S-CHI, and there is no explanation for Cohen's d.

This means that the odds for the participants to continue to use the **c-context** in the intervention group were **e-OR-value** times higher than for participants to continue to use the **c-context** in the baseline group.

E.3 Interview guidelines

Bold text varies depending on the conventional effect size, context or effect size measure present in the vignette.

E.3.1 S-CHI-i

Demographics are collected using the demographics questions from S-CHI-s in E.4.1.2. The questions were displayed on a tablet and the interviewee selected the answers. We asked follow-up questions when participants changed their answers or hesitated. These questions are written in italics and were answered only verbally.

[Present first vignette with the warning adherence context]

Interviewer: "Please say out loud what you are thinking as you look at the example. We are interested in your thought process."

- How severe do you think it is, if a user ignores the warning presented in the scenario above, and clicks through it?
[Not severe at all, Slightly severe, Moderately severe, Very severe, Extremely severe]
- In the study, the effect size was an odds ratio of **e-OR-value**. How would you rate this effect considering the information you received above? *[Small, Medium, Large]*

- How sure are you about your answer to the previous question?
[I am certain, Very sure, somewhat sure, somewhat unsure, Very unsure, It was a guess]
- How would you rate the importance of this effect?
[unimportant, important]
 - *Why do you think this effect is [prior judgment]?*
 - *Which information helped you make this decision?*
 - *From which source is this information?*

[Explanation of OR]

- How did your understanding of OR change after seeing this explanation?
- In the context of this study, what is the lowest odds ratio which you would still consider important? *[freetext]*

[Present second vignette with the password manager or the consistency checker context]

Interviewer: “When answering questions about this context, please compare it to the example before. Again, please say out loud what you are thinking as you look at the example. We are interested in your thought process.”

- How severe do you think it is, if a user does not adopt the software presented in the scenario above?
[Not severe at all, Slightly severe, Moderately severe, Very severe, Extremely severe]
- In the study, the effect size was an odds ratio of **e-OR-value**. How would you rate this effect considering the information you received above? *[Small, Medium, Large]*
- How sure are you about your answer to the previous question?
[I am certain, Very sure, somewhat sure, somewhat unsure, Very unsure, It was a guess]
- How would you rate the importance of this effect?
[unimportant, important]
 - *Why do you think this effect is less/more important than the effect in the prior example?*

[Explanation of OR]

- In the context of this study, what is the lowest odds ratio which you would still consider important? *[freetext]*

E.3.2 S-SOUPS-i

Demographics are collected using the demographics questions from S-SOUPS-s in E.4.2.2, and participants answered these in private on a tablet. All non-demographic questions were directly asked by the interviewers and answered only verbally.

[Present first vignette with the warning adherence context]

[Allow time for reading, answer questions if they arise]

1. Please explain the effect size used in this scenario, a Cohen's d / OR of [e-d-value/e-OR-value]?

- What does it mean?
- How do you understand the given number?
- How would “no effect” be shown in this effect size? What number would this be?
- What is the maximum possible value for this effect size?
- Have you used it before (in your own work)?
 - If yes: How did you calculate it? / Which tool did you use to calculate it?
 - If yes: Where did you learn about this effect size?

Phrasing help:

- P: I would first do research...
I: Where / How would you look for information on effect sizes?
- P: I need more information...
I: Which information exactly do you need?
- P: I don't have any idea about this effect size.
I: Show explanation and then ask questions about understanding afterwards.
- P says something about participant number
I: What is the relationship between participant number and effect size?

[show explanation]

- How does this information change your understanding of the effect size?
- Can you explain the effect size in your own words?

2. Do you think the effect in this scenario is important?

- Why / Why not? Please explain how you came to this conclusion.

Phrasing help:

- P: I don't have enough information to decide.
I: What components/factors described in the scenario are missing for you to be able to make an informed decision?

3. What would be a topic / research question in the domain that you work in / do research in, where it would be important, even to find a small effect?

- Describe this research question, please.
- What makes an effect there important?

[Present second vignette with either the password manager or the consistency checker context]

Repeat questions 1 and 2.

4. Now that we have discussed another context, would you amend your description of the most important effect in your field?

- How?

E.4 Surveys

Text in quotes varies depending on the conventional effect size, context or effect size measure present in the vignette.

