



How Transparent is Usable Privacy and Security Research? A Meta-Study on Current Research Transparency Practices

Jan H. Klemmer ^C Juliane Schmäuser ^C Fabian Fischer ^C Jacques Suray ^C
Jan-Ulrich Holtgrave ^C Simon Lenau ^C Byron M. Lowens ^{*} Florian Schaub [†]
Sascha Fahl ^C

^C*CISPA Helmholtz Center for Information Security, Germany,*
{jan.klemmer, juliane.schmueser, fabian.fischer, jacques.suray,
jan-ulrich.holtgrave, lenau, fahl}@cispa.de
^{*}*Indiana University Indianapolis, USA, bylowe@iu.edu*
[†]*University of Michigan, USA, fschaub@umich.edu*

Abstract

Transparent research reporting is crucial to understanding and assessing research, its results and validity, and for fostering replication. While other research fields investigated reporting and transparency practices, similar meta-research is missing for the usable privacy and security (UPS) community, which combines security, privacy, and human research. To gain insights into current research transparency practices and their development in the UPS community, we analyzed 200 UPS publications from twelve venues (including USENIX Security, IEEE S&P, CCS, SOUPS, and CHI) from 2018 to 2023. Additionally, we evaluated those venues’ 81 calls for papers (CfPs) and 20 calls for artifacts (CfAs). We find that most papers report on many of 52 analyzed transparency criteria, but none achieve full transparency. Moreover, we uncover several areas that need improvements: essential artifacts like questionnaires are frequently missing and hinder replication, some information is reported inconsistently, and dead links further reduce availability. Our regression analysis indicates that paper length and the number of studies described in a paper impact reporting transparency, while we observed no effect of publication year and artifact evaluation (AE). Finally, we provide recommendations for authors, venues, and PC chairs to improve research transparency practices and suggest transparency guidelines.

1 Introduction

Transparent research reporting in scientific publications is fundamental to good scientific practice [74]. In response to challenges with reproducing results [7, 70, 49] and declared reproducibility crises (e.g., in medicine [11] and psychology [106]), research communities and organizations [40, 113] are increasing their efforts towards better reproducibility, and, as a prerequisite, transparent research reporting [104]. Transparent reporting enables other researchers to reliably review and assess findings [103], attempt replication [104], or gain

a comprehensive understanding of results, their context, and validity. Thus, a publication should ensure transparency by detailing methods and results, disclosing limitations, and providing artifacts such as study materials or datasets.

Research is not always reported transparently [49, 7, 74]. Overall, scholars have identified many transparency issues due to reporting shortcomings in computer science [92, 39, 37, 94, 16, 25, 95, 23]. For the security and privacy (SP) community, early studies on reporting completeness at CCS and IEEE S&P suggest that papers lack details like research objectives and limitations [15, 13]. Efforts like artifact evaluation (AE) did not significantly affect artifact availability in machine learning (ML) security papers [80]. Nonetheless, transparent reporting is essential for considering SP a scientific discipline—especially as security research has been accused of lacking scientific principles including transparency [43].

In 2025, the USENIX Security call for papers (CfP) encouraged meta-science, including research practices, transparency, and reproducibility, for the first time [9]. This paper contributes transparency insights in the usable privacy and security (UPS) subfield. We focus on UPS as it is deeply rooted in SP and uniquely blends with human-centered research, combining diverse methods and producing a myriad of research artifacts. Moreover, the focus on one subfield allows an in-depth analysis. Prior studies have indicated a lack of information in UPS publications (e.g., on demographics [41] and risk representation [24]). Recently, Klemmer et al. found that UPS researchers value transparency, apply reporting practices based on implicit community standards, and perceive an overall improvement over time—while highlighting challenges and desiring further transparency improvements [59]. We complement their qualitative insights with the first quantitative analysis of reporting transparency in UPS publications.

To assess research transparency practices in the UPS community, we conducted a meta-study of UPS papers from twelve major venues that publish UPS research, including SOUPS, USENIX Security, IEEE S&P, and CCS (see [Sec-](#)

tion 3 for details). While many papers already transparently report relevant information and materials, we identified gaps and inconsistencies that indicate room for improvement. Importantly, our work is not intended as finger-pointing or criticism of individual UPS researchers; rather, we aim to constructively contribute to improved transparency practices in the UPS research community. The following research questions guided our meta-research:

RQ1. *How transparent is research reporting in UPS?* Based on an extensive literature review, we investigated what information and materials (*transparency criteria*) are currently available in UPS publications and identified shortcomings.

RQ2. *What factors impact transparency and research reporting practices in UPS?* With our meta-study results and the bibliographic and venue metadata obtained by analyzing CfPs and calls for artifacts (CFAs), we explored factors influencing transparency and reporting practices in a regression analysis.

In this work, we make the following contributions:

List of Transparency Criteria. We identified 52 transparency criteria relevant to reporting UPS research (Table 3). Each criterion represents information or materials a paper should report if applicable to its methodology. These criteria can serve authors and reviewers as a transparency checklist.

Status Quo of Transparency. We analyzed 200 papers to assess the availability and reporting location of the 52 transparency criteria (Section 6). Overall, papers reported transparently on many criteria. Nonetheless, we uncovered potential for reporting improvements with 30.9% of criteria missing, potentially hindering replication of 28.0% of papers, and inconsistencies on what different papers report. Moreover, we found issues with the long-term availability of online artifacts.

Factors that Impact Transparency. We built a transparency score (TS) (Section 6.3) for transparency assessment based on the 52 criteria. Using a regression model (Section 6.4), we found that longer papers exhibit significantly better transparency, whereas reporting on multiple methods in a single paper significantly reduces transparency.

Recommendations. We derived ten recommendations (Section 7) for authors, venues, and program committee (PC) chairs to advance transparency in UPS research.

UPS Transparency Dataset. To support Open Science and transparency, we publish artifacts, including our analysis results and the list of all and the analyzed UPS publications at the twelve venues (2018–2023).

2 Related Work

We discuss related work on (i) transparency guidelines for researchers, (ii) research transparency perspectives, and (iii) transparency-related literature reviews.

2.1 Transparency Guidelines for Researchers

Guidelines outlining expected transparency practices are fundamental to evaluating, fostering, and unifying transparent research reporting. General guidelines, such as the *Transparency and Openness Promotion Guidelines (TOP)* [76, 17] and the *FAIR principles* [123, 30], provide overarching frameworks for science. Specific fields [72, 3] have also developed tailored guidelines, such as the *EQUATOR Network* [102, 101, 28] in medical research. In computing research, the software engineering (SE) community, among others, has established several guidelines for the execution and reporting of experiments to ensure relevant information is available and supports replicability [60, 51, 50, 100], e.g., ACM SIGSOFT’s *Empirical Standards* [93, 27] and CRA’s submission and review policies [21]. Nonetheless, existing guidelines are rarely followed. In human–computer interaction (HCI), for example, scholars discussed guideline adoption [19, 120], but important HCI journals lack comprehensive transparency guidelines [8]. Finally, following other communities, the computer security community has recently introduced artifact evaluation (AE) to advocate transparency and reproducibility (cf. Table 1) [98].

These developments demonstrate an increasing focus on transparent research reporting in adjacent fields, but specific guidelines for UPS are currently rare. In 2010, Schechter outlined how to write a UPS paper, emphasizing reporting of study details [97]. Distler et al. derive reporting guidelines from their analysis of risk representation [24], and Klemmer et al. report on transparency practices endorsed by UPS researchers [59]. Recently, Ortloff and Martius reflected on meta-science in UPS and HCI and presented several reporting and transparency recommendations [82].

However, it is unclear to what extent UPS researchers follow guidelines or best practices. Therefore, we derive 52 transparency criteria from available guidelines, and use them to evaluate the transparency of recent UPS publications.

2.2 Research Transparency Perspectives

Several studies with researchers across disciplines provide insights into experiences and challenges with reproducibility and transparency. Barriers to transparent reporting include high effort, institutional pressure, technical difficulties, and a lack of motivation and incentives for researchers [31, 34, 32, 61]. For example, a survey of CHI authors identified researcher motivation, resource constraints, reliable hosting, and difficulties with protecting personal identifiable information (PII) as reasons why sharing artifacts was uncommon [119]. Moreover, concerns regarding deanonymization and privacy [29] and methodological and ethical implications of secondary data usage [18] hinder publishing data.

For security conferences, interviews found that reviewers value transparency, stating that missing details are a weakness that may lead to the rejection of a submission [103]. Recently,

Klemmer et al. conducted an interview study on UPS researchers’ transparency practices and experiences. While they valued transparency and reported several practices, these were based on implicit community standards rather than explicit, field-wide guidelines [59]. We complement this prior qualitative work by conducting a literature analysis that provides empirical evidence for the extent of transparent reporting in recent UPS publications and identify areas for improvement.

2.3 Transparency Literature Analyses in CS

Transparency issues are widespread across computer science subdisciplines. Text mining studies [79], information system journals [92], ML publications [91, 38, 39], and web measurement research [23] often lack critical details. Similarly, SE has transparency issues [37, 94], including that desired replication studies are hard to publish [69] and face transparency issues themselves [16]. Common gaps include missing datasets, algorithm specifications, experiment code [79], bias discussion [53], statistical details like effect sizes [94], and documentation of execution processes [38, 39]. In HCI, only few papers from 2016–2017 openly released software artifacts [25]. A review of 2017 and 2022 CHI publications found improved reporting transparency in some areas, e.g., sharing interview guides, but no significant changes in others, e.g., sharing software artifacts [95]. In two studies with a focus on HCI (including UPS), Orloff et al. found that effect sizes are often not consistently and thoroughly reported and interpreted in publications [81, 83]. The authors provide recommendations, e.g., to report all effects sizes regardless of their significance [83]. In computer security, an analysis of ML security papers at Tier 1 conferences suggests a lack of computational reproducibility from provided artifacts, on which the introduction of AE had no significant impact [80]. A review of CCS and IEEE S&P papers suggests a lack of clearly defined research objectives, limitations, and relevant details of data collection and analysis [15, 13].

While a comprehensive analysis of UPS transparency is missing, prior studies provide some insights. Distler et al. noted insufficient methodological details in risk representation studies [24] and Hasegawa et al. found frequent omissions of demographic data in UPS papers, particularly in non-HCI and non-UPS venues [41]. Suray et al. suggest that future work statements are commonly reported in UPS, but often unspecific or hard to find [108, 109]. To extend prior work, we conduct a comprehensive quantitative transparency-focused systematic literature analysis of UPS papers, outlining current transparency practices and areas for improvement.

3 Considered Publication Venues & Years

To investigate the *status quo*, we considered recent conference proceedings from 2018–2023 in our literature analysis (Section 5). We did not include 2024 proceedings, because

not all were published at the time of analysis. For the CFP analysis (Section 4), we did include 2024, as all CFPs were available.

We included proceedings of twelve UPS-related conferences in our study. Those include the four top-tier security venues *IEEE S&P*, *CCS*, *USENIX Security*, and *NDSS*, as well as *EuroS&P* and the privacy venue *PETS/PoPETs*. In addition, we included both *CHI* and *CSCW* as HCI conferences that regularly publish UPS research. Moreover, we included both *SOUPS* and *EuroUSEC*¹ as designated UPS venues. In addition, we included *ICSE* and *WWW* since they published some UPS research in the past.

We argue that the chosen venues are some of the most popular venues that regularly publish UPS research under the highest quality standards, but do not claim that this list is exhaustive. Moreover, our twelve venues are commonly included in other analyses of UPS research [24, 41, 59, 54, 124, 36, 121, 110]. While *SOUPS* and *EuroUSEC* are designated UPS venues, UPS is one among many topics for the other venues. This mix of venues results in a broad overview with diverse insights from different perspectives such as privacy, security, HCI, and UPS. We excluded journals because most UPS research is published in conference proceedings. We also excluded workshops, as they often publish only short papers (i.e., potentially less room for details).

4 Call for Papers Analysis

To better understand external transparency requirements for UPS publications and uncover potential influence factors on reporting transparency, we analyzed the CFPs and CfAs of the 12 considered venues (Section 3). We focused on factors that might support or constrain transparent research reporting, such as submission requirements, page limits, and AE.

4.1 Methodology

We analyzed 81 CFPs from 2018–2024 for the 12 venues listed in Table 1. We included the 2024 CFPs because those were already published, in contrast to some 2024 proceedings. Two researchers jointly analyzed each CFP (and any linked submission guidelines). We focused only on the main technical track if multiple tracks with different CFPs existed. If the venue had AE, we also analyzed the CfA ($n = 20$), as AE can support transparency and reproducibility. Our analysis considered factors that might promote or hinder transparent research reporting. We evaluated if CFPs include requirements for transparency and reproducibility, calls for replication papers, page limits, and whether reviewers are expected to review appendices and supplementary materials. For CfAs, we recorded whether AE is optional, for which papers artifacts can be submitted, and whether badges and awards are assigned.

¹For EuroUSEC we consider only 2021 and later, as EuroUSEC turned from workshop to symposium in 2021.

4.2 Results

Table 1 gives an overview of the CfP analysis results. We provide detailed results for each year and venue as supplementary material (see [Open Science](#)).

Transparency & Reproducibility Requirements. In 2024, 9 of 12 venues had requirements for transparency or reproducibility in some form. Some venues are explicit, like CHI stating: “*accepted papers will excel in [...] validity [and] reproducibility*” [86]. Others are more indirect, such as CCS stating to “*promote the reproducibility*” [2] through introducing AE. Overall, we see a substantial increase in venues expressing the importance of transparency and reproducibility. In 2018, only ICSE requested transparency as “*sufficient information for the results to be independently replicated*” [46]. Most venues have since included transparency or reproducibility requirements in their CfPs, suggesting that the communities increasingly prioritize transparency and move towards formally supporting transparent reporting.

