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

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

Matthias Fassl
matthias.fassl@cispa.de
CISPA Helmholtz Center for Information Security
Germany
Saarland University
Germany

Alexander Ponticello
alexander.ponticello@cispa.de
CISPA Helmholtz Center for Information Security
Germany
Saarland University
Germany

Florin Martius
martius@uni-bonn.de
University of Bonn
Germany

Anne Mertens
anne.mertens@uni-bonn.de
University of Bonn
Germany

Katharina Krombholz
krombholz@cispa.de
CISPA Helmholtz Center for Information Security
Germany

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

ABSTRACT

When conducting qualitative research it is necessary to decide how many researchers should be involved in coding the data: Is one enough or are more coders beneficial? To offer empirical evidence for this question, we designed a series of studies investigating qualitative coding. We replicated and extended a usable security and privacy study by Ion et al. to gather both simple survey data and complex interview data. We had a total of 65 students and seven researchers analyze different parts of this data. We analyzed the codebook creation process, similarity of outcomes, inter-rater reliability, and compared the student to the researcher outcomes. We also surveyed five years of SOUPS-PC members about their views on coding. The reviewers view on coding practices for complex and simple data are almost identical. However, our results suggest that the coding process can be different for the two types of data, with complex data benefiting more from interaction between coders.

CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580766

CCS CONCEPTS

• **General and reference → Reliability**; **Empirical studies**; • **Security and privacy → Human and societal aspects of security and privacy**.

KEYWORDS

qualitative analysis, quality criteria, reliability

ACM Reference Format:
Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. 2023. Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of an Interview and Survey Study on Security Advice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3544548.3580766

1 INTRODUCTION

Qualitative content analysis is a common research method in the field of Human Computer Interaction (HCI). Different forms of qualitative analysis are used to generate theory [e.g. 53], investigate novel topics [e.g. 68] and inform follow-up quantitative analyses [e.g. 22, 60] or add depth to existing ones [e.g. 15, 43]. 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 [65].

In this paper, we will examine aspects of this coding process, including the effect 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 Usable Security and Privacy (USP). The sub-field was chosen since the authors have been involved in the paper reviewing and acceptance process as reviewers, program committee (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 [25] in HCI and for adhering to perceived statistical standards [3], e.g. regarding sample size [4] 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 [42]. 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:

RQ1.1: How does simple vs. complex data affect the similarity of results of two coders within a group of coders?

RQ1.2: How does simple vs. complex data affect the similarity of results between groups of two coders?

RQ2: How does coders' previous experience with qualitative data analysis affect the results?

RQ1 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 [12] and more relaxed recommendations [42]. RQ2 arose when we found differences between student and researcher coders. Prior work has examined the influence of researcher characteristics, such as epistemological stance [21, 35], experience [21] and background on qualitative analysis outcomes when using different qualitative analysis methods [14]. 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:

RQ3: 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 [42], we replicate and extend parts of the "No one can hack my mind" study [13, 32] 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 Symposium on Usable Privacy and Security (SOUPS) program committee 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.

2 RELATED WORK

This work investigates how researchers and the type of collected data influence the results of the qualitative analysis process. We briefly introduce qualitative analysis, including common terminology, and discuss coding practices in USP. Additionally, we present works investigating qualitative analysis methods on a meta-level.

2.1 A Short Introduction to Qualitative Analysis

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. 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 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 [10]. When the analysis involves multiple researchers, they usually discuss disagreements about code assignments. Sometimes they also calculate and report an *inter-rater reliability* value.

2.2 Coding Practices

Many different coding practices are used in the field of USP. Recommendations like those of McDonald et al. [42] suggest that only one coder codes when a researcher has special expertise, which is common in ethnographic research [e.g. 46, 53, 61], or where data is simple, e.g., when coding survey responses [e.g. 22]. However, most published work in USP involves multiple coders coding the same data. These may code everything, either independently [e.g. 38, 62] or jointly [e.g. 55]. Multiple coders may also code a subset of the data, before splitting up the remaining data [e.g. 5, 6, 22, 56, 69]. The subset of double coded data ranges from 8.7% of responses [69], through 20% of data [56], 7 of 16 interviews (43.8%) [22] and one third (33%) of transcripts [6] to 16 of 26 (61.5%) transcripts [5]. 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. 5] or the assignment of concrete codes to data [e.g. 56]. Different measures of agreement may be reported for specific topics in the codebook [e.g. 49], or as an overall agreement [e.g. 56]. Agreement procedures are often iterative and include phases of coding, discussions, and adjusting the codebook [e.g. 5]. They can also involve calculating inter-rater reliability (IRR) or similar statistical measures [e.g. 22]. 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 2, 23], or whether disagreements remain [e.g. in 56, 60], and if and how these disagreements are discussed.

Researchers also refer to different methodological approaches, such as Grounded Theory [e.g. 16, 41, 45], the General Inductive Approach [e.g. 46, 61], or Thematic Analysis [e.g. 28, 47, 62]. 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 Quality criteria for qualitative research

The review process ensures quality control of research before further dissemination. Quantitative research has well-established quality criteria: Usually, reviewers focus on reliability, validity, and generalizability [11]. If and how these criteria apply to qualitative research is part of an ongoing discussion [11, 26, 36, 40].

2.4 Meta-Analyses of Qualitative Methods

Prior work comparing different methods of qualitative analysis focused on specific methods or approaches. For example, Blair subjectively compared open coding and template coding [9], 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 [18]. Thematic Analysis and Rapid Analysis were found to produce largely similar outcomes with much overlap but also some distinct findings [57]. However, these may be attributed to different levels of immersion into the topic for the different groups of researchers [57]. Wertz et al. analyzed the same data using five different approaches: Phenomenological Psychology, Grounded Theory, Discourse Analysis, Narrative Research, and Intuitive Inquiry [65]. 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 [65]. 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 [65]. 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 [14], e.g. variants of phenomenological analysis [35], variants of narrative analysis [20], or more different methods, like grounded theory, interpretative phenomenological analysis, Foucauldian discourse analysis and narrative analysis [21]. A single researcher can use multiple methods in their analysis [20, 67], or the different methods are applied by different researchers [21, 35]. 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 [50]. However, their coding was not intended to aid sense-making [50].

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 [7]. Reliability of heuristics in the common usability method heuristic evaluation was also evaluated in various contexts [29], e.g., basic user interface elements [34], websites [54], and gaming [66].

To date, the similarity of results when different researchers qualitatively analyze the same data using the same method has not yet been investigated.

3 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 1 shows an overview of the involved participants. Table 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.

3.1 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 [42], 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.

We conducted two studies, see Table 2: (1) content-study-simple, a survey on security advice and related behavior, replicating Ion et al.'s and Busse et al.'s work [13, 32], and (2) content-study-complex,

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 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 2: Overview of the studies conducted for the empirical evaluation of the coding process.