E.4.1 S-CHI-s

After the participant has consented to participation, first vignette related questions are asked, then demographics. Next, the participant is asked whether they want to stop or continue:

- If you want to end the survey now, please select end survey, if you want to continue with another scenario please select continue. If you continue you have an additional chance in the raffle for 100 €. [*End survey / Continue*]

If they want to end the survey, they are able to provide feedback:

- Thank you for taking the time to take part in our study! If you feel anything was unclear during this survey or you have other feedback, please let us know below: [*free text*]

Otherwise they answer questions on a second vignette and after that may provide feedback.

E.4.1.1 Vignette-related

[vignette with either the password manager or the consistency checker]

- How severe do you think it is, if the technology is not adopted?
[*Not severe at all, Slightly severe, Moderately severe, Very severe, Extremely severe*]

- In the study, the effect size was an odds ratio of “e-OR-value”. How would you rate this effect considering the information you received above?
[Small, Medium, Large]
- How sure are you about your answer to the previous question?
[I am certain, Very sure, somewhat sure, somewhat unsure, Very unsure, It was a guess]
- How would you rate the importance of this effect?
[unimportant, important]

[Explanation of OR]

- For this study, what is be the **lowest odds ratio** you would still consider a **large** effect? I.e. what is the **lower limit** for an effect size to be considered **large** in the context of this study? Please enter your answer below. *[freetext]*
- How sure are you about your answer to the previous question?
[I am certain, Very sure, somewhat sure, somewhat unsure, Very unsure, It was a guess]
- For this study, what is be the **highest odds ratio** you would still consider a **small** effect? I.e. what is the **upper limit** for an effect size to be considered **small** in the context of this study? Please enter your answer below. *[freetext]*
- How sure are you about your answer to the previous question?
[I am certain, Very sure, somewhat sure, somewhat unsure, Very unsure, It was a guess]
- In the context of this study, what is the lowest odds ratio which you would still consider important? *[freetext]*

E.4.1.2 Demographics

- Have you previously taken part in our interview on the same topic?
[Yes; I am not sure; No]
- Please select your gender:
[Man; Woman; Non-binary; Prefer not to say; Prefer to self-describe: [freetext]]
- What is your age (in years)?
[Enter your age please [freetext], Prefer not to say]
- What is the highest level of education which you have completed?
[Elementary school; High School (or equivalent); Technical; trade school, vocational training, or apprenticeship; Undergraduate degree (Bachelor’s); Master’s degree; Doctoral degree; Prefer not to say]
- What is your current occupation?
[Undergraduate student (Bachelor’s student or equivalent); Undergraduate student (Master’s student or equivalent); Phd student; Post-doctoral researcher; Adjunct / Assistant / Junior Professor; Full professor; Industry Researcher; Other, please specify [freetext]]

- Do you consider yourself mostly a
[qualitative researcher; quantitative researcher; both qualitative and quantitative researcher; other, please specify: [freetext]]
- What is the focus of your research? (CHI'23 subcommittees - multiple answers possible)
[Accessibility and Aging; Blending Interaction: Engineering Interactive Systems & Tools; Building Devices: Hardware, Materials, and Fabrication; Computational Interaction; Critical Computing, Sustainability, and Social Justice; Design; Games and Play; Health; Interacting with Devices: Interaction Techniques & Modalities; Interaction Beyond the Individual; Learning, Education, and Families; Privacy and Security; Specific Applications Areas; Understanding People - Quantitative Methods; Understanding People - Qualitative Methods; Understanding People - Mixed and Alternative Methods; User Experience and Usability; Visualization; Other, please specify [freetext]]
- Which of the following effect sizes have you used in your work?
[Cohen's d; Odds Ratio (OR); Other, please specify [freetext]]

E.4.2 S-SOUPS-s

The flow of this survey is the same as in S-CHI-s (see E.4.1)

E.4.2.1 Vignette-related

[vignette with either the password manager or the consistency checker]

- How beneficial do you think the “usability improvement for/the adoption of” the software above is, to a user of the software?
[Not beneficial at all, Slightly beneficial, Moderately beneficial, Very beneficial, Extremely beneficial]
- In the study, the effect size was an “effect size measure” of “effect size value”. How would you rate this effect considering the information you received above?
[Small, Medium, Large]
- How sure are you about your answer to the previous question?
[I am certain, Very sure, somewhat sure, somewhat unsure, Very unsure, It was a guess]
- How would you rate the importance of this effect?
[unimportant, important]