Replication Studies. Although replication is closely related to transparency and reproducibility, we found little encouragement for authors to submit replication studies to the analyzed venues. Only the dedicated UPS venues, SOUPS and EuroUSEC, explicitly call for replication studies in their CfPs since at least 2018; CHI and USENIX Security introduced this recently in 2024 [86] and 2025 [9]. This absence does not mean that other venues do not accept replication studies. For ICSE and CHI, we found public instructions for reviewers to consider replication studies for publication. However, the scarcity of explicit calls for replication studies highlights that, despite advancements in reproducibility standards, replication efforts remain undervalued in much of the research landscape.

Page Limits. Many venues impose page limits on publications to encourage focused and concise reporting, with unclear consequences for transparency. Most security venues have separate page limits for submissions and for camera-ready versions. Page limits are often distinguished between the main part/body, typically 12 or 13 pages,² and appendices and references, which are often unlimited in submissions, but more commonly restricted for camera-ready versions. The total allowed length for camera-ready papers ranges between 12 and 20 pages. Only CSCW and CHI impose no strict limits, as the CSCW CfP states: “*While no minimum or maximum length is imposed [...] Reviewers will be instructed to weigh the contribution of a paper relative to its length if that exceeds 15 pages.*” [35]. That said, the limit is softer and suggests a range of typically expected length. Similarly, CHI evaluates the length concerning a paper’s contribution: “*Papers, where the length is incommensurate with contributions, will be desk rejected.*” [86]. We observed few changes in page limits since 2018. WWW and EuroS&P introduced page limits in 2022, and PETS/PoPETs reduced the main body limit by 3 pages in

²Specific page limits are not directly comparable among venues due to requiring different paper templates, resulting in different lengths (in pages).

2023 when starting self-publishing and switching templates. Overall, page limits are inconsistent across venues (although not directly comparable due to different templates), raising questions about the impact of space restrictions on transparent reporting.

Artifact Evaluation. Eleven of 12 venues do not require reviewers to consider appendices, supplementary materials, or artifacts, except ICSE. In 2019, ICSE was the first venue in our analysis to introduce AE. In the following years, other venues followed, until in 2024, 6 of 12 venues offered AE. Currently, AE is mostly optional, although often encouraged. Moreover, a lack of AE does not imply that artifacts are not encouraged, e.g., EuroS&P has no AE, but authors are still expected to publish artifacts [48]. In the end, AE results in assigning artifact badges to a paper (e.g., *Available*, *Functional*, *Reusable*, or *Results Reproduced* [1]). However, not all do so for the officially published version. For example, CCS 2023 performed AE after the papers’ camera-ready deadline, so badges could only be added to the author’s version—not the version published in the ACM Digital Library (ACM DL). While CHI has no formal AE, they grant *Artifact Available* badges [5]. Besides badges, four of six venues present awards for outstanding artifacts. Last, we note that the descriptions of AE mainly target technical artifacts, like data analysis scripts or software implementations, but rarely encourage non-technical artifacts that are typical for UPS research (e.g., survey instruments, interview guides, experimental materials). While the trend towards more AE illustrates increased attention and effort towards better availability of materials, transparency, and reproducibility, it is not currently geared towards UPS specifically, as UPS researchers pointed out [59]. Hence, the actual effects of the current mainly voluntary AE format are unclear.

Venues Foster or Constrain Transparency. Venues impact transparency through CfP requirements and reviewing. In recent years, they increasingly required transparency and reproducibility and introduced AE. However, page limits, lacking calls for replication, and optional reviewing of appendices and supplementary materials potentially disincentivize transparent reporting.

5 Systematic Literature Analysis – Method

We systematically analyzed a random sample of 200 UPS papers (stratified by venue and year) from twelve relevant venues between 2018–2023 to investigate reporting transparency. To facilitate a regression analysis (Section 6.4), we estimated sample size using *a priori* power analysis (Section 5.2). Finally, we calculated a transparency score (TS) for each paper and performed the regression to identify influence factors. Below, we report how we collected papers and analyzed their transparency among 52 criteria, which we identified based on an extensive literature review (Section 5.3.1).

Table 1: Results from the CfP analysis (including CfAs if applicable) for 12 UPS-related venues for 2018–2024. The table presents the state for 2024, while we provide information on past changes in braces, e.g., year of AE introduction.

Venue	CfP					AE/CfA			
	Transparency/Reproducibility ¹	Calls for Replication	Reviewing Appendices ²	Page Limit Subm. ³	Page Limit CR ³	Offers AE	Mandatory	Badges	Awards
IEEE S&P	○	○	○	13/ 5/18	13/ 5/18	○	—	—	—
CCS	● ('23)	○	○	12/ ∞/ ∞	—/—/15 [†]	● ('23)	○	●	● ('24)
USENIX Security	○	○	○	13/ ∞/ ∞*	13/ 5/18	● ('20)	○	●	● ('22)
NDSS	○	○	—	13/—/ ?	?/ ?/ ?	● ('24)	○	●	●
EuroS&P	● ('22)	○	○	13/ ∞/ ∞	?/ ?/ ?	○	—	—	—
PETS/PoPETS	● ('22)	○	○	12/ ∞/ ∞*	12/ ∞/ ∞*	● ('22) [¶]	○	●	●
CHI	● ('21)	● ('24)	○	● [‡]	● [‡]	○	—	● ('24) [§]	—
CSCW	● ('20)	○	○	—/—/ ∞	—/—/ ∞	○	—	—	—
SOUPS	● ('22)	●	○	12/ ∞/ ∞	12/ 8/20	○	—	—	—
EuroUSEC ⁴	●	●	○	16/—/ ?	?/ ?/ ?	○	—	—	—
ICSE	●	○ ('21)	● ('23) ^Δ	10/ 2/12 [#]	11/ 2/13 [#]	● ('19)	○	●	●
WWW	● ('19)	○	○	8/ ∞/ ∞	8/ 4/12	● ('23)	○	● [◊]	○

● Yes. ● Partially. ○ No. ¹ CfP requires transparency or reproducibility. ² Whether reviewers are required to review appendices and supplementary materials material. ³ Page limit for submission and camera-ready versions; *[main part] / [refs. + appendices] / [total]*. ? indicates no publicly available information. ⁴ Only analyzed 2021 and later after EuroUSEC turned from conference to symposium. Revisions can submit one page more of main text. [†] Not from CfP. [‡] No limit, must fit the overall paper contribution (typically 7,000–8,000 words; >12,000 might result in desk reject). [§] Only *Artifact Available* badge, is granted without formal evaluation. ^{||} Length must fit contribution, if >15 pages. [¶] Started 2020 as an experiment. [#] Main part limit includes appendices. ^Δ One reviewer checks the artifacts. [◊] Only *Artifact Available* badge.

5.1 UPS Literature Dataset

We constructed a dataset of 957 UPS papers by extending Klemmer et al.’s UPS dataset (only papers from 2018–2022) [59, 58] with newer papers: We collected the 2023 proceedings and included all papers related to UPS.

DBLP. Like the original dataset, we leveraged the bibliographic metadata provided by the *dblp computer science bibliography (DBLP)* [22] (as of 2024-02-01). DBLP provides high-quality bibliographic metadata independent of individual publishers. Within the DBLP database, we filtered by venues and years (Section 3) to retrieve all publications. This yielded 13,806 publications, with 2,979 new ones from 2023.

Inclusion & Exclusion Criteria. To make the inclusion and exclusion decision transparent and consistent with the original dataset, we leveraged Klemmer et al.’s definition [59]: “*We consider a publication a usable security and privacy paper, if it (1) covers the topics security and/or privacy, and (2) is human subjects research. [...]*”. The latter is typically fulfilled for common UPS methods like interviews, surveys, experiments, user evaluations, or other methods in which humans are the primary data source. Moreover, we only included full papers (e.g., from the main/technical track), and excluded SoKs, short papers, posters, and (extended) abstracts.

We based our assessment on the title and abstract. If those did not allow for a clear decision, the researchers inspected the papers’ full text. To validate the inclusion/exclusion assessment, three researchers reviewed 100 random publications, labeling papers as being a UPS paper or not based on the above criteria. They agreed in 97 of 100 cases, resulting in *almost perfect* [64] inter-rater reliability (IRR) (Fleiss’ $\kappa = 0.89$). Due to the high agreement and the decision re-

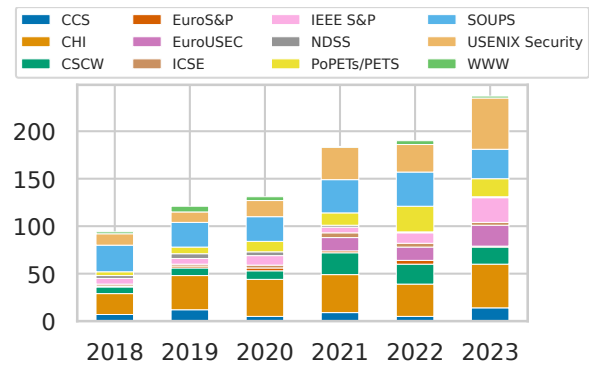


Figure 1: Published UPS papers per year in our dataset.

quiring only little interpretation [68], we split the remaining dataset between three researchers to identify UPS papers.

Final UPS Literature Dataset. Overall, we identified 237 new UPS papers according to the above inclusion/exclusion criteria, resulting in a total of 957 UPS papers between 2018–2023. This corresponds to 6.94% of the initial 13,806 papers from DBLP. Table 2 shows that, although the relative share of UPS papers at larger venues like CHI and USENIX Security is relatively low, they contribute significantly to the total number. Furthermore, the overall publication volume increased for the considered years. As illustrated in Figure 1, there has been a general increase in total UPS publications since 2018. Additionally, shifts in composition are apparent, with USENIX Security capturing an increasing share of UPS papers, while CHI and SOUPS exhibit modest growth in UPS papers. The list of all 13,806 publications, including the marked 957 UPS papers, is available in our supplementary material (see Open Science).

Table 2: Number of UPS papers in our dataset and sample.

Venue	All	UPS		Sample		
		#	% All	% Venue	#	%
CHI	4,387	217	1.6	4.9	33	16.5
WWW	2,031	19	0.1	0.9	11	5.5
CSCW	1,642	86	0.6	5.2	17	8.5
USENIX Security	1,298	157	1.1	12.1	24	12.0
CCS	1,238	52	0.4	4.2	12	6.0
ICSE	939	19	0.1	2.0	12	6.0
IEEE S&P	710	65	0.5	9.2	15	7.5
PoPETs/PETS	536	81	0.6	15.1	16	8.0
NDSS	514	16	0.1	3.1	10	5.0
EuroS&P	258	13	0.1	5.0	10	5.0
SOUPS	187	182	1.3	97.3	32	16.0
EuroUSEC	66	50	0.3	75.8	8	4.0
Total	13,806	957	6.9	n/a	200	100.0

5.2 Sampling

Since we analyzed paper content, which involved an extensive manual review, we decided to investigate a subsample (stratified by venue and year). Analyzing all 957 UPS papers would be an immense effort due to the substantial content analysis time required per paper (typically 30–120 minutes depending on the paper’s write-up and its number of studies).

To estimate the sample size, we conducted a power analysis for our sampling design to achieve a statistical power of 0.8 for testing the effects (coefficients) of venue, year, and study type in our regression analysis (cf. Section 6.4). The required sample size to achieve this power was determined through a Monte-Carlo simulation [10, 63]. Based on this simulation, for $n = 200$ papers, we can reliably detect effects increasing or decreasing the TS (cf. Section 6.3) by at least 10% between years, 15% between venues, and 17.5% between study types. Consequently, we have less power if effects are smaller, but consider this an acceptable trade-off, as we are interested in major effects. We provide details on the sample size estimation online (cf. Open Science section).

We randomly sampled 200 papers for the literature analysis (20.9% of the 957 UPS papers), stratified by year and venue. Proportionally allocating our total sample size across these strata ensures that our sample mirrors the distribution of all UPS papers in terms of venues and years [96]. We note that proportionality is not perfect, as we can only sample whole papers (requiring integer sample sizes)³ and need at least two papers per stratum for valid inference.

³To avoid changes to the overall sample size due to simple rounding, we applied the following: Round each (fractional) stratum target to the nearest integer. If this does not meet the required sample size, pick a stratum closest to being rounded in the needed direction (up or down) and round in that direction. We repeated this until all stratum sizes were integers. If the total sample size was off by 1, we increased or decreased the smallest or largest stratum, respectively.

5.3 Transparency Analysis

For our systematic analysis, we reviewed each paper and analyzed its content for 52 transparency criteria, the degree to which the corresponding information and materials are available or not (availability rating), and the reporting locations (e.g., in main paper, appendix, or online artifacts).

5.3.1 Transparency Criteria

For the analysis, we were interested in information and artifacts that authors should provide to support transparency. We call these *transparency criteria*. Three researchers compiled a list of 52 transparency criteria in an extensive literature review of related work as well as research reporting and transparency guidelines across fields (e.g., including medicine and other disciplines) to ensure comprehensiveness and relevance of the criteria. We carefully considered whether and how criteria from other disciplines apply to UPS (e.g., excluding criteria on biospecimen). In total, the 52 transparency criteria are based on 56 guidelines and other publications. Table 3 contains all transparency criteria along with the related literature references. Overall, we considered criteria in several areas, including ethics, study design, sample & recruitment, instruments & artifacts, data analysis (qualitative and quantitative), and results reporting (qualitative and quantitative).

In addition to the criteria, we apply an availability rating to analyze *whether* and to *what degree* the information or artifacts for a criterion are available. These include *available* and *unavailable*, but also *partially available* (e.g., when only a subset of survey or interview questions is available). We subdivided some availability ratings and obtained seven sub-ratings as shown in Table 5, e.g., to investigate the reasons for unavailability, such as dead URLs of online artifacts.