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 **s**imple 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 [13, 32]. We consider this type of data, which we collected through interviews, **c**omplex 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 [63], assigning security scores [24, 45] 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 [33, 37] or exploratory data analysis of interviews about implementing cryptography [27]. In general, when it is not clear what aspects of the data should be coded, as in open coding [52] 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 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 [59].

To investigate the usefulness of multiple coders, we wanted to compare the outcomes of the coding process within (RQ1.1) and across multiple different groups of coders (RQ1.2). 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 (RQ2), 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 (RQ1), and empirically evaluate the guidelines put forward by McDonald et al. [42], we collected security advice in a survey like Busse et al. [13] 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.

3.3 Interview Study (Complex data)

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

3.3.1 *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 [13] 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?

3.3.2 *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 [17]. As shown in Figure 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 1) was advised to base their coding on the research questions named in Section 3.3.1. Each student chose a research partner to work with on the coding task and followed a general approach to inductive coding [59]. 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 2.

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.

3.3.3 *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 [1]. 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 RQ1.2 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

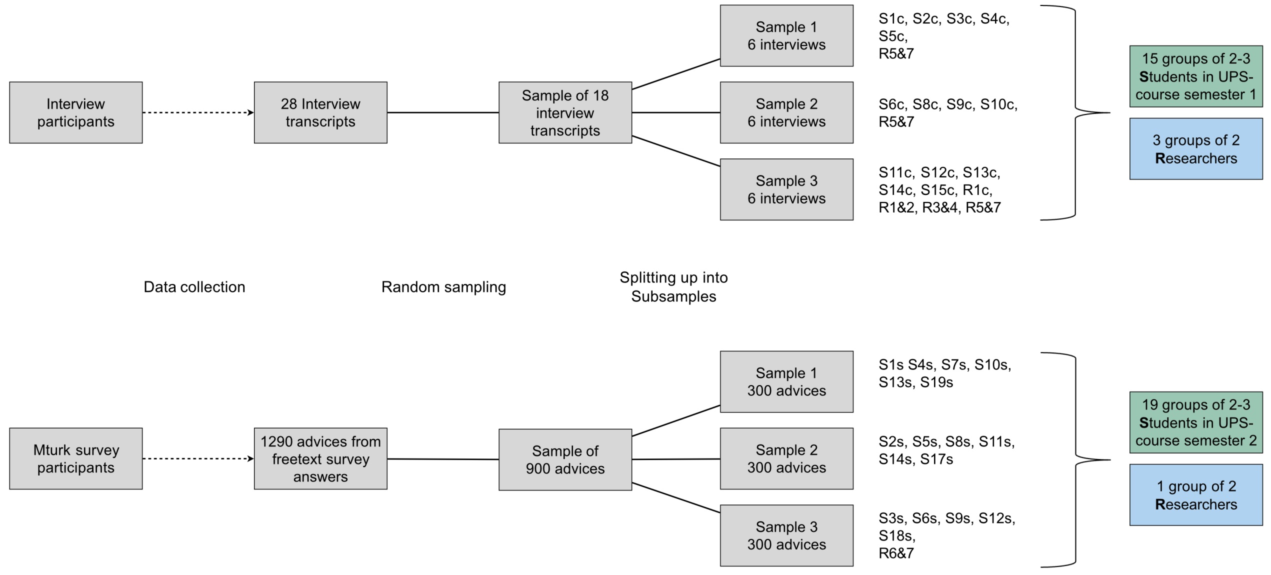

Interview participants
28 Interview transcripts
Sample of 18 interview transcripts
Sample 1
6 interviews
Sample 2
6 interviews
Sample 3
6 interviews
S1c, S2c, S3c, S4c, S5c, R5&7
S6c, S8c, S9c, S10c, R5&7
S11c, S12c, S13c, S14c, S15c, R1c, R1&2, R3&4, R5&7
15 groups of 2-3 Students in UPS-course semester 1
3 groups of 2 Researchers
Data collection
Random sampling
Splitting up into Subsamples
Mturk survey participants
1290 advices from freetext survey answers
Sample of 900 advices
Sample 1
300 advices
Sample 2
300 advices
Sample 3
300 advices
S1s S4s, S7s, S10s, S13s, S19s
S2s, S5s, S8s, S11s, S14s, S17s
S3s, S6s, S9s, S12s, S18s, R6&7
19 groups of 2-3 Students in UPS-course semester 2
1 group of 2 Researchers

Figure 1: Process of Data Preparation and Visualization for Empirical Evaluation

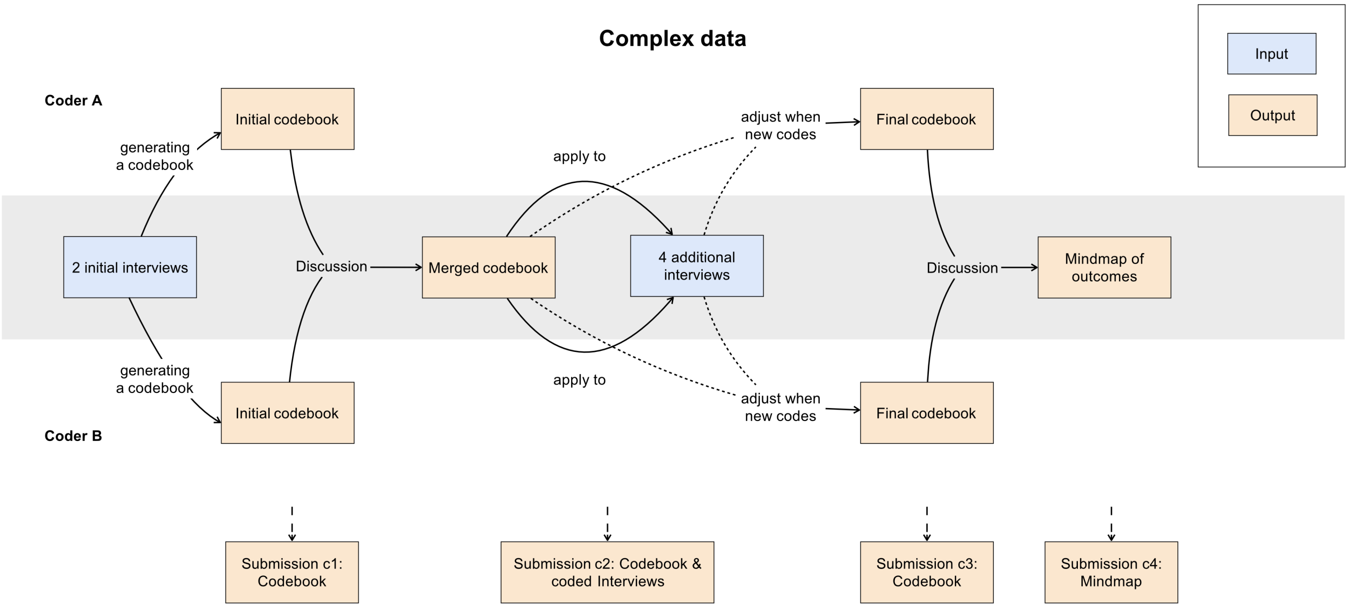


Figure 2: Interview coding process and submitted outcomes

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 (RQ1.1), 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.