[Explanation of “effect size measure”]

- For this study, what is be the **lowest “effect size measure”** you would still consider a **large** effect? I.e. what is the **lower limit** for an effect size to be considered **large** in the context of this study? Please enter your answer below. *[freetext]*

- How sure are you about your answer to the previous question?
[I am certain, Very sure, somewhat sure, somewhat unsure, Very unsure, It was a guess]
- For this study, what is be the **highest “effect size measure”** you would still consider a **small** effect? I.e. what is the **upper limit** for an effect size to be considered **small** in the context of this study? Please enter your answer below. *[freetext]*
- How sure are you about your answer to the previous question?
[I am certain, Very sure, somewhat sure, somewhat unsure, Very unsure, It was a guess]
- In the context of this study, what is the lowest odds ratio which you would still consider important? *[freetext]*

E.4.2.2 Demographics

We only include questions that are not the same as the demographic questions for S-CHI-s (Section E.4.1.2).

- What is the topical focus of your research (Please provide 1-5 keywords) *[freetext]*
replaces What is the focus of your research? (CHI’23 subcommittees - multiple answers possible)

Appendix F

A Re-replication Study About Experts' and Non-experts' Security Practices and Advice

This appendix contains additional material related to the work presented in Chapter 7. It contains the survey questions shown to the different groups of participants in Section F.1. Additional material related to data collection and data analysis is on OSF: <https://osf.io/j43ws/>. This includes the interview guidelines and graphs used as prompts in the interviews, as well as the codebook used to code security practices and advice, and anonymized data sets with analysis scripts.

F.1 Survey Material

All multiple-choice questions were single answer only. The questions were identical for the Expert, Expert (Freelancer only), and Non-expert survey, unless otherwise stated. The questions marked "(Experts only)", "(Expert Freelancers only)" or "(Non-experts only)" were asked in only one of the surveys.

- (*Experts only*) What are the top 3 pieces of advice you would give to a non-tech-savvy user to protect their security online? (*open-ended*)
- What are the 3 most important things you do to protect your security online? (*open-ended*)
- How did you learn about the things you listed above? (*open-ended*)
- Do you use a laptop or desktop computer that you or your family owns (i.e., not provided by school or work)? (*multiple-choice*)
 - Yes
 - No

- The following question apply to the laptop or desktop computer that you or your family owns. If you have more than one, please answer for the one you use most often for personal use.
- When did you get that computer? (*multiple-choice*)
 - Less than 1 year ago
 - At least 1 but less than 2 years ago
 - At least 2 but less than 3 years ago
 - At least 3 but less than 5 years ago
 - 5 or more years ago
 - I don't know
- How soon after you discover that a new version of your operating system (OS) software is available do you (or somebody else managing your computer) install it? (*multiple-choice*)
 - OS updates are installed automatically
 - Immediately
 - Soon after
 - Eventually
 - OS updates are never installed
 - Other (*open-ended*)
- Do you use anti-virus software on that computer? (*multiple-choice*)
 - Yes
 - No
 - I don't know
 - Other (*open-ended*)
- Which anti-virus software do you use? (*open-ended*)
- How do you keep track of your passwords for your online accounts? (*grid question*)
Answer options: For ALL of my accounts, For MOST of my accounts, For SOME of my accounts, For NONE of my accounts
 - Remember them
 - Write them down on paper
 - Save them in a local file on my computer
 - Have my password manager (e.g., 1Password, LastPass) remember them
 - Use the same password on multiple accounts
- If you use a password manager, which one do you use? (*open-ended*)
- (optional) What other things, if any, do you do to keep track of your passwords? (*open-ended*)