Finally, we specified six tags for the reporting location of criteria to investigate *where* authors provide information and artifacts. This includes in the publication itself within the *main paper* body or in an *appendix*, but also online artifacts, including *publisher materials* (i.e., artifacts that the publisher hosts alongside the publication) and author-hosted materials (i.e., on an author’s website). We distinguish the latter based on the usage of Digital Object Identifiers (DOIs) into *external materials with DOI* (e.g., typical when using OSF or Zenodo) and *external materials without DOI* (e.g., the authors’ websites). Additionally, we consider it a (pseudo) location when authors state to provide further information or artifacts *upon request*. We did not consider talks and slide decks as external materials, as those are not available for all conferences and often do not contain additional information.

Both availability level and location further included a *does not apply* tag to indicate that a specific criterion did not apply to a given paper and its methodology. We provide the list of all availability ratings and location tags online (Open Science). To validate our analysis approach, three researchers independently analyzed ten initial papers, then met to discuss

Table 3: List of all 52 transparency criteria, including a short description and references on which the criteria are based. We also present how many studies a criterion was not applicable to (n/a) and the availability for applicable studies.

Criteria	Short Description ¹ & References	n/a ²	Availability		
			●	◐	○
Ethical Considerations		—	53.7%	3.2%	43.1%
Ethics	Are ethical aspects discussed? [89, 87, 97, 23, 103, 125, 78, 77]	1.6%	41.8%	4.9%	53.3%
IRB	Is IRB/ERB approval described or is it justified why it was not obtained/granted? [95, 89, 87, 97, 24, 59]	2.0%	77.5%	1.7%	20.8%
Consent	Is described how consent was obtained? [95, 60, 24, 59]	13.4%	62.6%	3.3%	34.1%
Consent Form	Is a consent form provided? [59]	19.8%	11.3%	1.0%	87.7%
Study Compensation	Is participants' compensation explained? [95, 14, 60, 97, 24, 60, 59]	17.0%	69.3%	3.5%	27.2%
Vulnerability Disclosure	Is the vulnerability disclosure process described? [59]	90.1%	77.3%	0.0%	22.7%
Anon	Are any data anonymization or pseudonymization techniques reported? [95]	27.3%	53.4%	5.1%	41.6%
Study Design		—	62.6%	6.5%	30.9%
RQ	Are research questions or research goals stated? [89, 87, 57, 97, 50, 38, 103, 16, 125, 44, 77, 12, 59]	0.4%	95.6%	2.4%	2.0%
Method Choice	Is a justification for the method choice provided? [4, 87, 57, 38, 39, 77, 59]	0.4%	93.1%	4.0%	2.8%
Limitations and Validity Threats	Are limitations or threats to validity described? [87, 57, 97, 23, 105, 53, 50, 125, 44, 118, 78, 77, 12, 112, 71, 59]	0.8%	86.7%	3.2%	10.1%
Prereg	Is study pre-registration reported? [95, 89, 76, 87, 119, 94, 12, 59]	2.0%	1.2%	0.0%	98.8%
Positionality	Is a positionality statement provided? [4, 125, 77, 111, 59]	2.8%	4.1%	3.7%	92.2%
Study Protocol	Is the study protocol (i.e., a description of all steps and their order) available? [95, 89, 14, 87, 14, 119, 97, 74, 119, 39, 24, 125, 118, 77, 111, 59]	0.8%	97.1%	2.0%	0.8%
Study Piloting	Is information on piloting or testing the study protocol and resulting changes available? [59]	12.6%	41.9%	5.1%	53.0%
Study Duration	Is the duration of the study provided? [59]	9.1%	73.5%	9.0%	17.5%
Study Context	Is the context of the study described? (e.g., time of execution, setting, circumstances) [57, 97, 74, 99, 57, 125, 118, 111, 59]	0.8%	61.1%	27.9%	11.1%
Condition Assignment	Is condition assignment described (e.g., grouping, randomness, blinding)? [95, 89, 87, 52, 57, 60, 97, 38, 50, 53, 125]	66.8%	83.3%	11.5%	5.1%
Sample & Recruitment		—	66.1%	5.0%	28.9%
Sampled Population	Is the population that was analyzed/sampled from stated? [57, 16, 44]	1.6%	91.8%	2.5%	5.8%
Sample Description	Are sample characteristics described, e.g., participant demographics? [95, 89, 122, 60, 97, 99, 16, 125, 118, 77, 12, 112, 71, 59]	0.8%	85.2%	6.6%	8.2%
Sampling Procedure	Are the recruitment/sampling procedures reported, including exclusion and inclusion criteria? [4, 89, 87, 14, 57, 60, 97, 24, 50, 105, 99, 16, 125, 118, 77, 111, 12, 112, 71, 59]	1.2%	89.8%	5.3%	4.9%
Sampling Materials	Are the recruitment materials (texts, scripts, etc.) available? [59]	13.8%	8.1%	4.3%	87.6%
Sampling Success Rate	Is the success rate of sampling (e.g., invitations sent, participants, drop-outs, no-shows, etc.) provided? [14, 57, 50, 125, 118, 111]	33.2%	38.9%	12.3%	48.8%
Sample Size	Is the sample size reported? [95, 89, 87, 14, 57, 16, 118, 111, 112, 71]	0.4%	98.8%	0.8%	0.4%
Sample Size Justification	Is the sample size justified (e.g., with saturation, power-analysis)? [20, 4, 89, 87, 14, 52, 57, 99, 75, 125, 118, 12, 112, 71]	6.3%	30.7%	5.6%	63.6%
Instruments & Artifacts		—	43.5%	8.5%	47.9%
Experiment Materials	Are experimental materials available? [95, 89, 60, 97, 74, 119, 38, 39, 99, 16, 125]	61.7%	47.8%	31.5%	20.7%
Survey	Are questionnaires or surveys available? [95, 115, 119, 50, 77, 59]	38.7%	73.6%	5.4%	20.9%
Interview Guide	Is the interview guide available? [95, 4, 119, 77, 111, 59]	63.6%	66.3%	8.1%	25.6%
Share Data	Is the raw study data shared? [95, 33, 4, 76, 14, 61, 57, 52, 74, 119, 38, 39, 23, 75, 87, 119, 50, 75, 59]	0.4%	15.0%	1.2%	83.7%
Share Software	Is the source code of the software shared/its versions and dependencies specified? [95, 33, 89, 119, 38, 39, 116, 59]	66.8%	37.2%	7.7%	55.1%
Share Hardware	Is the hardware specification shared? [95, 33, 119, 38, 39, 116]	85.0%	65.6%	15.6%	18.8%
Data Analysis		—	41.8%	6.1%	52.1%
Data Preprocessing	Are data cleaning/processing procedures described (e.g., transcription, text processing, outlier treatment)? [14, 74, 119, 50, 77, 112, 71, 59]	5.1%	72.4%	6.9%	20.7%
Data Analysis Software	Is any data analysis code shared or used software described (e.g., version number)? [95, 33, 89, 76, 61, 74, 119, 38, 59]	1.6%	12.7%	5.3%	82.0%
Data Analysis – Qualitative		—	71.5%	10.3%	18.2%
Qual Analysis Method	Is qualitative data analysis approach named or explicitly described? [95, 4, 14, 78, 77, 111, 59]	31.6%	86.7%	6.6%	6.6%
Qual Analysis Reliability	Is inter-rater reliability or another trust measure for qualitative analysis discussed? [68, 57, 50, 99, 77, 59]	32.4%	56.1%	14.0%	29.9%
Data Analysis – Quantitative		—	85.6%	4.4%	10.0%
Quant Analysis Method	Is the statistical procedure for data analysis named and described? [95, 89, 87, 14, 52, 57, 97, 105, 99, 50, 75, 125, 118, 78, 112, 71, 59]	40.7%	94.4%	2.1%	3.5%
Quant Variables	Are the experiment's variables named and defined? [87, 60, 74, 38, 39, 99, 118]	41.5%	93.1%	6.2%	0.7%
Quant Variable Dependencies	Are all variables labeled as (in)dependent?	54.9%	73.1%	9.3%	17.6%
Quant Hypotheses	Are the hypotheses stated? [89, 87, 122, 57, 97, 38, 39, 99, 50, 57, 125, 118, 12]	61.7%	88.0%	2.2%	9.8%
Quant Assumptions	Are required statistical assumptions discussed per test/method? [95, 89, 87, 75]	64.8%	70.4%	1.2%	28.4%
Result Reporting – Qualitative		—	72.0%	8.1%	19.9%
Qual Analysis Codebook	Is the codebook of the qualitative analysis provided? [111, 59]	34.8%	51.3%	15.8%	32.9%
Qual Empirical Evidence	Is evidence for the results (e.g., quotes, field notes, text excerpts, photographs) provided? [77, 111, 95, 59]	32.0%	92.1%	0.6%	7.3%
Result Reporting – Quantitative		—	78.1%	2.6%	19.3%
Quant Descriptive (Central Tendency)	For each key dependent variable on the interval or ratio scale, is the sample's central tendency reported? [95, 89, 87, 57]	55.7%	97.2%	0.0%	2.8%
Quant Descriptive (Variability)	For each key dependent variable is the sample's variability reported? [95, 89, 87]	59.7%	63.5%	2.1%	34.4%
Quant Descriptive (Categorical Data)	For each key dependent variable on the nominal or ordinal scale, are the sample's values reported? [95, 89]	49.0%	97.6%	2.4%	0.0%
Quant Parameters (df)	Are degrees of freedom reported? [95, 89]	69.6%	43.7%	1.4%	54.9%
Quant Parameters (Test Value)	Are the test statistic and all test parameters reported (e.g., F-value)? [95, 89]	62.5%	77.5%	5.6%	16.9%
Quant Parameters (p-Value)	Are p-values reported? [95, 89, 122, 74]	57.3%	96.0%	2.0%	2.0%
Quant Effect Size	For statistical effects, are effect sizes reported? [95, 89, 87, 14, 122, 57, 99, 125, 107]	57.3%	64.4%	4.0%	31.7%
Quant CI	For statistical effects, are their confidence intervals reported? [95, 89, 87, 57, 74, 105, 125, 118, 12, 112, 71]	65.2%	43.2%	4.9%	51.9%
Quant Significant/Insignificant	Are both significant and non-significant results reported? [89, 59]	56.9%	96.1%	2.0%	2.0%
Miscellaneous		—	77.9%	0.4%	21.7%
Author Contact Information	Is contact information for at least one author provided? [125, 111]	0.0%	75.0%	0.4%	24.6%
Conflicts Of Interest and Funding	Are conflicts of interest declared? [89, 87, 125, 44, 118, 112, 71, 78, 77, 12, 112, 71]	0.8%	80.8%	0.4%	18.8%

¹ A detailed description of each criterion is available online (cf. [Open Science](#) section). ² Number of studies where the transparency criteria did not apply.

● Available: Proportion of studies providing this information/materials (ignoring not applicable ones). ◐ Partially Available. ○ Unavailable.

any conflicts and improve analysis criteria, availability ratings, and location tags. This only resulted in minor changes in the definitions of criteria, availability, and locations.

5.3.2 Analysis Procedure

Six researchers analyzed the 200 UPS papers for transparency, using the transparency criteria, availability ratings, and location tags mentioned above. For each paper, two researchers independently evaluated all 52 transparency criteria. First, they determined whether a criterion applied to the publication. If a criterion was not applicable, they assigned the criterion and *does not apply*. If the criterion applied, the researchers assigned it with (1) an availability rating to indicate the extent to which related information/artifacts are available or missing, and (2) a location tag to indicate where it is reported. For papers that included multiple studies (e.g., a survey and an interview study), we analyzed all studies separately.

Many papers did not follow a structured reporting approach, which made it easy to overlook relevant information. Therefore, two researchers analyzed each paper independently to ensure consistent results. Both researchers met to resolve any disagreements to obtain the final analysis results.

From the results, we computed the TS as an indicator of a publication's overall transparency based on the applicable transparency criteria for a paper (Section 6.3). We leverage a regression model to explore factors that influence and explain the TS. See Section 6.4 for details.

6 Systematic Literature Analysis – Results

This section presents the results of our systematic transparency analysis of 200 UPS papers. We report on the analyzed transparency criteria, their availability, and respective locations. Based on our data, we also present a transparency score (TS) to assess the overall reporting transparency of papers. Lastly, we leverage our TS in a regression to identify impacting factors. We provide all obtained data online (cf. Open Science).

6.1 Availability of Transparency Criteria

Overall, we assigned 12,851 availability ratings across all papers analyzed. On average, 65.0% of the papers' applicable criteria are *available* (median: 65.6%, std: 11.2%). At least 21.9% and at most 87.5% of a paper's applicable criteria are *available*.

In the following, we report the availability ratios per transparency criterion. The 200 papers in our sample report on 253 studies. Therefore, we present the availability for each criterion per study rather than per paper. To calculate availability ratios per criterion, we ignored studies to which the criterion did not apply. Moreover, we only considered the highest availability rating for each criterion, i.e., if it was *partially*

available in one location and *available* in another, we only counted the latter. This allows us to depict the completeness of expected information and materials accurately. In total, 63.9% of all criteria were *available*, 5.2% *partially available*, and 30.9% *unavailable*.

6.1.1 Reporting Consistency of Criteria

While some criteria are reported in a large majority of studies or barely at all, for others, their availability varies strongly (cf. availability heatmap in Table 3). We leverage *Shannon Entropy* S to measure consistency for the criteria (higher values indicating inconsistent reporting across practices). Providing experiment materials is the most inconsistent criterion ($S = 1.5$), followed by providing codebooks in qualitative analysis ($S = 1.4$), reporting sampling success rates ($S = 1.4$), or discussing IRR in qualitative analysis ($S = 1.4$). We discuss the individual criteria more deeply in Section 6.1.2. Overall, the high consistency for some criteria supports the existence of implicit community standards for transparency that Klemmer et al. describe [59].