To summarize, to answer part of RQ1, 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 (RQ1.2) and within a group of data analysts, e.g. by calculating IRR or conducting change tracking from the initial codebooks to the mindmaps (RQ1.1). To answer RQ2, in our comparisons between groups of data analysts, we specifically focused on differences between groups of researchers and groups of students.

3.4 Survey Study (Simple Data)

3.4.1 Survey Procedure. We re-ran two surveys from Busse et al. on Amazon Mechanical Turk [13]. 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.

3.4.2 Content-Study Analysis. Figure 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 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).

3.4.3 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 [13, 32] 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 [13, 32]. Like in the replication by Busse et al. [13], 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.

3.5 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 2) here.

3.5.1 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 [59] to code originally written material, like survey answers, and one had spent about 160 hours applying critical discourse analysis [19] 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 **c**omplex data from the interviews and *Ss* for those analyzing more **s**imple 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.

3.5.2 Researchers. A total of 7 researchers participated in data analysis for this project, as shown in Table 1. We will refer to groups by adding an ampersand between the researchers' identifiers, e.g., R5&7 for the 2021 course instructors.

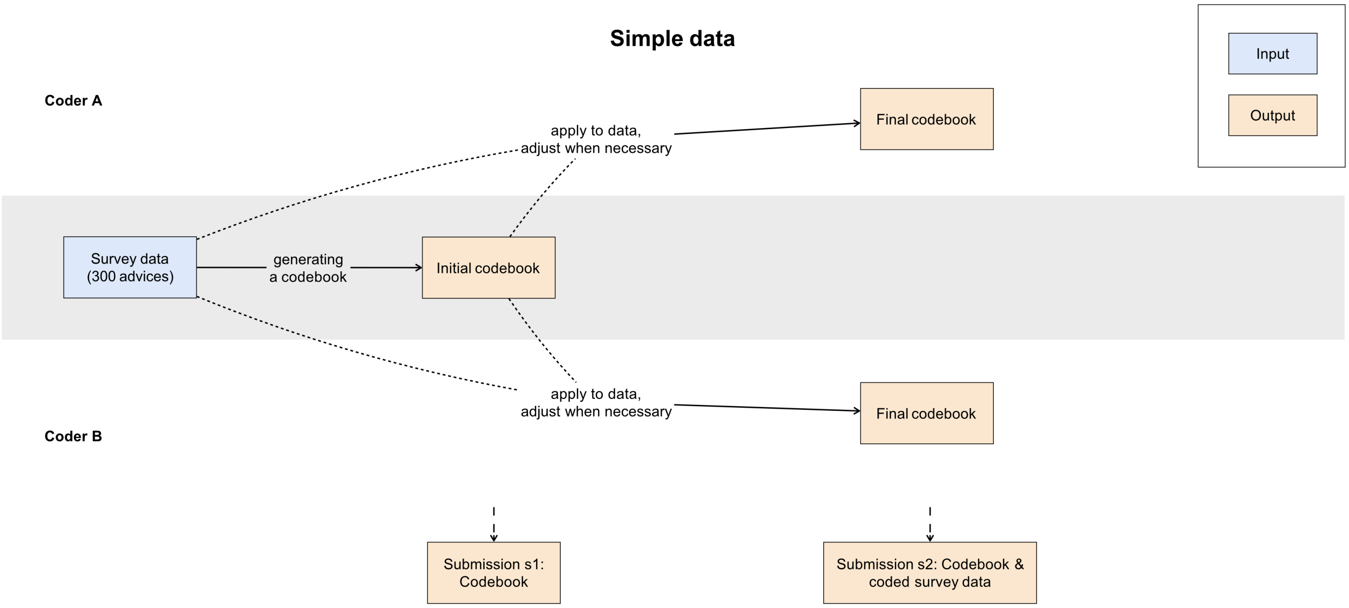


Figure 3: Survey coding process and submitted outcomes

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.

3.6 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.

4 RESULTS

To investigate the similarity of outcomes within and between groups (RQ1), 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 RQ2.

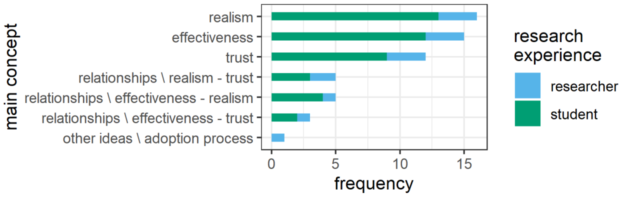


Figure 4: Frequency of main concepts in mindmaps based on complex data (N=16)

4.1 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 (RQ2) and between the interview and the survey study (RQ1.2).

4.1.1 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 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 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 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 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 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 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.

4.1.2 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 [13, 32], 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 7.

All five categories of advice appeared in all final codebooks except for S13s and S2s, as shown in Figure 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 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