-
- Do you use two-factor authentication (e.g., 2-Step Verification) for at least one of your online accounts? (*multiple-choice*)
 - Yes
 - No
 - I don't know
 - Other (*open-ended*)
 - Do you use a hardware token (e.g. Yubikey) for authentication with at least one of your online accounts?
 - Yes
 - No
 - I don't know
 - Other (*open-ended*)
 - Do you use an authenticator app (e.g. Google Authenticator, Banking Authenticator app) for at least one of your online accounts?
 - Yes
 - No
 - I don't know
 - Other (*open-ended*)
 - Do you use passkeys (e.g. Windows Hello) for at least one of your online accounts?
 - Yes
 - No
 - I don't know
 - Other (*open-ended*)
 - Do you use biometrics-based authentication (e.g. FaceID, fingerprint scanning) for at least one of your online accounts?
 - Yes
 - No
 - I don't know
 - Other (*open-ended*)
 - Do you check permissions of the apps you use?
 - Yes, often
 - Yes, sometimes
 - Yes, rarely
 - No

- I don't know
 - Other (*open-ended*)
- Do you look at the URL bar to verify that you are visiting the website you intended to? (*multiple-choice*)
 - Yes, often
 - Yes, sometimes
 - Yes, rarely
 - No
 - I don't know
 - Other (*open-ended*)
- Google began in January 1996 as a research project. Its initial public offering took place on August 19, 2004. Did the initial public offering of Google take place in 1996? (*multiple-choice*)
 - Yes
 - No
 - Other (*open-ended*)
- Do you check if the website you're visiting uses HTTPS? (*multiple-choice*)
 - Yes, often
 - Yes, sometimes
 - Yes, rarely
 - No
 - I don't know
 - Other (*open-ended*)
- Do you visit websites you have not heard of before? (*multiple-choice*)
 - Yes, often
 - Yes, sometimes
 - Yes, rarely
 - No
 - I don't know
 - Other (*open-ended*)
- When you click on a link in an email and that link takes you to a website that asks for your password, do you enter it? (*multiple-choice*)
 - Yes, often
 - Yes, sometimes

- Yes, rarely
- No
- I don't know
- Other (*open-ended*)

Do you open emails you receive from people or companies you don't know? (*multiple-choice*)

- Yes, often
- Yes, sometimes
- Yes, rarely
- No
- I don't know
- Other (*open-ended*)

- Do you click on links that people or companies you don't know send you? (*multiple-choice*)
 - Yes, often
 - Yes, sometimes
 - Yes, rarely
 - No
 - I don't know
 - Other (*open-ended*)
- Do you share unnecessary personal information online, e.g. on social media?
 - Yes, often
 - Yes, sometimes
 - Yes, rarely
 - No
 - I don't know
 - Other (*open-ended*)
- (*Expert Freelancers only*) We are interested in your personal knowledge in the following questions. Please do not use Google, ChatGPT or other outside resources to answer this survey. If you don't know the answer to a question, answer "I don't know".
- (*Expert Freelancers only*) For each of the following tasks in asymmetric encryption, select the appropriate key to use to securely complete this task. (*grid question*)
Answer options: private key, public key, I don't know
 - Encryption of messages
 - Decryption of messages
 - Signing

- Verifying a signature
- (*Expert Freelancers only*) What does TLS stand for? (*multiple-choice*)
 - Transmission Level Security
 - Trusted Link Security
 - Transport Level Security
 - Transport Layer Security
 - Trusted Layer Security
 - I don't know
- (*Expert Freelancers only*) Which type of vulnerability is generally considered the most dangerous? (*multiple-choice*)
 - remote code execution
 - integer overflow
 - integer underflow
 - heap overflow
 - use-after-free
 - double free
 - race conditions
 - I don't know
- (*Experts only*) For each of the following pieces of advice, please rate on a scale from 1 to 5 how good (in terms of both EFFECTIVE at keeping the user secure, as well as REALISTIC that the user can follow it) you think they are at protecting a non-tech-savvy user's security online. (*grid question*)
 Scale: 5 (Very good), 4, 3, 2, 1 (Not at all), I don't know
 - Use anti-virus software
 - Install the latest operating system updates
 - Turn on automatic software updates
 - Update applications to the latest version
 - Clear your Web browser cookies
 - Check the permissions of the apps you use.
- (*Experts only*) (optional) Please use this space to clarify any of the above. (*open-ended*)
- (*Experts only*) For each of the following pieces of advice, please rate on a scale from 1 to 5 how good (in terms of both EFFECTIVE at keeping the user secure, as well as REALISTIC that the user can follow it) you think they are at protecting a non-tech-savvy user's security online. (*grid question*)
 Scale: 5 (Very good), 4, 3, 2, 1 (Not at all), I don't know