6.1.2 Transparency Criteria & Categories

Below, we provide in-depth insights into all transparency criteria, grouped into seven categories. While we cannot discuss all details due to space constraints, information on all criteria and their availability are contained in Table 3.

Ethical Considerations. This category covers ethics and includes discussions of ethical implications, institutional review board (IRB) approval and consent, the availability of consent forms, the participant compensation, and the description of vulnerability disclosure and anonymization procedures. 90.8% of studies reported at least one ethics criterion, but only 8.4% provided information on all applicable ethics criteria. The most commonly reported criterion was IRB approval at 77.5%. Vulnerability disclosure applied to 9.9% of the studies, and related information was *available* in 77.3% of these. At 69.3% and 62.6% availability, compensation and consent procedure were described in about two-thirds of the studies. Yet, only 11.3% of studies provided a consent form. Finally, descriptions of anonymization of the reported data were *available* in 53.4% of studies, and a discussion of ethical implications of a study appeared in 41.8% of studies.

Study Design. This category encompasses criteria about reporting research questions (RQs) and research goals, method choice, limitations, and study details such as the overall protocol, duration, context, pilot testing, or condition assignment. While RQs (95.6%), method choice (93.1%), and the general study protocol (97.1%) were widely *available*, 10.1% omitted limitations. While it is common to report study duration and context, especially the context is often only partially reported (27.9%), e.g., lacking the year of a study. Reporting on study piloting was missing in 53.0% of cases. It remains unclear

whether authors forgot to report pilot testing, or if many UPS studies are not piloted. We assumed that positionality statements and pre-registration can be applied to most studies to investigate their overall prevalence. However, both are not widely used and reported by the UPS community.

Sample & Recruitment. This category is about a study's sample and how it was obtained. This includes the investigated population, sample characteristics (e.g., demographics), the sampling procedure and materials, the sampling success rate, and finally, the sample size and its justification (e.g., power analysis). Overall, 66.1% of sampling information was *available*, 5.0% was *partially available*, and 28.9% was *unavailable*. However, this varied across criteria. Sample size was the most reported criterion of all categories at 98.8%; sampled population, sampling procedure, and sample description were all provided in more than 85% of studies, indicating that this information is widely expected in the UPS community. In contrast, sampling success rates were described in 38.9% of studies, and sample size justification in 30.7%. Sampling materials were rarely provided (8.1%), showcasing these are not currently viewed as critical information. Merely 10.8% of UPS studies reported on all applicable sampling criteria.

Instruments & Artifacts. This category encompasses the instruments used in a study's execution, including experimental materials, questionnaires, interview guides, and other artifacts used, studied, or produced during the research (e.g., datasets, software artifacts, and hardware specifications). Only 10.7% provided all applicable materials. The share of *partially available* instruments and artifacts is relatively high (8.5%) among categories, indicating a tendency to share excerpts. We found differences among study methods: While about two thirds of survey instruments (73.6%) and interview guides (66.3%) were provided, only 47.8% of experimental materials were fully *available*. However, 31.5% of experimental materials were *partially available*, e.g., as screenshots. This was less common for interview guides (8.1%) and surveys (5.4%), indicating that their size and format may make it easier to provide them. For artifacts, only 15.0% of studies dealt with hardware artifacts, which most (65.6%) fully described, and many (15.6%) partially specified, showcasing a trend to share this information, but not all details. Software artifacts were only *available* in 37.2% of cases and *partially available* (i.e., as code excerpts or UI screenshots) in 7.7% of cases. Finally, 15.0% shared their data or justified why this was not possible, typically due to PII concerns. This demonstrates that publishing datasets is not common practice in UPS research.

Data Analysis. We distinguish (1) qualitative and (2) quantitative data analysis. However, two criteria apply to both: 20.7% of studies did not report on any data pre-processing, e.g., outlier handling or transcription—leaving open whether and how any pre-processing was applied. Even though the data analysis software, including authors' scripts, creates and impacts results, it was *unavailable* for 82.0% of studies.

(1) *Qualitative:* 6.6% of qualitative studies reported their analysis method only shallow or in part (e.g., how coding worked, or the codebook was created), and entirely lacked this information in some cases (6.6%). While missing in 29.9% of cases, many authors (56.1%) report IRR or disclose why they did not use it, e.g., IRR is not meaningful for all types of qualitative analysis [68]. Overall, 52.0% of studies reported all relevant qualitative analysis criteria.

(2) *Quantitative:* Like qualitative analysis, most authors describe their respective quantitative methods (94.4%), and define variables (93.1%). In contrast, the variable dependencies are not always explicitly reported (17.6% *unavailable*). Similarly, not all studies that test hypotheses state those (9.8% *unavailable*). While essential to assess the soundness, 28.4% did not report on any assumptions made for statistical tests or whether those were fulfilled. 61.2% of studies provided all applicable quantitative data analysis criteria.

Result Reporting. Like for data analysis, we split result reporting into (1) qualitative and (2) quantitative reporting.

(1) *Qualitative:* For qualitative result reporting, the transparency criteria include providing a codebook and empirical evidence for the results (e.g., quotes). 68.4% of the studies reported qualitative results, and the vast majority of them (92.1%) provided empirical evidence. However, only about half (51.3%) of codebooks were fully *available*. About one in six codebooks (15.8%) were *partially available*, meaning that only excerpts or high-level codes were provided. 32.9% of codebooks were entirely *unavailable*.

(2) *Quantitative:* The quantitative reporting category encompasses criteria for descriptive results (central tendency, variability, and categorical data) and statistical results (e.g., p-values, effect sizes, confidence intervals, and significant and non-significant results). 64.4% of all studies reported quantitative results, and 38.7% of those reported on all applicable criteria. In descriptive reporting, providing central tendencies (97.2%), like mean or median, and categorical data (97.6%) was very prevalent, which is expected as these are typically the main results. In contrast, variability (e.g., standard deviation) as a metric that provides further context was missing in 34.4%. For statistical reporting, reporting p-values appears to be a norm that almost all studies (96.0%) follow. Most studies also reported both significant and insignificant results (96.1%), and resulting values of statistical tests (77.5%). Less common was the reporting of effect sizes (necessary to understand the magnitude of an effect, apart from its significance [107]), which were missing in 31.7% of studies, confidence intervals (missing in 51.9%), and degrees of freedom (missing in 54.9%), suggesting that their reporting is not perceived as essential or meaningful in the UPS community.

Miscellaneous. Not all authors provide contact email addresses on their papers, 24.6% only included names and affiliations. This makes it difficult to contact them with questions or requests for sharing materials. In general, reporting con-

flicts of interests (COIs) and funding is common. While 18.8% did not do so, we suspect that this is likely due to not having any funding or COIs to declare.

Reporting is Transparent, but Inconsistent. Our analysis of 52 transparency criteria (Table 3) indicates a clear recognition of research transparency: 63.9% are fully available and 5.2% partially. Nonetheless, there is room for improvement, with 30.9% of information and materials missing. While many criteria are reported consistently (e.g., sample size, study protocol), others are reported inconsistently (e.g., study’s timespan/year, codebooks), and some, even if relevant, are commonly not reported (e.g., study piloting, sharing data, sample size justification).

6.2 Reporting Locations

Most criteria were provided as part of the paper, either in the main body (5,698, 89.5%) or appendix (414, 6.5%). The remaining information and artifacts were provided online. We distinguish publisher materials (i.e., artifacts that the publisher archives and provides alongside a paper) (40, 0.6%), external materials that are assigned and referenced via a DOI (60, 0.9%), and any other external materials without DOI (148, 2.3%). Rarely, authors mentioned that some information or materials are available upon request (6, 0.1%). In 55 instances, a study made criteria *partially available* in the paper’s main body, but then fully *available* in the appendix or online.

Missing Online Artifacts & Dead Links. The analyzed papers pointed to online locations for information or artifacts 248 times. In 14.5% of these cases, what was claimed to be *available* online was in fact *unavailable*. Either the link to the online resources was dead (11.3%), or they did not contain the promised materials (3.2%). Of all online artifacts, publisher-hosted materials had the highest availability (93.3%). Still, they were also the least common, used by only 14 papers, all from the ACM DL.⁴ Still, 7.5% (3) of artifacts promised in publisher-hosted materials were *unavailable* there, indicating a lack of consistency checks. Information and materials published online by the authors are less available, primarily when hosted on a website without a DOI, such as an author’s website, a project website, or a GitHub repository. Of those, 18.9% (28) were *unavailable* because of a dead link, and 1.4% (2) were not part of the provided online resources. This corresponds to 16 papers that contained dead links to online resources. When authors used a hosting platform that supports DOIs (e.g., OSF or Zenodo) and assigned a DOI, all links worked, but 5.0% (3) of materials were *unavailable* in the online resources.

Reporting Locations are Diverse and Impact Transparency. Most information is reported in the papers’ main bodies. Appendices or online hosting are used for artifacts (e.g., questionnaires,

⁴Likely because the ACM is the only publisher among the analyzed venues that offers this option; the IEEE provides a separate platform [47].

codebooks). We identify a notable lack of long-term availability for self-hosted supplementary materials due to dead links and absence of DOIs.

6.3 Transparency Score

While our systematic literature analysis provides detailed insights on 52 individual criteria for transparent research reporting, we were also interested in the broader picture of paper transparency overall. Therefore, we computed a transparency score (TS) for each paper based on our analysis.

6.3.1 Transparency Score Construction

The TS ranges from 0 to 1, where zero indicates that all applicable transparency criteria were unavailable in a paper and one indicates that all were available. To calculate the TS, we ignored criteria that did not apply.

For each applicable transparency criterion, we considered the highest availability (e.g., when some information is *partially available* in the main body but *available* in the appendix, we considered the criteria *available*). The highest availability was mapped to a value between 0 (*unavailable*) and 1 (*available*) as stated in Table 5. For papers with multiple studies, we aggregated the highest availabilities of all sub-studies for each criterion to a mean for the whole paper. For example, if a criterion was *available* (1) in study 1 and *partially available* (0.5) in study 2 of a paper, we used the mean $((1 + 0.5)/2 = 0.75)$. Lastly, we summed the highest availability values and divided them by the number of applicable criteria, i.e., we calculated a mean over all applicable criteria.

6.3.2 Transparency Score Insights

The TS shows that papers have an average transparency of 0.677 ± 0.103 (median = 0.681) that ranges from 0.281 to 0.885. The TS is almost normally distributed.

Overall, the TS indicates a considerable variance in transparency and reporting practices among the papers in our sample. While some papers have very detailed reporting and achieve high transparency of about 85–90%, no paper transparently reported all applicable criteria. The average paper did not report about 32% of the expected criteria.

Transparency can be Further Improved. Using the availability of applicable criteria to calculate an overall transparency score (TS), we find that on average, papers report only about two-thirds of the information and materials needed for full transparency.

6.4 Exploratory Transparency Regression

Below, we describe our regression analysis that explores factors predicting and explaining papers’ TS.

Table 4: Results of the linear mixed effects regression. Significant effects highlighted in bold.

Variable	95% CI
Intercept	0.560 ± 0.107
Venue (Baseline: <i>SOUPS</i>)	
<i>CHI</i>	0.077 ± 0.096
<i>USENIX Security</i>	0.040 ± 0.117
<i>IEEE S&P</i>	0.103 ± 0.104
<i>NDSS</i>	0.124 ± 0.129
<i>ICSE</i>	0.109 ± 0.098
<i>CCS</i>	-0.011 ± 0.113
<i>CSCW</i>	-0.012 ± 0.070
<i>PETS/PoPETs</i>	0.064 ± 0.077
<i>EuroUSEC</i>	0.068 ± 0.101
<i>WWW</i>	0.086 ± 0.112
<i>EuroS&P</i>	0.117 ± 0.087
Year	0.014 ± 0.018
Main Method (Baseline: <i>Interview</i>)	
<i>Survey</i>	0.021 ± 0.090
<i>Experiment</i>	-0.021 ± 0.083
<i>Focus Groups</i>	0.050 ± 0.094
<i>Data Logs</i>	-0.001 ± 0.121
<i>Observation</i>	0.040 ± 0.169
<i>Workshop</i>	-0.006 ± 0.092
<i>Other</i>	0.022 ± 0.124
No. of pages (normalized)	0.061 ± 0.054
No. of methods	-0.024 ± 0.023
Has AE badge	-0.021 ± 0.118
Marginal R ²	0.381

6.4.1 Method and Regression Model

To assess how the TS (our dependent variable) can be explained by a set of independent variables while controlling for all other of these variables, we used a linear mixed model [88]. We considered the following independent variables (see Appendix A for details): *Venue* (reference category: *SOUPS*), *Year*, *Main Method* used in the publication (reference: *interview*), *# Methods* in a publication, *Paper Length* (normalized by length of a reference paper in the respective template), and whether the paper has *AE Badges*. To account for dependencies among papers by the same (senior) author, we included random intercepts for last authors of all papers [42]. The stratified sampling design (Section 5.2) is accounted for in point estimation and inference [65, 66], computing confidence intervals from rescaling bootstrap variance estimates [90]. To quantify the model’s predictive power, we used marginal R² as coefficient of determination [73].

6.4.2 Regression Results

We present the regression results for the TS in Table 4. An effect is marked as significant at a 5% significance level (bolded) if the 95% confidence interval (estimated coefficient ± margin of error) does not contain 0.

The paper length has a significant positive relation to transparency. This is intuitive, as longer papers can fit more content and allow authors to provide more details that might be relevant for transparency. Additionally, the number of methods used in a paper shows a significant association: papers with

more methods tend to report less transparently. Considering that most venues impose page limits, including more methods reduces the space to report on each one. We hypothesize that this prevents the reporting of many details, thereby worsening transparency.