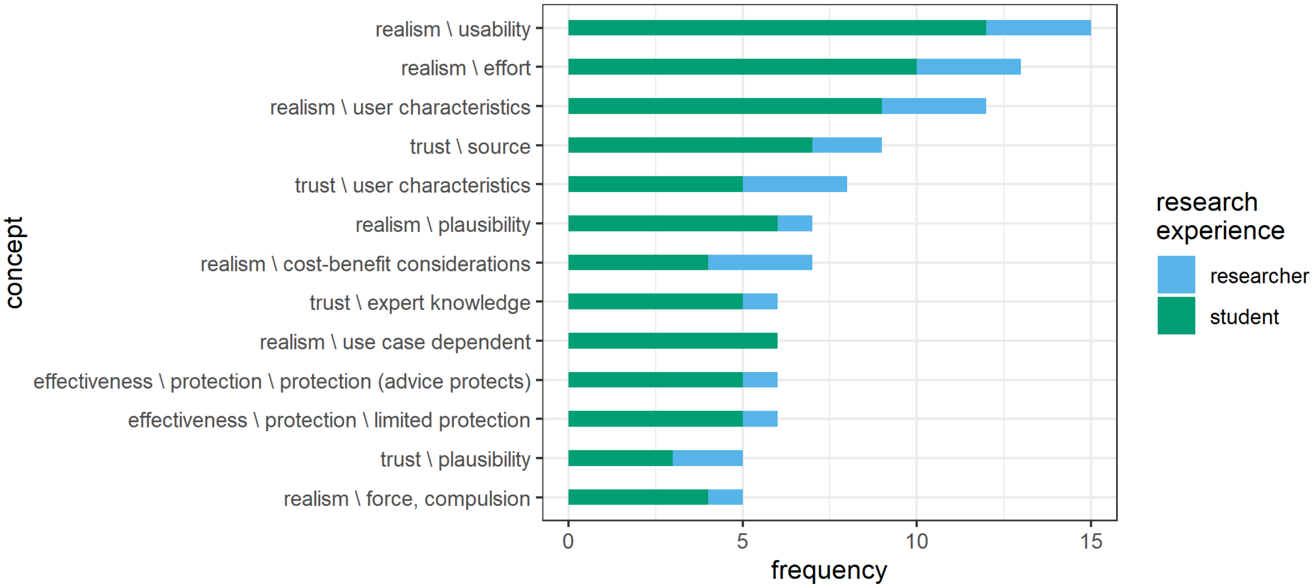


Figure 5: Frequency of concepts in interview-based mindmaps (only showing concepts that were present in the mindmap of at least 5 of the 16 groups)

to *security software* are 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.

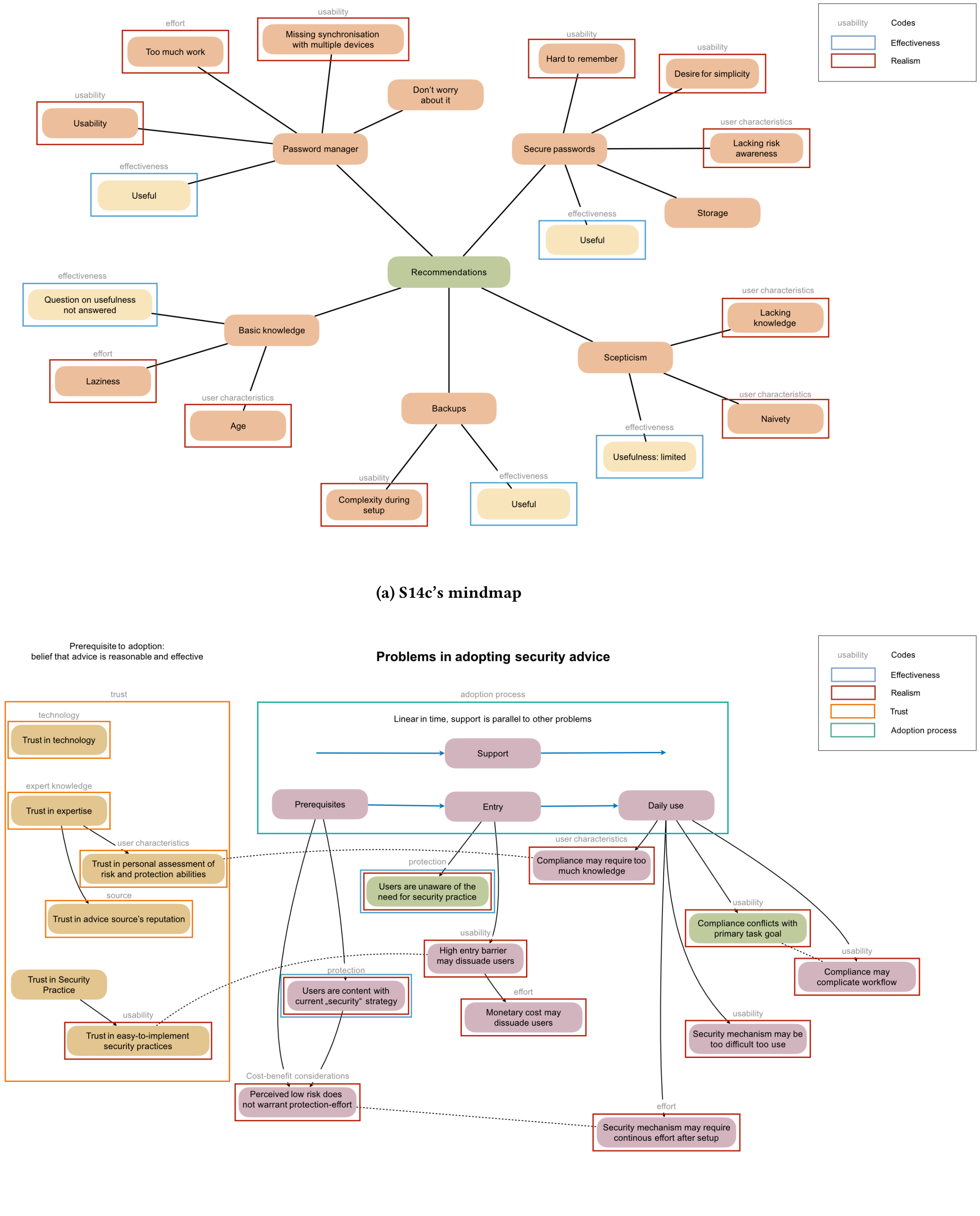
Likewise, some concepts which featured prominently in prior work and were thus included in the codebook from the beginning, did not appear at all. In some cases, e.g., *Use Linux*, this may have been caused by our participant sample, which only consisted of MTurk workers, and likely did not include experts like in prior work [13, 32]. 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.

4.1.3 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 (RQ2).

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 (RQ1.2).

4.2 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


usability
Codes
Effectiveness
Realism
effort
Too much work
usability
Missing synchronisation with multiple devices
Don't worry about it
usability
Usability
effectiveness
Useful
Password manager
usability
Hard to remember
usability
Desire for simplicity
Secure passwords
user characteristics
Lacking risk awareness
effectiveness
Useful
Storage
Recommendations
effectiveness
Question on usefulness not answered
Basic knowledge
effort
Laziness
user characteristics
Age
Backups
usability
Complexity during setup
effectiveness
Useful
Scepticism
user characteristics
Lacking knowledge
user characteristics
Naivety
effectiveness
Usefulness: limited

(a) S14c's mindmap

Prerequisite to adoption:
belief that advice is reasonable and effective
Problems in adopting security advice
usability
Codes
Effectiveness
Realism
Trust
Adoption process
trust
technology
Trust in technology
expert knowledge
Trust in expertise
user characteristics
Trust in personal assessment of risk and protection abilities
source
Trust in advice source's reputation
Trust in Security Practice
usability
Trust in easy-to-implement security practices
adoption process
Linear in time, support is parallel to other problems
Support
Prerequisites
Entry
Daily use
user characteristics
Compliance may require too much knowledge
protection
Users are unaware of the need for security practice
usability
High entry barrier may dissuade users
protection
Users are content with current „security" strategy
effort
Monetary cost may dissuade users
usability
Compliance conflicts with primary task goal
usability
Compliance may complicate workflow
usability
Security mechanism may be too difficult too use
Cost-benefit considerations
Perceived low risk does not warrant protection-effort
effort
Security mechanism may require continous effort after setup

(b) R3&4's mindmap

Figure 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.