- Use different passwords for each account
 - Use passwords that are not easy to guess
 - Don't write down passwords on paper
 - Save your passwords in a local file on their computer
 - Use a password manager (e.g., 1Password, LastPass)
 - Write down passwords on paper
- (*Experts only*) (optional) Please use this space to clarify any of the above. (*open-ended*)
 - (*Experts only*) For each of the following pieces of advice, please rate on a scale from 1 to 5 how good (in terms of both EFFECTIVE at keeping the user secure, as well as REALISTIC that the user can follow it) you think they are at protecting a non-tech-savvy user's security online. (*grid question*)
Scale: 5 (Very good), 4, 3, 2, 1 (Not at all), I don't know
 - Check if the website you're visiting uses HTTPS
 - Be skeptical of everything when online
 - Be suspicious of links received in emails or messages
 - Visit only websites you've heard of
 - Use two-factor authentication for your online accounts
 - Don't share unnecessary personal information online, e.g. on social media
 - (*Experts only*) (optional) Please use this space to clarify any of the above. (*open-ended*)
 - (*Experts only*) For each of the following pieces of advice, please rate on a scale from 1 to 5 how good (in terms of both EFFECTIVE at keeping the user secure, as well as REALISTIC that the user can follow it) you think they are at protecting a non-tech-savvy user's security online. (*grid question*)
Scale: 5 (Very good), 4, 3, 2, 1 (Not at all), I don't know
 - Use a hardware token (e.g. a Yubikey) for authenticating your online accounts.
 - Use an authenticator app (e.g. Google Authenticator, banking authenticator app) for your online accounts.
 - Use passkeys (e.g. Windows Hello) for your online accounts.
 - Use biometrics (e.g. FaceID, fingerprint scanning) to log in to your online accounts.
 - (*Experts only*) (optional) Please use this space to clarify any of the above. (*open-ended*)
 - (*Experts only*) For each of the following pieces of advice, please rate on a scale from 1 to 5 how good (in terms of both EFFECTIVE at keeping the user secure, as well as REALISTIC that the user can follow it) you think they are at protecting a non-tech-savvy user's security online. (*grid question*)
Scale: 5 (Very good), 4, 3, 2, 1 (Not at all), I don't know
 - Don't click on links that people or companies you don't know send you

- Don't enter your password when you click on a link in an email and that link takes you to a website that asks for your password
- Pay attention when taking online surveys. We appreciate your input. To let us know you're paying attention, select four for this response
- Look at the URL bar to verify that you are visiting the website you intended to
- Don't open email attachments from people or companies you don't know
- Use AI (e.g. ChatGPT) to get security advice
- *(Experts only)* (optional) Please use this space to clarify any of the above. *(open-ended)*
- *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how EFFECTIVE you think the advice would be at protecting your security online, IF YOU FOLLOWED IT. *(grid question)*
Scale: 5 (Very good), 4, 3, 2, 1 (Not at all), I don't know
 - Use anti-virus software
 - Install the latest operating system updates
 - Turn on automatic software updates
 - Update applications to the latest version
 - Clear your Web browser cookies
 - Check the permissions of the apps you use
- *(Non-experts only)*(optional) Please use this space to clarify any of the above. *(open-ended)*
- *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how LIKELY YOU WOULD BE TO FOLLOW the advice, if you heard it would help protect your security online. *(grid question)*
Scale: 5 (Very good), 4, 3, 2, 1 (Not at all), I don't know
 - Use anti-virus software
 - Install the latest operating system updates
 - Turn on automatic software updates
 - Update applications to the latest version
 - Clear your Web browser cookies
 - Check the permissions of the apps you use
- *(Non-experts only)* (optional) Please use this space to clarify any of the above. *(open-ended)*
- *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how EFFECTIVE you think the advice would be at protecting your security online, IF YOU FOLLOWED IT. *(grid question)*
Scale: 5 (Very good), 4, 3, 2, 1 (Not at all), I don't know
 - Use different passwords for each account
 - Use passwords that are not easy to guess

