Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice

NORA MCDONALD, Drexel University, USA
SARITA SCHOENEBECK, University of Michigan, USA
ANDREA FORTE, Drexel University, USA

What does reliability mean for building a grounded theory? What about when writing an auto-ethnography? When is it appropriate to use measures like inter-rater reliability (IRR)? Reliability is a familiar concept in traditional scientific practice, but how, and even whether to establish reliability in qualitative research is an oft-debated question. For researchers in highly interdisciplinary fields like computer-supported cooperative work (CSCW) and human-computer interaction (HCI), the question is particularly complex as collaborators bring diverse epistemologies and training to their research. In this article, we use two approaches to understand reliability in qualitative research. We first investigate and describe local norms in the CSCW and HCI literature, then we combine examples from these findings with guidelines from methods literature to help researchers answer questions like: "should I calculate IRR?" Drawing on a meta-analysis of a representative sample of CSCW and HCI papers from 2016-2018, we find that authors use a variety of approaches to communicate reliability; notably, IRR is rare, occurring in around 1/9 of qualitative papers. We reflect on current practices and propose guidelines for reporting on reliability in qualitative research using IRR as a central example of a form of agreement. The guidelines are designed to generate discussion and orient new CSCW and HCI scholars and reviewers to reliability in qualitative research.

CCS Concepts: • **Human-Centered Computing** → HCI design and evaluation methods; HCI theory, concepts, and models

KEYWORDS

Qualitative methods; interviews; content analysis; inter-rater reliability; IRR

ACM Reference format:

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.*, No. CSCW, Article 72 (November 2019), 23 pages. https://doi.org/10.1145/3359174

1 INTRODUCTION: TO IRR OR NOT TO IRR

"Should we calculate inter-rater reliability (IRR)?"

Researchers have asked this question many times in the course of designing studies, and its sequel, "should they have calculated IRR?" while conducting reviews. Qualitative researchers who use methods that are not compatible with IRR sometimes find themselves defending their methods

Author addresses: Nora McDonald, nkm39@drexel.edu, Drexel University, 3141 Chestnut St., Philadelphia, PA, 19104, USA; Sarita Schoenebeck, yardi@umich.edu, University of Michigan, 3376 North Quad, 105 S. State St., Ann Arbor, MI, 48109, USA; Andrea Forte, aforte@drexel.edu, Drexel University, 3141 Chestnut St., Philadelphia, PA, 19104, USA.

2573-0142/2019/November - 72 $15.00

https://doi.org/10.1145/3359174

to reviewers who have been trained to expect it. The authors of this paper themselves have been challenged repeatedly to explain alternate approaches to reliability and have discussed whether IRR is appropriate in conversations online, in personal communications, in reviews, and in dissertation defenses. These conversations are laborious, but they encapsulate tensions implicit in scholarly communities like CSCW and HCI that unite diverse research traditions. We begin our investigation of reliability in qualitative CSCW and HCI literature with this deceptively simple question because answering it requires a thoughtful exploration of what reliability means in different methodological traditions. We show that CSCW and HCI qualitative researchers use the same terms and concepts in multiple, complex ways, and that readers and authors themselves may have little consensus about what was done and why.

IRR is a statistical measure of agreement between two or more coders of data. IRR can be confusing because it merges a quantitative method, which has roots in positivism and objective discovery, with qualitative methods that favor an interpretivist view of knowledge. Further, while there are many guidelines available across disciplines for *how* to do IRR, there are few guidelines for deciding *when* and *why* to do IRR.

In our informal discussions with CSCW and HCI researchers about this work, we have observed passionate and thoughtful perspectives on IRR and the concept of reliability more generally, but these perspectives are diverse and sometimes contradictory. In general, diversity is a celebrated hallmark of CSCW and HCI—Dourish noted that HCI is a "discipline that has often proceeded with something of a mix-and-match approach, liberally and creatively borrowing ideas and elements from different places" [28]. It is unsurprising, then, that researchers who write and evaluate descriptions of qualitative methods often face quandaries, like whether or not IRR is appropriate and how to communicate their choices to a diverse audience. Similar confusion has been reported and addressed by Caine when it comes to choosing sample sizes in HCI [18].