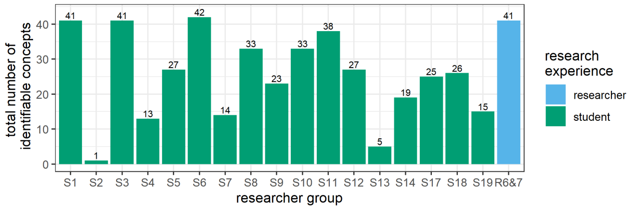


Figure 7: Number of identifiable concepts for each of the submitted final codebooks for simple data

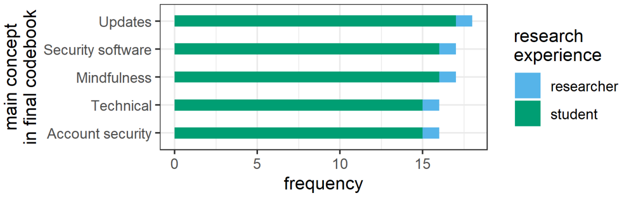


Figure 8: Frequency of main concepts in 18 final codebooks based on simple data

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 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 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 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 RQ2, 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 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.

4.3 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 (RQ1.1) for complex data, we compared the codebooks between different stages of the analysis process.

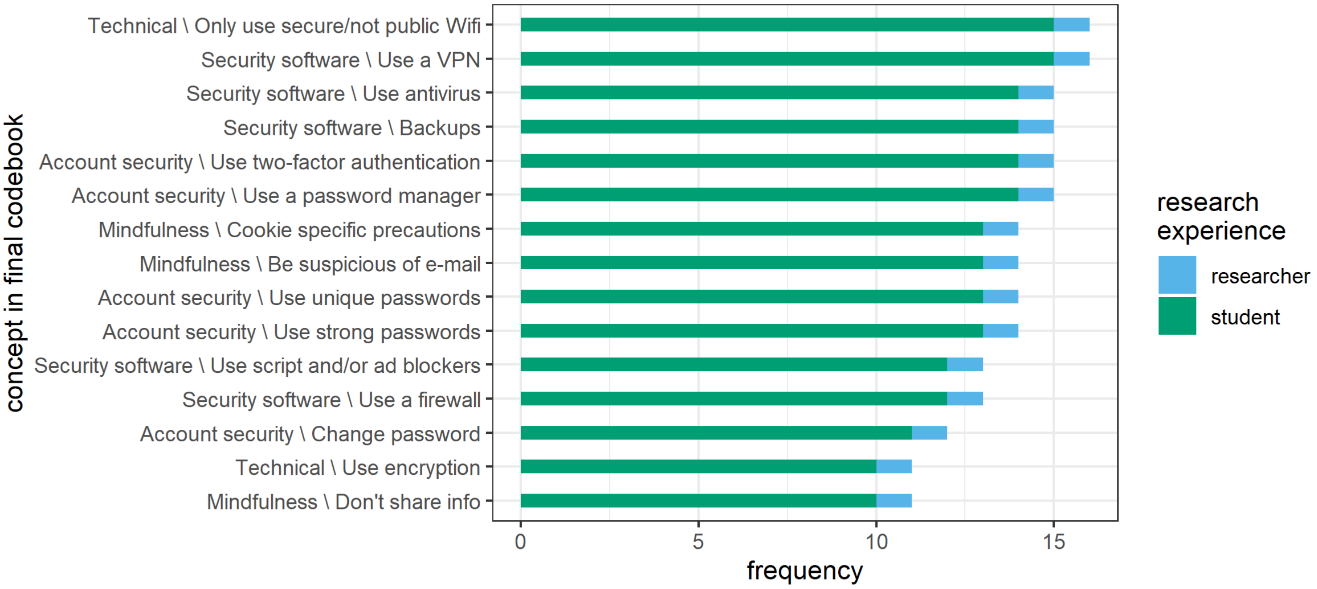


Figure 9: Frequency of lower-level concepts in 16 final codebooks based on simple data

4.3.1 *Development of Codebooks.*

We examined content differences between the initial codebooks, where meta-level participants worked alone, and the final outcomes, the mindmaps. In 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 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 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 RQ1.1, multiple coders were beneficial in analyzing complex data.

4.4 The Influence of Different Data Types and Researcher Experience on Calculated IRR

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 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 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 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

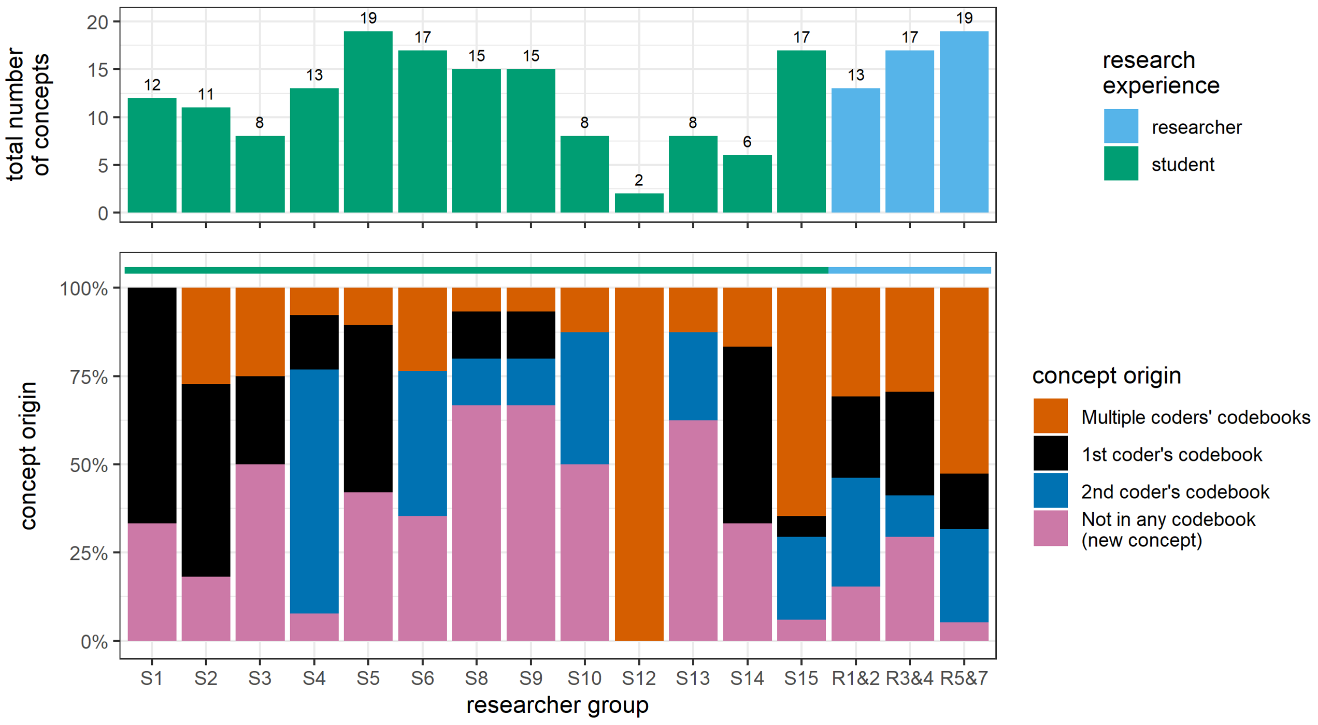


Figure 10: Overview of concept origin: Upper part of the graph shows the total number of concepts present in the mindmaps, while the lower part of the graph shows the proportion of different origins for the concepts in the mindmap.