- Don't write down passwords on paper
 - Save your passwords in a local file on their computer
 - Use a password manager (e.g., 1Password, LastPass)
 - Write down passwords on paper
- *(Non-experts only)*(optional) Please use this space to clarify any of the above. *(open-ended)*
 - *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how **LIKELY YOU WOULD BE TO FOLLOW** the advice, if you heard it would help protect your security online. *(grid question)*
Scale: 5 (Very likely), 4, 3, 2, 1 (Not at all), I don't know
 - Use different passwords for each account
 - Use passwords that are not easy to guess
 - Don't write down passwords on paper
 - Save your passwords in a local file on their computer
 - Use a password manager (e.g., 1Password, LastPass)
 - Write down passwords on paper
 - *(Non-experts only)* (optional) Please use this space to clarify any of the above. *(open-ended)*
 - *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how **EFFECTIVE** you think the advice would be at protecting your security online, **IF YOU FOLLOWED IT**. *(grid question)*
Scale: 5 (Very effective), 4, 3, 2, 1 (Not at all), I don't know
 - Check if the website you're visiting uses HTTPS
 - Be skeptical of everything when online
 - Be suspicious of links received in emails or messages
 - Visit only websites you've heard of
 - Use two-factor authentication for your online accounts
 - Don't share unnecessary personal information online, e.g. on social media.
 - *(Non-experts only)* (optional) Please use this space to clarify any of the above. *(open-ended)*
 - *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how **LIKELY YOU WOULD BE TO FOLLOW** the advice, if you heard it would help protect your security online. *(grid question)*
Scale: 5 (Very likely), 4, 3, 2, 1 (Not at all), I don't know
 - Check if the website you're visiting uses HTTPS
 - Be skeptical of everything when online
 - Be suspicious of links received in emails or messages
 - Visit only websites you've heard of

- Use two-factor authentication for your online accounts
- Don't share unnecessary personal information online, e.g. on social media.
- *(Non-experts only)* (optional) Please use this space to clarify any of the above. *(open-ended)*
- *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how EFFECTIVE you think the advice would be at protecting your security online, IF YOU FOLLOWED IT. *(grid question)*
Scale: 5 (Very effective), 4, 3, 2, 1 (Not at all), I don't know
 - Use a hardware token (e.g. a Yubikey) for authenticating your online accounts.
 - Use an authenticator app (e.g. Google Authenticator, banking authenticator app) for your online accounts.
 - Use passkeys (e.g. Windows Hello) for your online accounts.
 - Use biometrics (e.g. FaceID, fingerprint scanning) to log in to your online accounts.
- *(Non-experts only)* (optional) Please use this space to clarify any of the above. *(open-ended)*
- *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how LIKELY YOU WOULD BE TO FOLLOW the advice, if you heard it would help protect your security online. *(grid question)*
Scale: 5 (Very likely), 4, 3, 2, 1 (Not at all), I don't know
 - Use a hardware token (e.g. a Yubikey) for authenticating your online accounts.
 - Use an authenticator app (e.g. Google Authenticator, banking authenticator app) for your online accounts.
 - Use passkeys (e.g. Windows Hello) for your online accounts.
 - Use biometrics (e.g. FaceID, fingerprint scanning) to log in to your online accounts.
- *(Non-experts only)* (optional) Please use this space to clarify any of the above. *(open-ended)*
- *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how EFFECTIVE you think the advice would be at protecting your security online, IF YOU FOLLOWED IT. *(grid question)*
Scale: 5 (Very effective), 4, 3, 2, 1 (Not at all), I don't know
 - Don't click on links that people or companies you don't know send you
 - Don't enter your password when you click on a link in an email and that link takes you to a website that asks for your password
 - Pay attention when taking online surveys. We appreciate your input. To let us know you're paying attention, select four for this response
 - Look at the URL bar to verify that you are visiting the website you intended to
 - Don't open email attachments from people or companies you don't know
 - Use AI (e.g. ChatGPT) to get security advice
- *(Non-experts only)* (optional) Please use this space to clarify any of the above. *(open-ended)*