Among the venues, only *ICSE* and *EuroS&P* showed a significantly higher TS compared to the baseline (*SOUPS*). An explanation for the mainly insignificant differences might be that many authors publish at multiple venues, including resubmissions of the same paper at different venues. Based on our data, we cannot infer a conclusive explanation for the higher transparency of *ICSE* and *EuroS&P*. However, *ICSE* and the wider SE community introduced transparency expectations and guidelines earlier than the other venues.

The model indicates no significant effect for the primary study type (baseline: *interviews*). Considering that many papers use a mix of methods (115 papers, 57.5%) or consist of multiple studies (39 papers, 19.5%), the influence of individual methods might be negligible.

The model indicates no significant effect of the publication year. However, considering prior research, like the qualitative interview insights from Klemmer et al. [59] and findings on *CHI* paper transparency [95], a small effect might still exist. A six-year range might be too short to reliably detect a small effect, given that changes in research reporting take time. Nonetheless, we observed a slight improvement over the analyzed years, but also an increase in TS variance (Figure 2).

Regarding *AE*, the model indicates no significant association of awarded badges and TS. However, only a few papers have *AE* badges (8 papers, 4.0%), limiting reliable conclusions due to large error margins. This may be because *AE* is currently voluntary and not yet widespread (Section 4).

Space for Reporting Impacts Transparency. The regression reveals a significantly positive association of transparency with a paper’s length and fewer methods. Most venues do not differ significantly from the baseline (*SOUPS*). The model indicates no significant relation with *AE*, a paper’s primary method, and publication year.

7 Discussion

We start by answering our guiding research questions. Afterward, we derive community recommendations (R1–10), contextualize our findings with related work, discuss limitations, and give an outlook on future work.

RQ1: How transparent is research reporting in UPS? Our meta-study reveals that about 70% of relevant transparency criteria are reported at least partially, indicating that the community cares about transparency—while also leaving room for future improvements. Our TS (mean = 0.677) reflects this, and reveals that no paper reported all applicable criteria. However, considering that our analysis is very detailed with 52 individual transparency criteria, it is unlikely that papers

reach the maximum score of 1.0. Self-hosting artifacts regularly results in dead links, causing artifact unavailability and hindering transparency. Moreover, reporting practices were inconsistent for some criteria, while the community rarely reports others (e.g., justification of sample size).

RQ2: What factors impact transparency and research reporting practices in UPS? We fitted a regression model on our TS to explore factors influencing transparency. We found a significant positive association between transparency and paper length, and a negative one with the number of methods in a paper. Except for ICSE and EuroS&P, no venue was significantly different compared to SOUPS (baseline). While we observe a slight improving trend over time in our sample, the effect on transparency is insignificant.

7.1 Transparency Reporting Practices

Our meta-study on transparency in UPS literature revealed several areas for improvement:

7.1.1 Inconsistency Demands Transparency Guidelines

Overall, papers differ in how and what transparency criteria are reported (Section 6.1). Besides different practices among authors, this highlights that venues and their reviewers do not consistently enforce transparency either, confirming Klemmer et al.’s qualitative results [59]. Further, they found an implicit community standard for transparency. Our findings suggest that an implicit standard exists, leading to inconsistency.

R1: Develop Transparency Guidelines for UPS. We support Klemmer et al.’s [59] recommendation to establish explicit community guidelines on transparency for authors and reviewers, suggesting our 52 criteria as a foundation. Table 3 could serve as an initial checklist that the community can use and build on.

7.1.2 Low Effort Transparency Improvements

We identified two transparency gaps that we believe can be addressed with little effort. This appears especially important given the misalignment between incentives and required effort, a significant obstacle to transparency in UPS [59].

Making all Study Materials Available. We found materials were missing that the authors had already prepared for conducting their studies, and could have provided as is (besides anonymization for review). These include consent forms, recruitment materials, experimental materials, and surveys.

R2: Make Existing Materials Available. Authors can and should improve transparency by providing all materials that they prepared for study execution in the appendix or online.

Transparency Means being Explicit. Some criteria (e.g., compensation, study piloting, data pre-processing, cf. Section 6.1) have relatively high *unavailable* rates that we suspect are in some cases because it was not part of the study. Unfortunately, readers cannot know, e.g., whether participants were not compensated or whether authors did not report it. However, being explicit often requires only one sentence and improves transparency. The reasons for a lack of comprehensive and transparent reporting are unclear. Besides authors forgetting to report a criterion, not having done the respective practice, or lacking the space to report it, another explanation is that the community generally does not report on some criteria. Implicit community standards [59] might influence whether and why authors report on individual criteria. For example, UPS researchers care about data protection but are uncertain how to process personal data [67], and commonly do not publish datasets without explicitly stating so [59].

R3: Practice Explicit Reporting. Authors should make their reporting explicit, e.g., briefly stating if they skipped a common methodology step. Making community guidelines explicit (Rec. 1) can help authors not to forget relevant information.

7.1.3 Better Supporting Transparency

However, not all measures that could significantly boost transparency are low effort. The provision of analysis software and datasets improves transparency, but they are rarely *available*. For example, self-written analysis scripts typically require preparation to make them useful for others, but contain critical information on analysis details and facilitate replication. Similarly, ethically publishing UPS datasets, which often contain PII, is not trivial [59]. Still, there are approaches in, e.g., the social sciences, to address these problems, such as archives with access control protections for sensitive datasets [45].

R4: Develop Author Support for Sharing Data and Scripts. The community should support authors in providing datasets and analysis scripts by developing best practices and tool support.

7.2 External Impact on Transparency

External factors influence the transparency of UPS publications. Below, we discuss their impact and give recommendations for adaptation to support transparent reporting in UPS.

7.2.1 Low Prevalence of Artifact Evaluation

In our sample, only 8 of 200 papers had an AE badge. The negligible effect of AE on artifact availability (Section 6.4) aligns with a meta-study in ML security [80]. Not all venues have established AE, yet—while many did in the last years. And those who did, often only encourage but do not require AE (Section 4). UPS authors, with their (often non-technical) artifacts like interview guides, might hesitate to participate in AE without clear processes and expectations for UPS [59].

R5: Adapt AE for UPS. Venues should consider whether AE is the right approach for UPS or adapt it accordingly.

Venues typically separate paper review and (optional) AE, and reviewers are often not required to review appendices and artifacts (Section 4). However, we argue that a thorough review often requires considering artifacts. For example, it is essential in UPS to review interview guides and survey instruments, e.g., to assess risks of priming or biasing participants.

R6: Review Appendices and Artifacts. The program committee (PC) should review appendices and artifacts if papers rely on them. Authors should provide artifacts for peer-review, not only AE.

7.2.2 Space Constraints can Decrease Transparency

Paper length had a significant positive relation to transparency in our sample (Table 4). However, many venues impose limits on the papers' main parts and appendices, which may require cutting details that would improve transparency. At the same time, researchers do appreciate page limits on the main part to facilitate brevity [59]. The number of methods in a paper had a significant negative effect on transparency. Reporting multiple methods or studies means less space for each, forcing authors to omit important details. Nonetheless, papers with multiple studies or mixed methods might provide richer insights.

R7: Lift Appendix Page Limits. PC chairs should lift or generously increase appendix page limits. Alternatively, they might set no limit and assess whether a paper's length fits its contribution.

Moreover, we suspect that decreasing page limits for the camera-ready versions (as some venues do, cf. Section 4) causes authors to cut appendices. As this occurs after paper acceptance, it may explain why critical transparency criteria are lacking in published papers—while reviewers no longer have the opportunity to express their concerns.

R8: Do not Decrease Page Limits for CR. Venues should not decrease any page limit for the camera-ready versions, compared to the submitted versions. Page limits should at least remain the same to avoid incentivizing the cutting of important details.

7.2.3 Poor Long-Term Artifact Availability

The loss of an essential artifact can render a publication incomplete. Thus, we argue that publications and their artifacts should be published and archived as a unit. Unfortunately, only a few publishers offer such a feature (e.g., ACM DL), and not all venues leverage available options (e.g., CCS 2024).

R9: Offer Publisher Artifact Hosting. Publishers should offer options to publish artifacts alongside publications to ensure accessibility and long-term availability. PC chairs and venues should leverage these options to avoid fragmenting papers and their artifacts.

Author-organized artifacts hosting is necessary if publishers do not offer hosting, but it does not work well: In 16 papers, the links were dead (e.g., to private homepages). Since we analyzed papers since 2018, we expect more artifacts of the 42 papers that are currently accessible without DOIs to become unavailable due to dying links. Researchers in biology have made similar observations, finding that dataset availability decreases by 17% each year [117]. We deem it unlikely that authors themselves can ensure long-term availability, e.g., by self-hosting. Recently, USENIX Security recognized this problem and now forbids self-hosting, requiring the use of Zenodo or similar platforms as part of the mandatory artifact availability evaluation [114].

R10: Use Long-Term Artifact Archival. In the absence of publisher artifact hosting, authors should use DOI-capable archival platforms (e.g., OSF or Zenodo). They should assign and reference artifacts using DOIs. Authors should avoid personal resources (e.g., author's website) or options not intended for long-term archival (e.g., GitHub).

7.3 Contextualization with Prior Work

We cannot statistically confirm an improvement of transparency over time (Section 6.4), that prior studies on CHI [95] and UPS [59] suggest. However, we observe a slight increase in transparency over time in our sample, and believe that transparency in UPS improves slowly.

We can overall confirm findings from literature analyses of IEEE S&P (2015, 2016) and CCS (2015) publications concluding that key information is often missing, but see some differences. Burcham et al. found that about 30% of empirical studies did not report threats to validity [13]. In our more current sample, this rate is considerably lower at 10.1%.

We notice interesting similarities and differences between CHI and UPS papers. In their analysis of CHI papers from 2017 and 2022, Salehzadeh Niksirat et al. found a significant improvement for 12 reporting criteria, but none for their other 17 [95]. Similar to our data, availability was high for study protocols, condition assignment, and sample sizes and descriptions, and low for pre-registration, and data and software sharing. In contrast, interview guides and surveys were shared at much lower rates in CHI papers [95], even though about 25% of them were missing in our UPS sample. Similar to recent CHI papers [81, 83], we can confirm shortcomings in the reporting of effect sizes.

Our results confirm Klemmer et al.'s qualitative insights that research reporting in UPS is already quite transparent, but needs further improvement. For example, many interviewees reported providing materials like questionnaires and interview guides [59], but we find that about 20% of studies lack those.

Intransparency Hinders Meta-Studies and Replication. In UPS, two recent meta-studies explored specific aspects of study design, namely risk representation [24] and participant samples [41]. Both reported uncertainty in their results

caused by a lack of information in the papers they analyzed. For example, Distler et al. described how about one in three papers did not discuss IRB approval, which does not necessarily mean that none was obtained [24]. They proposed a guide on what information on risk representation and related aspects (e.g., ethics and recruitment) authors should include. Nonetheless, we found that UPS papers often lack some of that information, including IRB approval, ethics discussion, and especially consent forms. Moreover, sample descriptions and recruitment are more commonly reported than many other transparency criteria, but are not always complete. Hasegawa et al. note that 32% of their analyzed papers did not report participants' countries, and 14% omitted education level, leaving an incomplete picture of UPS participant samples [41] and showcasing that even *unavailable* rates below 10% can cause problems. Interviewed UPS researchers additionally reported that a lack of transparency had previously hindered them in replication attempts [59].

From our results, we can derive how many papers lack essential information (e.g., interview guides, survey, recruitment methods) required for replication by other researchers. 56 of 200 papers (28.0%) lack essential information or artifacts that are needed for replication. Thus, about every fourth paper might contribute to a replication crisis in UPS, if the authors cannot provide those upon request. This illustrates the problems caused by a lack of research transparency, and highlights the need to increase transparency in UPS.

7.4 Limitations

For the CfP and systematic literature analysis, we focused on various SP, UPS, HCI and related venues (Section 3), of which many are highly ranked [26]. Hence, we may have missed relevant UPS research published in other venues and our results might not generalize to those. However, we argue that we included the most influential venues in our dataset, and can therefore provide comprehensive insights. We limited our literature review to research published between 2018–2023. However, our representative cross-sectional sample provides valuable insights and reflects the transparency practices of the UPS research community for that time frame.

The reported availability might be slightly higher in reality, as one can try to request information and artifacts from a paper's authors. We did not contact authors, as we argue that papers should be transparent on their own. Prior research also suggests that the success rate would be poor [62]. Moreover, 23.0% of papers included no contact email addresses, and requesting information would cause unnecessary work for the respective authors.

7.5 Future Work & Outlook

Our insights represent the state of transparency in UPS for 2018–2023. So far, they do not indicate larger differences

in transparency among venues (Section 6.4), despite minor transparency-related policy changes among the venues in past years (Section 4). However, there were considerable changes recently, e.g., USENIX Security's new mandatory open science policy in 2025 [9]. It would be valuable to repeat our analysis in a few years to monitor the overall transparency development and how these policies impacted transparency. The dataset that we provide can serve as a baseline and might lay the foundation to train and evaluate automated approaches, e.g., to assess transparency and flag issues. Besides preventing extensive manual annotation effort, automated transparency analysis tools might benefit authors as transparency checkers for submissions or assist reviewers with automated checks.

Further opportunities for future work are accessibility, pre-registration, and other scientific quality measures that go hand-in-hand with transparency. Regarding accessibility and language, both are relevant aspects beyond the scope of our analysis. We only focused on whether relevant transparency criteria were available, not on how accessible or understandable they were. Those aspects of research reporting are potential future work. Regarding pre-registration, it adds transparency about data analysis and if it was conducted as initially planned or whether deviations occurred. However, it could face similar transparency problems as the typical post-hoc reporting of papers. Currently, our results suggest that pre-registration is rare in our community. A potential reason is that pre-registration is not appropriate for all study types; it is most beneficial for hypothesis-driven, confirmatory, quantitative research, but not for more exploratory or qualitative research [74]. Therefore, generally requiring pre-registration is not advisable, but it could be encouraged more for respective studies.