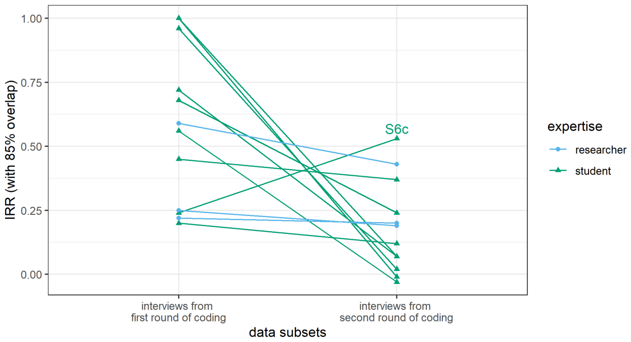


Figure 11: IRR for different subsets of interviews

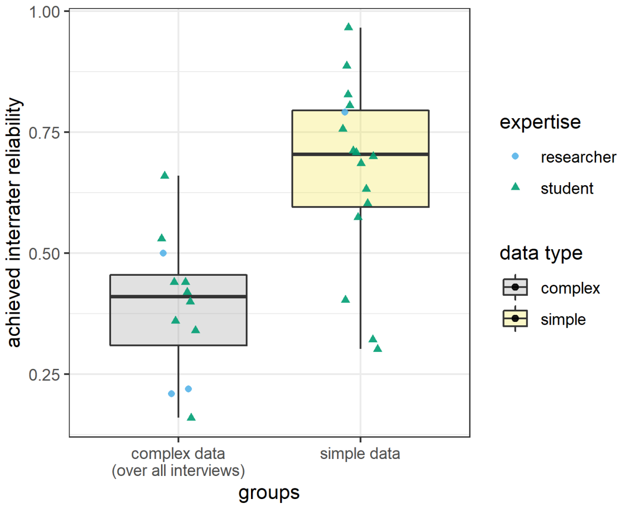


Figure 12: IRR compared for different data types and experience levels

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 12. This difference becomes even more pronounced when comparing to the interviews coded in the second round of coding (see Figure 11). Regarding RQ1.1, 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 RQ2, 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’ 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.

4.5 Summary: RQ1 and RQ2

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 (RQ1.2). 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 (RQ1.1).

Regarding RQ2, 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.

5 SURVEY OF SOUPS PC 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. [42] 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 [8, 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.

5.1 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 amongst participants who wished to join the raffle.

5.2 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 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.

5.3 Data Analysis

Two authors used the General Inductive Approach [59] 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>.

5.4 Important Quality Criteria

We asked PC members to judge the importance of different criteria in their reviewing process. Figure 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".

5.5 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 [32] 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 [42], we don't see any large differences between complex and simple data, e.g., Figure 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 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 mind 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 5.4, such as sufficient justification for the process, how the coding results are used, and the type of analysis conducted.

5.6 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 [39]. Figure 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

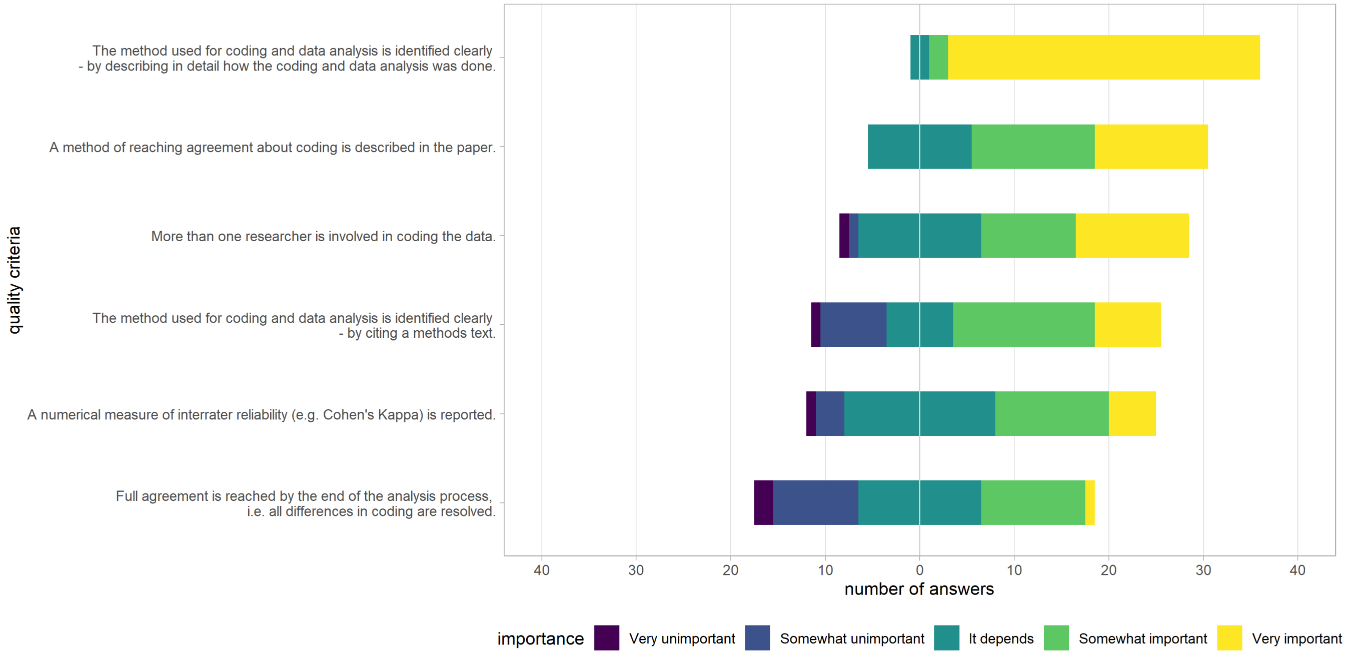


Figure 13: PC members' importance judgments for given quality criteria, There was one response of "I don't know" each for "agreement described" and "full agreement reached", which were excluded from this visualization

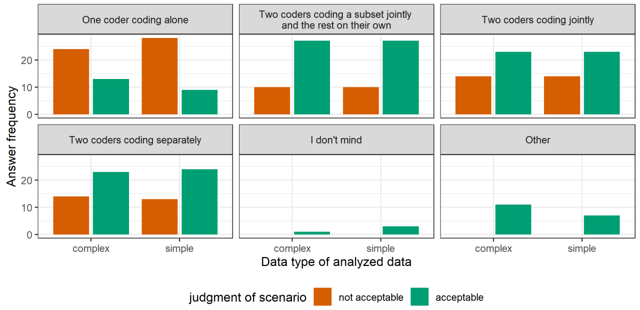


Figure 14: PC members' acceptance of different scenarios regarding the number of involved coders and the type of analyzed data. Non-answers for "left-over" categories of *I don't mind* and other are not graphed since they don't carry information.