- *(Non-experts only)* For each of the following pieces of advice, please rate on a scale from 1 to 5 how **LIKELY YOU WOULD BE TO FOLLOW** the advice, if you heard it would help protect your security online. *(grid question)*
Scale: 5 (Very likely), 4, 3, 2, 1 (Not at all), I don't know
 - Don't click on links that people or companies you don't know send you
 - Don't enter your password when you click on a link in an email and that link takes you to a website that asks for your password
 - Pay attention when taking online surveys. We appreciate your input. To let us know you're paying attention, select four for this response
 - Look at the URL bar to verify that you are visiting the website you intended to
 - Don't open email attachments from people or companies you don't know
 - Use AI (e.g. ChatGPT) to get security advice
- *(Non-experts only)* (optional) Please use this space to clarify any of the above. *(open-ended)*
- What is your gender? *(multiple-choice)*
 - Woman
 - Man
 - Non-binary
 - I prefer not to answer
 - I prefer to self-describe: *(open-ended)*
- What is your age? *(multiple-choice)*
 - 18-24 years old
 - 25-34
 - 35-44
 - 45-54
 - 55-64
 - 65 or older
 - I prefer not to answer
- What is the highest degree or level of school that you have completed? *(multiple-choice)*
 - Professional doctorate (for example, MD, JD, DDS, DVM, LLB)
 - Doctoral degree (for example, PhD, EdD)
 - Masters degree (for example, MS, MBA, MEng, MA, MEd, MSW)
 - Bachelor (for example, BS, BA; also German Berufsausbildung)
 - Associates Degree (or German Abitur)
 - Some college, no degree
 - Technical/Trade school

- Regular High School Diploma (or German Realschulabschluss)
 - GED or alternative credential
 - Some High School (or German Hauptschulabschluss)
 - I prefer not to answer
 - Other (*open-ended*)
- (*Experts only*) How many total years of experience do you have in computer security? 'Experience' includes years at work or studying in a security-related field. (*open-ended*)
- (*Experts only*) What is your current job role?
For example, Network Security Engineer, Penetration Tester (*open-ended*)
 - Researcher
 - Principal Architect
 - IT Strategist
 - CEO
 - Manager
 - Security Engineer
 - Engineer
 - Other (*open-ended*)
- (*Experts only*) Which of the following best characterizes your workplace? (*multiple-choice*)
 - University
 - Corporate research lab
 - Industry
 - Government
 - Self-employed
 - Other (*open-ended*)
- (*Experts only*) In what country do you work? (*multiple-choice*)
 - Australia
 - Canada
 - Germany
 - India
 - Sweden
 - United Kingdom
 - United States
 - Other (*open-ended*)
- (*Experts only*) In what state do you work? (*open-ended*)

- (*Experts only*) Do you remember taking a survey with similar questions in the past (ca. 2014, or ca. 2018)? (*multiple-choice*)
 - Yes
 - No
- (*Expert Freelancers only*) If you know the answer to the following questions, please summarize it briefly for each question below. It is okay to leave the fields blank. Please do not use any other resources to answer these questions.
- (*Expert Freelancers only*) How does quantum gravity work? (*open-ended*)
- (*Expert Freelancers only*) How do spin-locks work in the Linux operating system? (*open-ended*)
- (*Expert Freelancers only*) What is the exact depth of the mariana trench in meters? (*open-ended*)
- (*Expert Freelancers only*) It is okay if you did not know any answers to the previous three questions. We use them as additional checks and would only expect specialists to know the answers.
- (*Non-experts only*) Which describes your current employment status? (*multiple-choice*)
 - Employed full-time
 - Employed part-time
 - Self-employed
 - Care-provider
 - Homemaker
 - Retired
 - Student - Undergraduate
 - Student - Masters
 - Student - Doctoral
 - Looking for work / Unemployed
 - Other (*open-ended*)
- (*Non-experts only*) What is your occupation? (*open-ended*)
- (*Non-experts only*) If you know the answer to the following questions, please summarize it briefly for each question below. It is okay to leave the fields blank. Please do not use any other resources to answer these questions.
- (*Non-experts only*) What is the difference between the two encryption algorithms Rivest-Shamir-Adleman (RSA) and ECC (Elliptic Curve Cryptography)? (*open-ended*)
- (*Non-experts only*) How do spin-locks work in the Linux operating system? (*open-ended*)
- (*Non-experts only*) What is salting in the context of password storage? (*open-ended*)

- (*Non-experts only*) It is okay if you did not know any answers to the previous three questions. We use them as additional checks and would only expect IT professionals to know the answers.
- (Optional) Thank you for completing our survey. Is there anything else you'd like to add or clarify? (*open-ended*)