8 Conclusion

In a systematic literature analysis of 200 UPS publications from twelve UPS-publishing venues from 2018–2023, we investigated 52 transparency criteria, their reporting locations, and availability. The UPS community already practices transparent reporting for many criteria, with room for further improvements. For example, authors hosting artifacts on personal websites long-term often fails. Authors have different reporting practices on some transparency criteria while being consistent in others. Therefore, we recommend a community guideline. While authors can impact transparency, venues and PC chairs also constrain or foster transparency, e.g., through page limits or publisher hosting of supplementary materials.

Transparency is the first step towards improved reproducibility, a core value of science. The 52 transparency criteria and ten recommendations in this paper can serve as a foundation for authors, reviewers, and PC chairs for improved and transparent research reporting in UPS and beyond.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback and for helping us to improve this paper. We also thank Jannis Rautenstrauch for helpful feedback on a draft of this paper. The Hannover authors want to thank our feel-good manager, Cleo, for being the cutest office dog and always offering emotional support. This research was partially funded by VolkswagenStiftung Niedersächsisches Vorab – ZN3695.

Conflicts of Interest

The authors declare none.

Ethics Considerations

We followed the ethical principles of the *Menlo Report* [56, 55, 6]. Since our analysis does not involve human subjects and only utilized a dataset of publicly available publications, data, and artifacts, we did not obtain ethical review board (ERB) approval. As all papers are public (or can be retrieved from the respective publishers with the necessary subscriptions), we publish our dataset including the analysis results for transparency. Nonetheless, as highlighted in the introduction (Section 1), we do not refer to individual publications or authors in this publication, so as not to call out somebody for poor transparency compared to other papers. Instead, we aimed to establish an overview of the whole community to discuss and reflect on transparency and research reporting. Nonetheless, the dataset that provides analysis results for individual papers is available online (cf. Open Science section).

Open Science

By writing a paper on transparency, we inherently want to encourage and support open science (including USENIX Security’s new open science policy [9]) and thus followed the recommendations that we provided as well. Specifically, we share our dataset as well as analysis results. Please note, that we cannot share the PDFs of papers, because the publications are (in many cases) protected by publishers’ license restrictions that prevent us from distributing copies. Moreover, we share other artifacts; see “Availability” below.

To ensure a transparent and complete reporting for this paper, we followed the *PRISMA* checklist [85, 84]. We note that we adapted the checklist, as we systematically evaluate other paper’s reporting and not their content.

We did not preregister this study.

Availability. We could not fit all relevant details and artifacts into the paper or its appendix. Therefore, we provide the following artifacts online: (1) The detailed results table from the CfP analysis (Section 4), (2) the list of all UPS and non-UPS papers, (3) the main dataset containing all analysis

results for the analyzed sample of 200 UPS papers, (4) details of the sample size estimation, and (5) scripts used in our data analysis to calculate numbers, render figures, and perform regression analysis. The artifacts are available at: <https://doi.org/10.5281/zenodo.15532982>.

References

- [1] ACM. *Artifact Review and Badging*. Version 1.1. Aug. 24, 2020. URL: <https://www.acm.org/publications/policies/artifact-review-and-badging-current> (visited on 04/12/2024).
- [2] ACM CCS 2024. 2024. URL: <https://www.sigmac.org/ccs/CCS2024/call-for/call-for-papers.html> (visited on 04/11/2024).
- [3] Balazs Aczel et al. “A consensus-based transparency checklist”. In: *Nature Human Behaviour* 4.1 (Dec. 2019), pp. 4–6.
- [4] Herman Aguinis and Angelo M Solarino. “Transparency and replicability in qualitative research: The case of interviews with elite informants”. In: *Strategic management journal* 40.8 (2019), pp. 1291–1315.
- [5] *Artifacts at CHI 2024*. Feb. 8, 2024. URL: <https://chi2024.acm.org/2024/02/08/artifacts-at-chi-2024/> (visited on 04/12/2024).
- [6] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. “The Menlo Report”. In: *IEEE Security & Privacy* 10.2 (2012), pp. 71–75.
- [7] Monya Baker. “1,500 scientists lift the lid on reproducibility”. In: *Nature* 533 (2016).
- [8] Nick Ballou, Vivek R. Warriar, and Sebastian Deterding. “Are You Open? A Content Analysis of Transparency and Openness Guidelines in HCI Journals”. In: *Proc. 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. Yokohama, Japan: ACM, 2021.
- [9] Lujo Bauer and Giancarlo Pellegrino. *USENIX Security ’25 Call for Papers*. 2024. URL: <https://www.usenix.org/conference/usenixsecurity25/call-for-papers> (visited on 11/21/2024).
- [10] A. Alexander Beaujean. “Sample Size Determination for Regression Models Using Monte Carlo Methods in R”. In: *Practical Assessment, Research, and Evaluation* 19.12 (2014).
- [11] C. Glenn Begley and John P. A. Ioannidis. “Reproducibility in science: improving the standard for basic and preclinical research”. In: *Circulation Research* 116.1 (Jan. 2015), pp. 116–126.
- [12] Patrick M. Bossuyt et al. “STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies”. en. In: *BMJ* (Oct. 2015), h5527.
- [13] Morgan Burcham, Mahran Al-Zyoud, Jeffrey C. Carver, Mohammed Alsaleh, Hongying Du, Fida Gilani, Jun Jiang, Akond Rahman, Özgür Kafalı, Ehab Al-Shaer, and Laurie Williams. “Characterizing Scientific Reporting in Security Literature: An analysis of ACM CCS and IEEE S&P Papers”. In: *Proc. Hot Topics in Science of Security: Symposium and Bootcamp*. HoTSoS ’17. Hanover, MD, USA: ACM, 2017, pp. 13–23.
- [14] Tim Buthe, Alan M. Jacobs, Erik Bleich, Robert Pekkanen, Marc Trachtenberg, Katherine Cramer, Victor Shih, Sarah Parkinson, Elisabeth Jean Wood, Timothy Pachirat, et al. “Transparency in qualitative and multi-method research: A symposium”. In: *Qualitative and Multi-Method Research: Newsletter of the American Political Science Association’s QMMR Section* 13.1 (2015), pp. 2–64.

- [15] Jeffrey C. Carver, Morgan Burcham, Sedef Akinli Kocak, Ayse Bener, Michael Felderer, Matthias Gander, Jason King, Jouni Markkula, Markku Oivo, Clemens Sauerwein, and Laurie Williams. "Establishing a baseline for measuring advancement in the science of security: an analysis of the 2015 IEEE security & privacy proceedings". In: *Proc. Symposium and Bootcamp on the Science of Security*. HoTSoS '16. Pittsburgh, Pennsylvania: ACM, 2016, pp. 38–51.
- [16] Jeffrey C. Carver, Natalia Juristo, Maria Teresa Baldassarre, and Sira Vegas. "Replications of software engineering experiments". In: *Empirical Software Engineering* 19 (2014).
- [17] Center for Open Science (COS). *TOP Guidelines*. URL: <https://www.cos.io/initiatives/top-guidelines> (visited on 01/30/2024).
- [18] Amelia Chauvette, Kara Schick-Makaroff, and Anita E. Molzahn. "Open Data in Qualitative Research". In: *International Journal of Qualitative Methods* 18 (2019), pp. 1–6.
- [19] Lewis L. Chuang and Ulrike Pfeil. "Transparency and Openness Promotion Guidelines for HCF". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA '18. ACM, 2018, pp. 1–4.
- [20] Juliet Corbin and Anselm Strauss. *Basics of qualitative research*. Vol. 14. sage, 2015.
- [21] Lorrie Cranor, Kim Hazelwood, Daniel Lopresti, and Amanda Stent. *Conference Submission and Review Policies to Foster Responsible Computing Research*. Aug. 2024. URL: <https://cra.org/wp-content/uploads/2024/07/Report-Conference-Submission-and-Review-Policies.pdf> (visited on 08/05/2024).
- [22] *dblp*. 2024. URL: <https://dblp.org/> (visited on 05/14/2024).
- [23] Nurullah Demir, Matteo Große-Kampmann, Tobias Urban, Christian Wressnegger, Thorsten Holz, and Norbert Pohlmann. "Reproducibility and replicability of web measurement studies". In: *Proc. ACM Web Conference 2022*. 2022, pp. 533–544.
- [24] Verena Distler, Matthias Fassel, Hana Habib, Katharina Krombolz, Gabriele Lenzini, Carine Lallemand, Lorrie Faith Cranor, and Vincent Koenig. "A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research". In: *ACM Trans. Comput.-Hum. Interact.* 28.6 (Dec. 2021).
- [25] Florian Echter and Maximilian Häußler. "Open Source, Open Science, and the Replication Crisis in HCF". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA '18. Montreal QC, Canada: ACM, 2018, pp. 1–8.
- [26] Computing Research & Education. *ICORE Conference Portal*. 2023. URL: <https://portal.core.edu.au/conf-ranks/> (visited on 01/14/2025).
- [27] *Empirical Standards*. URL: <https://www.sigsoft.org/EmpiricalStandards/about/> (visited on 01/30/2024).
- [28] EQUATOR Network. *Enhancing the QUALity and Transparency of health Research*. URL: <https://www.equator-network.org/> (visited on 01/30/2024).
- [29] Benjamin Erb, Christoph Bösch, Cornelia Herbert, Frank Kargl, and Christian Montag. *Emerging Privacy Issues in Times of Open Science*. June 2021. URL: <https://doi.org/10.31234/osf.io/u236e>.
- [30] *FAIR Principles*. URL: <https://www.go-fair.org/fair-principles/> (visited on 01/30/2024).
- [31] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Albrecht Schmidt, and Paweł W. Woźniak. "Designing for Reproducibility: A Qualitative Study of Challenges and Opportunities in High Energy Physics". In: *Proc. 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland UK: ACM, 2019, pp. 1–14.
- [32] Sebastian S. Feger, Paweł W. Woźniak, Lars Lischke, and Albrecht Schmidt. "'Yes, I Comply!': Motivations and Practices around Research Data Management and Reuse across Scientific Fields". In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW2 (Oct. 2020).
- [33] Dror G. Feitelson. "From repeatability to reproducibility and corroboration". In: *ACM SIGOPS Operating Systems Review* 49.1 (2015), pp. 3–11.
- [34] Ben G. Fitzpatrick, Elena Koustova, and Yun Wang. "Getting personal with the 'reproducibility crisis': interviews in the animal research community". In: *Lab Animal* 47 (2018).
- [35] Call for papers - CSCW 2024. *Call For Papers*. 2024. URL: <https://cscw.acm.org/2024/index.php/submit-papers/> (visited on 04/12/2024).
- [36] Anjali Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. "SoK: Still Plenty of Phish in the Sea — A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research". In: *Proc. 17th Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX, Aug. 2021, pp. 339–358.
- [37] Jesús M. González-Barahona and Gregorio Robles. "On the reproducibility of empirical software engineering studies based on data retrieved from development repositories". In: *Empirical Software Engineering* 17 (2012).
- [38] Odd Erik Gundersen, Yolanda Gil, and David W. Aha. "On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications". In: *AI Magazine* 39.3 (Sept. 2018), pp. 56–68.
- [39] Odd Erik Gundersen and Sigbjørn Kjensmo. "State of the Art: Reproducibility in Artificial Intelligence". In: *Proc. AAAI Conference on Artificial Intelligence*. Vol. 32. 1. Apr. 2018.
- [40] Martin D. Halbert. "Advancing Reproducibility at the NSF". In: *Computer* 55.8 (2022), pp. 31–39.
- [41] Ayako A. Hasegawa, Daisuke Inoue, and Mitsuoaki Akiyama. "How WEIRD is Usable Privacy and Security Research?" In: *Proc. 33rd USENIX Security Symposium, August 14–16, 2024, Philadelphia, PA, USA*. USENIX, 2024.
- [42] Trevor J. Hastie, Robert J. Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. 2nd ed. New York: Springer, 2008.
- [43] Cormac Herley and Paul C van Oorschot. "SoK: Science, Security, and the Elusive Goal of Security as a Scientific Pursuit". In: *Proc. 38th IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 2017.
- [44] Don Husereau et al. "Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations". en. In: *BMJ* (Jan. 2022), e067975.
- [45] ICPSR. *Guide to Social Science Data Preparation and Archiving. Best Practice Throughout the Data Life Cycle: 6th Edition*. 2020. URL: <https://www.icpsr.umich.edu/files/deposit/dataprep.pdf> (visited on 02/08/2024).
- [46] *ICSE 2018 Technical Papers - ICSE 2018*. 2018. URL: <https://web.archive.org/web/20170903195010/http://www.icse2018.org/track/icse-2018-Technical-Papers> (visited on 04/11/2024).