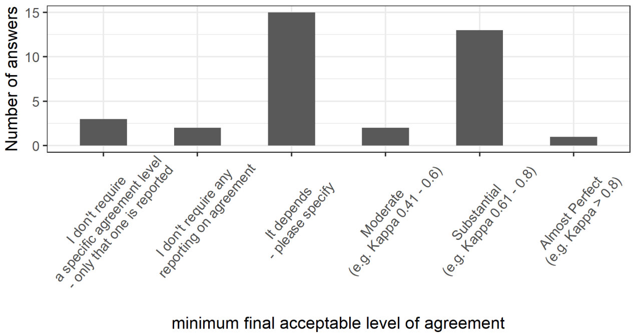


Figure 15: Minimum acceptable levels of numerical agreement

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 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.

6 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 [48]. 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 [65], 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 [44], 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.

7 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 [21]. 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 [31]. 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 [30]. In some instances HCI [58] and USP [64] 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 [51]. 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 3.1 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 [14, 21]. 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.

8 RECOMMENDATIONS FOR RESEARCH PRACTICE

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

(1) Use multiple coders when research questions and data are open-ended and complex. (RQ1)
(2) When data is simple, research questions are well-defined, and analysis is straightforward, use only a single coder to code the data after a codebook has been established in collaboration with multiple coders. (RQ1)
(3) When coding with multiple researchers, discussions and interaction between coders are a vital part of the analysis process. Do not analyze in isolation. (RQ1)
(4) Make clear who the analyzing researchers are and what level of expertise they have, both regarding the topic and the analysis methods used. (RQ2)
(5) 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, even when it is a well-established method. (RQ3)

Since our studies were designed to be representative of USP studies and included two types of data of a complexity typical for this field, our recommendations 1, 2 and 3 apply 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 3.1 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 our recommendations apply to their study context. 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 our recommendation 4 is also applicable to the broader HCI community, but note that making positionality explicit may be more wide-spread already outside USP. The sampling frame for our PC member survey was the SOUPS PC of the last five years. As such, our recommendation 5 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 [31], we believe that explicating these decisions is beneficial not only when using qualitative methods but also for quantitative work.

9 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.

ACKNOWLEDGMENTS

The authors wish to thank Alena Naiakshina, Anastasia Danilova and the student participants of the Usable Security and Privacy Exercises at the University of Bonn in summer semester of 2021 and 2022 for their help in data collection! This work was partially funded by the ERC Grant 678341: Frontiers of Usable Security.

REFERENCES

[1] 2022. MAXQDA Manual. https://www.maxqda.com/help-mx22/welcome.

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

[3] Peter Bacchetti. 2002. Peer Review of Statistics in Medical Research: The Other Problem. *BMJ : British Medical Journal* 324, 7348 (May 2002), 1271–1273.

[4] Peter Bacchetti. 2010. Current Sample Size Conventions: Flaws, Harms, and Alternatives. *BMC Medicine* 8, 1 (March 2010), 17.

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

[6] Khadija Baig, Reham Mohamed, Anna-Lena Theus, and Sonia Chiasson. 2020. "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, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[7] Christopher Bailey, Elaine Pearson, and Voula Gkatzidou. 2014. Measuring and comparing the reliability of the structured walkthrough evaluation method with novices and experts. In *Proceedings of the 11th Web for All Conference on - W4A '14*. ACM Press, Seoul, Korea, 1–10.

[8] Kathy Baxter, Catherine Courage, and Kelly Caine. 2015. *Understanding Your Users. A Practical Guide to User Research Methods* (second ed.). Morgan Kaufmann Publications.

[9] Erik Blair. 2015. A reflexive exploration of two qualitative data coding techniques. *Journal of Methods and Measurement in the Social Sciences* 6, 1 (2015), 14–29.

[10] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101.

[11] Virginia Braun and Victoria Clarke. 2013. *Successful Qualitative Research: A Practical Guide for Beginners*. SAGE.

[12] Loraine Busetto, Wolfgang Wick, and Christoph Gumbinger. 2020. How to Use and Assess Qualitative Research Methods. *Neurological Research and Practice* 2, 1 (May 2020), 14.

[13] Karoline Busse, Julia Schäfer, and Matthew Smith. 2019. 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 2019)*. USENIX Association, Santa Clara, CA.

[14] Nicola J. Clarke, Martin E. H. Willis, Jemima S. Barnes, Nick Caddick, John Cromby, Hilary McDermott, and Gareth Wiltshire. 2015. Analytical Pluralism in Qualitative Research: A Meta-Study. *Qualitative Research in Psychology* 12, 2 (April 2015), 182–201.

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

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

[17] Thorsten Dresing and Thorsten Pehl. 2017. *Praxisbuch Interview, Transkription & Analyse: Anleitungen und Regelsysteme für qualitativ Forschende* (7 ed.). self-published, Marburg.

[18] Isabelle F. Dufour and Marie-Claude Richard. 2019. Theorizing from Secondary Qualitative Data: A Comparison of Two Data Analysis Methods. *Cogent Education* 6, 1 (2019), 1690265.

[19] Norman Fairclough. 2013. Critical Discourse Analysis. In *The Routledge Handbook of Discourse Analysis*. Routledge, 9–34.

[20] Nollaig Frost. 2009. 'Do You Know What I Mean?': The Use of a Pluralistic Narrative Analysis Approach in the Interpretation of an Interview. *Qualitative Research* 9, 1 (Feb. 2009), 9–29.

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

[22] Kelsey R Fulton, Anna Chan, Daniel Votipka, Michael Hicks, and Michelle L Mazurek. 2021. Benefits and Drawbacks of Adopting a Secure Programming Language: Rust as a Case Study. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021) (SOUPS '21)*. USENIX Association, 597–616.

[23] Eva Gerlitz, Maximilian Häring, and Matthew Smith. 2021. Please Do Not Use !?_ or Your License Plate Number: Analyzing Password Policies in German Companies. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 17–36.