- [47] IEEE. *Research Reproducibility - IEEE Author Center Conferences*. 2024. URL: <https://conferences.ieeeauthorcenter.ieee.org/write-your-paper/research-reproducibility/#sharingdata> (visited on 12/06/2024).
- [48] IEEE EuroS&P. *Call For Papers*. Feb. 28, 2024. URL: <https://eurosp2024.ieee-security.org/cfp.html> (visited on 04/12/2024).
- [49] John P. A. Ioannidis. “Why Most Published Research Findings Are False”. In: *PLoS Medicine* 2.8 (Aug. 2005), pp. 696–701.
- [50] Andreas Jedlitschka, Marcus Ciolkowski, and Dietmar Pfahl. “Reporting Experiments in Software Engineering”. In: *Guide to Advanced Empirical Software Engineering*. Ed. by Forrest Shull, Janice Singer, and Dag I. K. Sjøberg. London: Springer, 2008, pp. 201–228.
- [51] Andreas Jedlitschka and Dietmar Pfahl. “Reporting guidelines for controlled experiments in software engineering”. In: *4th International Symposium on Empirical Software Engineering*. IEEE, 2005.
- [52] Biophysical Journal. *Guidelines for the Reproducibility of Biophysics Research*. Accessed: 2024-11-08. URL: <https://www.cell.com/pb/assets/raw/journals/society/biophysj/PDFs/reproducibility-guidelines.pdf>.
- [53] Vigdis By Kampenes, Tore Dybå, Jo E. Hannay, and Dag I. K. Sjøberg. “A systematic review of quasi-experiments in software engineering”. In: *Information and Software Technology* 51.1 (2009). Special Section - Most Cited Articles in 2002 and Regular Research Papers, pp. 71–82.
- [54] Mannat Kaur, Michel van Eeten, Marijn Janssen, Kevin Borgolte, and Tobias Fiebig. *Human Factors in Security Research: Lessons Learned from 2008-2018*. 2021. arXiv: [2103.13287](https://arxiv.org/abs/2103.13287) [cs.CY].
- [55] Erin Kenneally and David Dittrich. *Applying Ethical Principles to Information and Communication Technology Research: A Companion to the Menlo Report*. Tech. rep. U.S. Department of Homeland Security, Oct. 2013.
- [56] Erin Kenneally and David Dittrich. *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research*. Tech. rep. U.S. Department of Homeland Security, Aug. 2012.
- [57] Barbara A. Kitchenham, Shari Lawrence Pfleeger, Lesley M. Pickard, Peter W. Jones, David C. Hoaglin, Khaled El Emam, and Jarrett Rosenberg. “Preliminary guidelines for empirical research in software engineering”. In: *IEEE Transactions on software engineering* 28.8 (2002), pp. 721–734.
- [58] Jan H. Klemmer, Juliane Schmäuser, Byron M. Lowens, Fabian Fischer, Lea Schmäuser, Florian Schaub, and Sascha Fahl. *Artifacts of “Transparency in Usable Privacy and Security Research: Scholars’ Perspectives, Practices, and Recommendations”*. May 2024. URL: <https://doi.org/10.17605/OSF.IO/HY7EJ>.
- [59] Jan H. Klemmer, Juliane Schmäuser, Byron M. Lowens, Fabian Fischer, Lea Schmäuser, Florian Schaub, and Sascha Fahl. “Transparency in Usable Privacy and Security Research: Scholars’ Perspectives, Practices, and Recommendations”. In: *46th IEEE Symposium on Security and Privacy, IEEE S&P 2025, May 12-15, 2025*. IEEE, May 2025.
- [60] Amy J. Ko, Thomas D. LaToza, and Margaret M. Burnett. “A practical guide to controlled experiments of software engineering tools with human participants”. In: *Empirical Software Engineering* 20.1 (2015), pp. 110–141.
- [61] Markus Konkol, Christian Kray, and Max Pfeiffer. “Computational reproducibility in geoscientific papers: Insights from a series of studies with geoscientists and a reproduction study”. In: *International Journal of Geographical Information Science* 33.2 (2019), pp. 408–429.
- [62] Michal Krawczyk and Ernesto Reuben. “(Un)Available upon Request: Field Experiment on Researchers’ Willingness to Share Supplementary Materials”. In: *Accountability in Research* 19.3 (2012), pp. 175–186.
- [63] Levi Kumle, Melissa L.-H. Vö, and Dejan Draschkow. “Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R”. In: *Behavior Research Methods* 53.6 (Dec. 2021), pp. 2528–2543.
- [64] J. Richard Landis and Gary G. Koch. “The Measurement of Observer Agreement for Categorical Data”. In: *Biometrics* 33.1 (1977), pp. 159–174.
- [65] Thomas Lumley. “Analysis of complex survey samples”. In: *Journal of statistical software* 9 (2004), pp. 1–19.
- [66] Thomas Lumley and Xudong Huang. “Weighted composite likelihood for linear mixed models in complex samples”. In: *arXiv preprint arXiv:2311.13048* (2023).
- [67] Florin Martius, Luisa Jansen, Lukas Struck, Arthi Arumugam, Lisa Geierhaas, Anna-Marie Orloff, Matthew Smith, and Christian Tiefenau. “Out of Sight, out of Mind? Exploring Data Protection Practices for Personal Data in Usable Security & Privacy Studies”. In: *Proc. 2025 CHI Conference on Human Factors in Computing Systems*. CHI ’25. Yokohama, Japan: ACM, 2025.
- [68] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. “Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice”. In: *ACM on Human-Computer Interaction* 3.CSCW, 72 (2019), pp. 1–23.
- [69] James Miller. “Replicating software engineering experiments: a poisoned chalice or the Holy Grail”. In: *Information and Software Technology* 47.4 (2005), pp. 233–244.
- [70] Aaron Mobley, Suzanne K. Linder, Russell Braeuer, Lee M. Ellis, and Leonard Zwelling. “A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic”. In: *PLoS One* 8 (2013).
- [71] Karel G. M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P. A. Ioannidis, Petra Macaskill, Ewout W. Steyerberg, Andrew J. Vickers, David F. Ransohoff, and Gary S. Collins. “Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration”. In: *Annals of internal medicine* 162.1 (2015), W1–W73.
- [72] Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. “A manifesto for reproducible science”. In: *Nature Human Behaviour* 1.1. 0021 (Jan. 2017).
- [73] Shinichi Nakagawa and Holger Schielzeth. “A general and simple method for obtaining R^2 from generalized linear mixed-effects models”. In: *Methods in Ecology and Evolution* 4.2 (Feb. 2013). Ed. by Robert B. O’Hara, pp. 133–142.
- [74] National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. The National Academies Press, 2019.
- [75] nature publishing group. *Reporting Life Sciences Research*. Accessed: 2024-11-08. 2015. URL: <https://www.nature.com/documents/nr-reporting-life-sciences-research.pdf>.
- [76] B. A. Nosek et al. “Promoting an open research culture”. In: *Science* 348.6242 (2015), pp. 1422–1425. eprint: <https://www.science.org/doi/pdf/10.1126/science.aab2374>.
- [77] Bridget C. O’Brien, Ilene B. Harris, Thomas J. Beckman, Darcy A. Reed, and David A. Cook. “Standards for Reporting Qualitative Research”. In: *Academic Medicine* 89.9 (Sept. 2014), pp. 1245–1251.

- [78] Greg Ogrinc, Louise Davies, Daisy Goodman, Paul Batalden, Frank Davidoff, and David Stevens. “*SQUIRE 2.0 (Standards for QUality Improvement Reporting Excellence): revised publication guidelines from a detailed consensus process*”. en. In: *BMJ Quality & Safety* 25.12 (Dec. 2016), pp. 986–992.
- [79] Babatunde Kazeem Olorisade, Pearl Brereton, and Peter Andras. “*Reproducibility of studies on text mining for citation screening in systematic reviews: Evaluation and checklist*”. In: *Journal of Biomedical Informatics* 73 (2017), pp. 1–13.
- [80] Daniel Olszewski, Allison Lu, Carson Stillman, Kevin Warren, Cole Kitroser, Alejandro Pascual, Divyajyoti Ukirde, Kevin Butler, and Patrick Traynor. “*“Get in Researchers; We’re Measuring Reproducibility”: A Reproducibility Study of Machine Learning Papers in Tier 1 Security Conferences*”. In: *Proc. 2023 ACM SIGSAC Conference on Computer and Communications Security*. CCS ’23. Copenhagen, Denmark: ACM, 2023, pp. 3433–3459.
- [81] Anna-Marie Ortloff, Julia Angelika Grohs, Simon Lenau, and Matthew Smith. “*A Qualitative Study on How Usable Security and HCI Researchers Judge the Size and Importance of Odds Ratio and Cohen’s d Effect Sizes*”. In: *Proc. 2025 CHI Conference on Human Factors in Computing Systems*. CHI ’25. Yokohama, Japan: ACM, 2025.
- [82] Anna-Marie Ortloff and Florin Martius. “*Meta-Science in Usable Security and Privacy and HCF*”. In: *Meta-HCI: First Workshop on Meta-Research in HCI at CHI’25*. Yokohama, Japan, 2025.
- [83] Anna-Marie Ortloff, Florin Martius, Mischa Meier, Theo Raimbault, Lisa Geierhaas, and Matthew Smith. “*Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation*”. In: *Proc. 2025 CHI Conference on Human Factors in Computing Systems*. CHI ’25. Yokohama, Japan: ACM, 2025.
- [84] Matthew J. Page et al. “*PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews*”. In: *BMJ* 372 (2021).
- [85] Matthew J. Page et al. “*The PRISMA 2020 statement: an updated guideline for reporting systematic reviews*”. In: *BMJ* 372 (2021).
- [86] *Papers - CHI*. 2024. URL: <https://chi2024.acm.org/for-authors/papers/> (visited on 04/11/2024).
- [87] N. Percie du Sert, Viki Hurst, Amrita Ahluwalia, Sabina Alam, Marc T. Avey, Monya Baker, William J. Browne, Alejandra Clark, Innes C. Cuthill, Ulrich Dirnagl, et al. “*The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research*”. In: *PLoS Biol* 18.7 (2020), e3000410.
- [88] José Pinheiro and Douglas Bates. *Mixed-effects Models in S and S-PLUS*. New York: Springer Science & Business Media, 2000.
- [89] Eric M. Prager, Karen E. Chambers, Joshua L. Plotkin, David L. McArthur, Anita E. Bandrowski, Nidhi Bansal, Maryann E. Martone, Hadley C. Bergstrom, Anton Bespalov, and Chris Graf. “*Improving transparency and scientific rigor in academic publishing*”. In: *Journal of Neuroscience Research* 97.4 (2019), pp. 377–390.
- [90] John Preston. “*Rescaled Bootstrap for Stratified Multistage Sampling*”. In: *Survey Methodology* 35.2 (2009), pp. 227–234.
- [91] Edward Raff. “*A Step toward Quantifying Independently Reproducible Machine Learning Research*”. In: *Proc. 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.
- [92] Wullianallur Raghupathi, Viju Raghupathi, and Jie Ren. “*Reproducibility in Computing Research: An Empirical Study*”. In: *IEEE Access* 10 (2022), pp. 29207–29223.
- [93] Paul Ralph et al. *Empirical Standards for Software Engineering Research*. 2020. arXiv: 2010.03525 [cs.SE].
- [94] Rolando P. Reyes, Oscar Dieste, Efraín R. Fonseca, and Natalia Juristo. “*Statistical Errors in Software Engineering Experiments: A Preliminary Literature Review*”. In: *Proc. 40th International Conference on Software Engineering*. ICSE ’18. Gothenburg, Sweden: ACM, 2018, pp. 1195–1206.
- [95] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanatham, and Mauro Cherubini. “*Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022*”. In: *Proc. 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. ACM, 2023.
- [96] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.
- [97] Stuart Schechter. *Common Pitfalls in Writing about Security and Privacy Human Subjects Experiments, and How to Avoid Them*. 2010. URL: <https://cups.cs.cmu.edu/soups/2010/howtosoups.pdf>.
- [98] *Security Research Artifacts*. URL: <https://secartifacts.github.io/> (visited on 01/30/2024).
- [99] Mary Shaw. “*Writing good software engineering research papers*”. In: *25th International Conference on Software Engineering*. IEEE, 2003, pp. 726–736.
- [100] Forrest J. Shull, Jeffrey C. Carver, Sira Vegas, and Natalia Juristo. “*The role of replications in Empirical Software Engineering*”. In: *Empirical Software Engineering* 13.2 (Jan. 2008), pp. 211–218.
- [101] Iveta Simera, David Moher, Allison Hirst, John Hoey, Kenneth F. Schulz, and Douglas G. Altman. “*Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network*”. In: *BMC Medicine* 8.1 (Apr. 2010).
- [102] Iveta Simera, David Moher, John Hoey, Kenneth F. Schulz, and Douglas G. Altman. “*A catalogue of reporting guidelines for health research*”. In: *European Journal of Clinical Investigation* 40.1 (2010), pp. 35–53.
- [103] Ananta Soneji, Faris Bugra Kokulu, Carlos Rubio-Medrano, Tiffany Bao, Ruoyu Wang, Yan Shoshitaishvili, and Adam Doupé. “*“Flawed, but like democracy we don’t have a better system”: The Experts’ Insights on the Peer Review Process of Evaluating Security Papers*”. In: *2022 IEEE Symposium on Security and Privacy (SP)*. 2022, pp. 1845–1862.
- [104] Philip B. Stark. “*Before reproducibility must come preproducibility*”. In: *Nature* 557.7707 (May 2018), pp. 613–613.
- [105] Gretchen A. Stevens, Leontine Alkema, Robert E. Black, J. Ties Boerma, Gary S. Collins, Majid Ezzati, John T. Grove, Daniel R. Hogan, Margaret C. Hogan, Richard Horton, et al. “*Guidelines for accurate and transparent health estimates reporting: the GATHER statement*”. In: *The Lancet* 388.10062 (2016), e19–e23.
- [106] Jeffrey R. Stevens. “*Replicability and Reproducibility in Comparative Psychology*”. In: *Frontiers in Psychology* 8 (2017).
- [107] Gail M. Sullivan and Richard Feinn. “*Using Effect Size—or Why the P Value Is Not Enough*”. In: *Journal of Graduate Medical Education* 4.3 (Sept. 2012), pp. 279–282.
- [108] Jacques Suray, Jan H. Klemmer, Juliane Schmüser, and Sascha Fahl. *How the Future Works at SOUPS: Analyzing Future Work Statements and Their Impact on Usable Security and Privacy Research*. 2024. arXiv: 2405.20785 [cs.CR].
- [109] Jacques Suray, Jan H. Klemmer, Juliane Schmüser, and Sascha Fahl. “*Poster: Future Work Statements at SOUPS*”. In: *20th Symposium on Usable Privacy and Security (SOUPS ’24)*. Philadelphia, PA, USA: USENIX, 2024.