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

[25] Saul Greenberg and Bill Buxton. 2008. Usability Evaluation Considered Harmful (Some of the Time). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 111–120.

[26] Egon Guba. 1981. ERIC/ECTJ Annual Review Paper: Criteria for Assessing the Trustworthiness of Naturalistic Inquiries. 29, 2 (1981), 75–91.

[27] Julie M Haney, Mary F Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. 2018. "We Make It a Big Deal in the Company": Security Mindsets in Organizations That Develop Cryptographic Products. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018) (SOUPS '18)*. USENIX Association, Baltimore, MD, 357–373.

[28] Ayako A. Hasegawa, Naomi Yamashita, Tatsuya Mori, Daisuke Inoue, and Mitsuaki Akiyama. 2022. Understanding Non-Experts' Security- and Privacy-Related Questions on a Q&A Site. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. USENIX Association, Boston, MA, 39–56.

[29] Morten Hertzum and Niels Ebbe Jacobsen. 2001. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human–Computer Interaction* 13, 4 (2001), 421–443.

[30] Andrew Gary Darwin Holmes. 2020. Researcher Positionality - A Consideration of Its Influence and Place in Qualitative Research - A New Researcher Guide. *Shanlax International Journal of Education* 8, 4 (Sept. 2020), 1–10.

[31] 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. 2021. The Influence of Hidden Researcher Decisions in Applied Microeconomics. *Economic Inquiry* 59, 3 (2021), 944–960.

[32] Iulia Ion, Rob Reeder, and Sunny Consolvo. 2015. "...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 2015)*. USENIX Association, Ottawa, 327–346.

[33] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. 2015. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. USENIX Association, Ottawa, 39–52.

[34] Martin Kessner, Jo Wood, Richard F. Dillon, and Robert L. West. 2001. On the Reliability of Usability Testing. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (Seattle, Washington) *(CHI EA '01)*. Association for Computing Machinery, New York, NY, USA, 97–98.

[35] Nigel King, Linda Finlay, Peter Ashworth, Jonathan A. Smith, Darren Langdridge, and Trevor Butt. 2008. "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 (May 2008), 80–102.

[36] Laura Krefting. 1991. Rigor in Qualitative Research: The Assessment of Trustworthiness. 45, 3 (1991), 10.

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

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

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

[40] Lawrence Leung. 2015. Validity, reliability, and generalizability in qualitative research. *Journal of family medicine and primary care* 4, 3 (2015), 324–327. Publisher: Medknow Publications & Media Pvt Ltd.

[41] Alexandra Mai, Katharina Pfeffer, Matthias Gusenbauer, Edgar Weippl, and Katharina Krombholz. 2020. User Mental Models of Cryptocurrency Systems - A

Grounded Theory Approach. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, 341–358.

[42] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. 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 (Nov. 2019), 1–23.

[43] Sarah Nadi, Stefan Krüger, Mira Mezini, and Eric Bodden. 2016. Jumping through Hoops: Why Do Java Developers Struggle with Cryptography APIs?. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, Austin Texas, 935–946.

[44] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. 2020. 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*. ACM, Honolulu HI USA, 1–13.

[45] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. 2017. Why Do Developers Get Password Storage Wrong? A Qualitative Usability Study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS '17)*. Association for Computing Machinery, New York, NY, USA, 311–328.

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

[47] Nikhil Patnaik, Joseph Hallett, and Awais Rashid. 2019. Usability Smells: An Analysis of Developers' Struggle With Crypto Libraries. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019) (SOUPS '19)*. USENIX Association, Santa Clara, CA, USA, 245–257.

[48] Anne E Pezalla, Jonathan Pettigrew, and Michelle Miller-Day. 2012. Researching the researcher-as-instrument: an exercise in interviewer self-reflexivity. *Qualitative Research* 12, 2 (2012), 165–185.

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

[50] Carrie B. Sanders and Carl J. Cuneo. 2010. Social Reliability in Qualitative Team Research. *Sociology* 44, 2 (April 2010), 325–343.

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

[52] Anselm Strauss and Juliet M Corbin. 1997. *Grounded Theory in Practice*. Sage, London.

[53] Sathya Chandran Sundaramurthy, Alexandru G. Bardas, Jacob Case, Xinming Ou, Michael Wesch, John McHugh, and S. Raj Rajagopalan. 2015. A Human Capital Model for Mitigating Security Analyst Burnout. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. USENIX Association, Ottawa, 347–359.

[54] Alistair Sutcliffe. 2002. Assessing the reliability of heuristic evaluation for Web site attractiveness and usability. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*. 1838–1847.

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

[56] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. 2020. Understanding Privacy-Related Questions on Stack Overflow. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14.

[57] Beck Taylor, Catherine Henshall, Sara Kenyon, Ian Litchfield, and Sheila Greenfield. 2018. 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 (Oct. 2018).

[58] Jennyfer Lawrence Taylor, Alessandro Soro, Paul Roe, Anita Lee Hong, and Margot Brereton. 2018. "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*. ACM, Montreal QC Canada, 1–14.

[59] David R. Thomas. 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation* 27, 2 (June 2006), 237–246.

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

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

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

[63] Daniel Votipka, Kelsey R Fulton, James Parker, Matthew Hou, Michelle L Mazurek, and Michael Hicks. 2020. Understanding Security Mistakes Developers Make: Qualitative Analysis from Build It, Break It, Fix It. In *29th USENIX Security Symposium (USENIX Security 20) (USENIX Security '20)*. USENIX Association, 109–126.

[64] Miranda Wei, Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. 2022. Anti-Privacy and Anti-Security Advice on TikTok: Case Studies of Technology-Enabled Surveillance and Control in Intimate Partner and Parent-Child Relationships. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. USENIX Association, Boston, MA, 447–462.

[65] Frederick J. Wertz, Kathy Charmaz, Linda M. McMullen, Ruthellen Josselson, Rosemarie Anderson, and Emalinda McSpadden. 2011. *Five Ways of Doing Qualitative Analysis: Phenomenological Psychology, Grounded Theory, Discourse Analysis, Narrative Research, and Intuitive Inquiry* (1 ed.). The Guilford Press, New York.

[66] Gareth R White, Pejman Mirza-babaei, Graham McAllister, and Judith Good. 2011. Weak inter-rater reliability in heuristic evaluation of video games. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. Association for Computing Machinery, Vancouver, BC, Canada, 1441–1446.

[67] Corrine M. Wickens. 2011. The Investigation of Power in Written Texts through the Use of Multiple Textual Analytic Frames. *International Journal of Qualitative Studies in Education* 24, 2 (March 2011), 151–164.

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

[69] Yixin Zou, Kevin Roundy, Acar Tamersoy, Saurabh Shintre, Johann Roturier, and Florian Schaub. 2020. 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, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15.