- [110] Mohammad Tahaei and Kami Vaniea. “A Survey on Developer-Centred Security”. In: *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. 4th European Workshop on Usable Security, EuroUSEC 2019. IEEE, Aug. 2019, pp. 129–138.
- [111] Allison Tong, Peter Sainsbury, and Jonathan Craig. “Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups”. In: *International journal for quality in health care* 19.6 (2007), pp. 349–357.
- [112] *TRIPOD Checklist: Prediction Model Development and Validation*. 2015. URL: <https://www.tripod-statement.org/wp-content/uploads/2020/01/Tripod-Checklist-Prediction-Model-Development-and-Validation-PDF.pdf> (visited on 11/08/2024).
- [113] US National Institutes of Health (NIH). *Enhancing Reproducibility through Rigor and Transparency*. 2023. URL: <https://grants.nih.gov/policy/reproducibility/index.htm> (visited on 01/22/2024).
- [114] *USENIX Security '25 Call for Artifacts*. 2025. URL: <https://www.usenix.org/conference/usenixsecurity25/call-for-artifacts> (visited on 06/02/2025).
- [115] Judy Van Biljon. “A critical review on the reporting of surveys in transdisciplinary research: A case study in Information Systems”. In: *TD: The Journal for Transdisciplinary Research in Southern Africa* 7.2 (2011), pp. 337–350.
- [116] Erik van der Kouwe, Gernot Heiser, Dennis Andriess, Herbert Bos, and Cristiano Giuffrida. “SoK: Benchmarking flaws in systems security”. In: *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2019, pp. 310–325.
- [117] Timothy H. Vines, Arianne Y.K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. “The Availability of Research Data Declines Rapidly with Article Age”. In: *Current Biology* 24.1 (2014), pp. 94–97.
- [118] Erik von Elm, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gøtzsche, and Jan P. Vandenbroucke. “The Strengthening of Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies”. In: *The Lancet* 370.9596 (2007), pp. 1453–1457.
- [119] Chat Wacharamanatham, Lukas Eisenring, Steve Haroz, and Florian Echtler. “Transparency of CHI Research Artifacts: Results of a Self-Reported Survey”. In: *Proc. 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: ACM, 2020, pp. 1–14.
- [120] Chat Wacharamanatham, Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. “Special Interest Group on Transparent Statistics Guidelines”. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA '18. ACM, 2018, pp. 1–4.
- [121] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L. Mazurek, Manya Sleeper, and Kurt Thomas. “SoK: A Framework for Unifying At-Risk User Research”. In: *Proc. 2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 2344–2360.
- [122] Tracey L. Weissgerber, Vesna D. Garovic, Stacey J. Winham, Natasa M. Milic, and Eric M. Prager. “Transparent reporting for reproducible science”. In: *Journal of Neuroscience Research* 94.10 (2016), p. 859.
- [123] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.1 (Mar. 2016).
- [124] Yuxi Wu, W. Keith Edwards, and Sauvik Das. “SoK: Social Cybersecurity”. In: *Proc. 2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1863–1879.
- [125] Christina Yap et al. “Enhancing reporting quality and impact of early phase dose-finding clinical trials: CONSORT Dose-finding Extension (CONSORT-DEFINE) guidance”. In: *BMJ* 383 (2023).

A Regression Factors

In our regression model, we consider the following independent variables:

- **Venue:** We explore the differences between venues. Being the largest UPS-specific venue, we use SOUPS as the reference category (baseline) for comparison.
- **Year:** To explore the development of transparency over time, we consider the publication year of a paper.
- **Main Method:** In our systematic literature analysis, we tagged each paper with the methods that it entails. From these, we assigned the dominant method as the paper’s main method and use this information to evaluate the transparency differences between study types. The reference category here is “Interview.”
- **# Methods:** We consider the number of methods in a paper as a potential factor.
- **Paper Length:** We also consider the paper length. For each paper, we extracted the length in pages from the PDF metadata. As venues use different and change templates, we compiled a reference paper (i.e., always having the same content) in all \LaTeX templates used at the different venues throughout all relevant years. Using these reference lengths, we normalized each paper length accordingly to allow an objective comparison among venues.
- **AE Badges:** Finally, we include papers’ AE badges as a regression factor. This is a binary variable, distinguishing papers with no AE badge and papers with at least one, using no AE badge as the reference category.

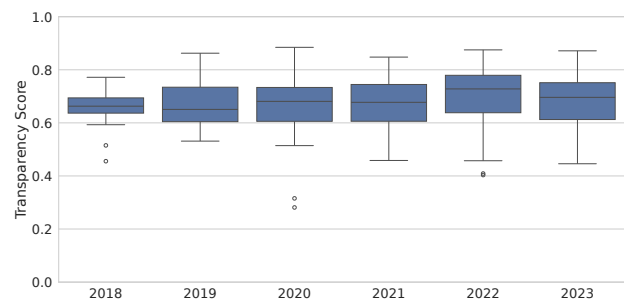


Figure 2: Transparency score over the years.

Table 5: Overview of availability ratings. We provide finer-grained distinctions for some. Each of those is assigned a value for the calculation of the transparency score. The value is used in the transparency score calculation.

Availability	Description	Value
Available		
Available	The material/information is available.	1.00
Unavailable, but justified	The material/information is unavailable, but a justification for why it was not made available is provided. (For the transparency score, we count this as available because the authors were transparent by providing a justification.)	1.00
Available, but not readable with free software	A file, which allegedly contains the materials/information, is provided but cannot be opened with free software.	0.75
Partially Available	Only parts of the material/information are available. (We do not further assess the degree of availability, as it is not always apparent how much is missing.)	0.50
Unavailable		
Unavailable	The material/information is generally unavailable, and the unavailability is not justified.	0.00
Not findable	The material/information could not be found because is not available, even if the paper claims so.	0.00
Link dead	Availability of materials/information online is claimed, but the provided link is not working.	0.00
Does not apply	The criterion does not apply to the study.	n/a

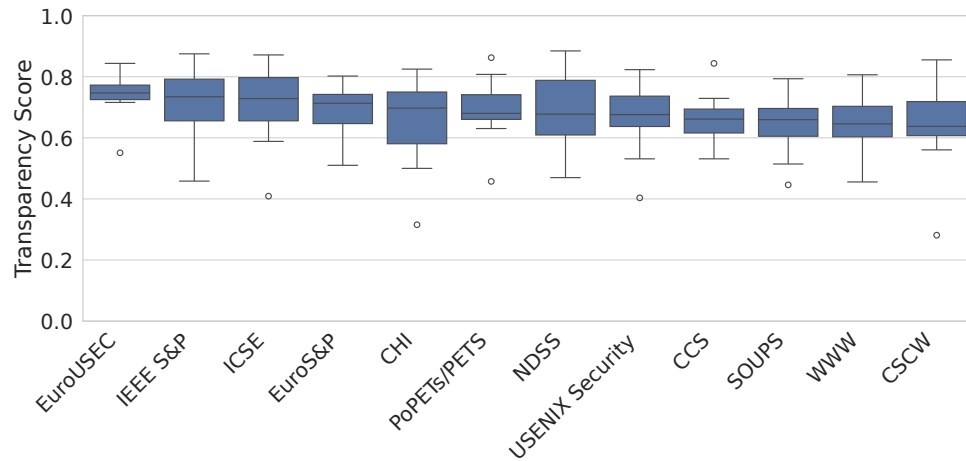


Figure 3: Transparency score among venues.



USENIX Security '25 Artifact Appendix: How Transparent is Usable Privacy and Security Research? A Meta-Study on Current Research Transparency Practices

Jan H. Klemmer ^C Juliane Schmüser ^C Fabian Fischer ^C Jacques Suray ^C
Jan-Ulrich Holtgrave ^C Simon Lenau ^C Byron M. Lowens ^{*} Florian Schaub [†]
Sascha Fahl ^C

^CCISPA Helmholtz Center for Information Security, Germany

^{*}Indiana University Indianapolis, USA

[†]University of Michigan, USA

A Artifact Appendix

A.1 Abstract

To support open science and transparency, we provide several artifacts that accompany our paper. The artifacts contain the data used for our analysis (including the raw analysis results, as well as computed data like the transparency score), analysis scripts to compute the transparency score, and results such as tables and figures for the paper, and the R code for the regression, as well as the generated outputs. Moreover, we include the PDFs that detail the results of our sample size estimation.

Please refer to the contained `README.md` for a more detailed overview of which artifacts are contained in which folders and files.

A.2 Description & Requirements

A.2.1 Security, privacy, and ethical concerns

Except for installing dependencies, the software does not receive/send Internet data. All computations are done locally. The code does not execute any destructive steps. There are no security, privacy, or ethical concerns or risks for the evaluators.

A.2.2 How to access

The artifacts are hosted at Zenodo and accessible via the following link: <https://doi.org/10.5281/zenodo.15532982>.

A.2.3 Hardware dependencies

None.

A.2.4 Software dependencies

The contained source code can be executed on any system that fulfills the following requirements:

- Python with Poetry as dependency manager (Python dependencies are listed in `code/pyproject.toml`). Poetry can be downloaded here: <https://python-poetry.org/>.
- Docker (Installation: <https://docs.docker.com/engine/install/>).

A.2.5 Benchmarks

Our dataset with analysis results is contained in the artifacts and does not need to be installed separately.

A.3 Set-up

Please confirm that the required software packages are installed on the system as stated above (i.e., Python, Poetry, and Docker), before proceeding with the following steps.

A.3.1 Installation

1. Visit <https://doi.org/10.5281/zenodo.15532982> and download the artifacts. Under the tab "Files" you can choose to download the `.zip` file.
2. Unarchive the `.zip` file on your local computer.
3. Run the following commands to install Python dependencies: `cd code/` and `poetry install`
4. For the regression that we implemented in R, we are using a Docker container. To build the docker container,

execute the following commands: `cd code/R/` and then `./build-container.sh`.

A.3.2 Basic Test

- For the Python scripts: Open a shell (from the `code/` directory) in the virtual environment with `poetry shell`. Then execute `jupyter notebook`. This should open Jupyter in your browser. (If that is not the case, you can manually navigate your browser to <http://localhost:8888/>). Try to open one of the notebooks (`.ipynb`).
- For the Docker container: `docker run -it --rm containr:latest /bin/bash` (opens a shell).¹

A.4 Evaluation workflow

A.4.1 Major Claims

(C1): Transparency Criteria Availability: For 52 transparency criteria, we quantify the transparency in terms of availability. On average, $65.0\% \pm 11.2$ of the papers' applicable criteria are available (cf. Section 6.1); in total, 63.9% of all criteria were available, 5.2% partially available, and 30.9% unavailable. We present the detailed availability for all 52 transparency criteria in Table 3.

(C2): Transparency Score: To assess the overall transparency of papers, we calculate a transparency score (TS). The TS per paper is on average 0.677 ± 0.103 (cf. Section 6.3.2).

(C3): Regression: The exploratory regression on the TS reveals a significantly positive association of transparency with a paper's length and fewer methods. Most venues do not differ significantly from the baseline (SOUPS). The model indicates no significant relation with AE, a paper's primary method, and publication year. (Exact results contained in Table 4 and Section 6.4.2.)

A.4.2 Experiments

(E1): [Transparency Analysis] [15 human-minutes + 0 compute-hour + <1GB disk]: Process the result dataset obtained from the manual paper annotation and calculate availabilities, transparency scores, etc.

Preparation: Open a shell (from the `code/` directory) in the virtual environment with `poetry shell`. Then execute `jupyter notebook`. This should open Jupyter

in your browser. (If that is not the case, you can manually try to navigate your browser to <http://localhost:8888/>).

Execution: First, open `transparency-score.ipynb` in Jupyter and execute all cells (“Run”, “Run All Cells”). This will transform the raw annotation results stored in `transparency-analysis.tsv`, calculate the transparency score and store all results in `results.tsv`. Next, open `main.ipynb` in Jupyter and execute all cells (“Run”, “Run All Cells”). Based on `results.tsv`, this notebook contains the actual analysis, e.g., to generate some of the figures and tables.

Results: The overall availability (C1) is presented in cells 19 and 20 of `main.ipynb`, and the table written to `output/tables/criteria.tex`. The results of the transparency score (C2) are written to `data/result.tsv` (column: “Transparency Score”), and the descriptive statistics are in cell 10.

Apart from that, all computed numbers are also stored (for usage in L^AT_EX) as key-value-pairs in `numbers-generated.tex`.

(E2): [Regression] [15 human-minutes + 30 compute-minutes + 6GB disk]: Run the regression model (and related other R code).

Preparation: Build the docker container as described above.

Execution: Run `docker run -it --rm -v $(pwd)/data:/data containr:latest /bin/bash` (from the `code/` directory) to open a shell in the docker container and mount your local `data/` directory. Once you entered the container environment, run `./main` to start the analysis.

Results: You can now retrieve all results in the `data/` folder on your local machine (or inside the container at `/data`). This also includes the regression table (C3).

A.5 Notes on Reusability

Our artifacts operate on the contained dataset. Therefore, the code is reusable on other datasets containing transparency analysis data (assuming that they are in the same format as our dataset). For example, future work might analyze the same transparency criteria, store the transparency analysis data in a similar `.tsv/.csv` file, and run our code to compute transparency scores.

A.6 Version

Based on the L^AT_EX template for Artifact Evaluation V20231005. Submission, reviewing and badging methodology followed for the evaluation of this artifact can be found at <https://secartifacts.github.io/usenixsec2025/>.

¹For Mac/Apple Silicon it might be necessary to setup emulation first.

Option A): Enable Emulation in Docker Desktop (should be enabled by default): `docker.desktop settings` → `general` → `virtual machine options`. Choose `Apple Virtualization framework` and activate `Rosetta`.

Option B): Set Platform Explicitly When Pulling or Running: If Docker does not find an ARM version of an image, it may fail unless you explicitly request the amd64 version: `docker run --platform linux/amd64 <image>`