In this article we first examine practices in recently published papers in the major CSCW and HCI publishing venues, the CSCW and CHI conferences[2], to create transparency around local norms for communicating reliability and related concepts in qualitative scholarship. The same diversity that makes these communities challenging to navigate methodologically also position them to provide guidance. Our research found that over 1/3 of CHI and almost 1/2 of CSCW papers from 2016 to 2018 used qualitative analysis as a primary method, highlighting the need for shared understandings of how to write about and evaluate qualitative research [74]. We found that IRR is relatively rare, but that most papers used some form of agreement to signal reliability. Based on foundational methods texts and illustrated by our findings, we propose guidelines for deciding when IRR is appropriate. This work contributes:

Conceptual work: We situate reliability and IRR in a wide spectrum of related practices and concepts, including rejection of the term reliability altogether, to widen our discussion of methods and approaches beyond our starting point of "Should we use IRR?"

Descriptive norms: We characterize how qualitative researchers in the CSCW and HCI community communicate reliability in their publications. Understanding local standards can both reveal misperceptions about community expectations and enable reflection on our practices as a community by newcomers and established researchers alike.

[2] We adopt the shorthand of these communities and refer to the Association for Computing Machinery (ACM) Special Interest Group on Computer-Human Interaction (SIGCHI) Conference on Computer-Supported Cooperative Work and Social Computing simply as *CSCW* and the ACM SIGCHI Conference on Human Factors in Computing Systems as *CHI*. Note that the CSCW conference proceedings became part of the journal series *Proceedings of the ACM: HCI* in 2017.

Guidelines for deciding when agreement and/or IRR is not desirable (and may even be harmful): The decision not to use agreement or IRR is associated with the use of methods for which IRR does not make sense. Pragmatic examples include when developing codes is part of the process, when there is a single researcher, when researchers are embedded in the research context, where analysis is driven by participants' own interpretations of their data, or when coding requires little interpretation.

Guidelines for deciding when agreement or IRR is useful: The use of agreement or IRR should be consistent with methodological choices and analytical goals, which might include ensuring consistency across multiple coders, for applying existing codebooks, and when researchers aim to report quantitative results.

2 RELATED LITERATURE

2.1 Qualitative Research in CSCW and HCI

Qualitative data is often gathered in "natural" [37], or non-experimental, settings, and tends to involve analysis of verbal responses, observed actions, videos, photos, documents, drawings, or other materials that provide data on how people make meaning in their lives. In CSCW and HCI, qualitative methods became common as researchers aimed to develop human-centered research practices that integrated and privileged the perspectives of system "users."

Creswell categorizes qualitative research in five classical traditions of inquiry: biography, phenomenology, grounded theory, ethnography, and case study [25]. These traditions are not mutually exclusive, and many researchers draw on multiple approaches, particularly in interdisciplinary fields like CSCW and HCI. Pinelle and Gutwin demonstrate that qualitative research was the predominant method in evaluations of groupware systems at CSCW from 1990-1998 [78]. However, Wallace et al.'s study of CSCW from 1990-2015 [88] note that mixed methods research became increasingly common at CSCW. Our work takes up where Wallace et al. leaves off, focusing on the importance of orienting both qualitative and mixed methods scholars to issues of reliability in their methods. Qualitative inquiries rely on overlapping sources of data and methods of analysis. For example, any of the five traditions above could draw from interview data, and most could also draw from field notes, audio and video recordings, and textual or visual content [25]. However, there can be substantial differences in data analysis processes and products. For example, ethnographic research may yield a detailed descriptive narrative or "thick description" [35], whereas thematic analysis may yield a lexicon of themes that help explain the phenomenon of interest [15]. The term "coding" generally refers to the practice of examining data and labeling units of text with descriptive text, or "codes." However, the terms "coding" and "codes" are sometimes used to describe a process of inductive interpretation—which in some cases denote the process of interpretation, and in others, the outcome of interpretation—and at other times are used to describe a process of labeling data with preexisting codes followed by interpretation. The term "theme" is typically applied to describe recurring topics or meanings that represent a phenomena; themes need not align with the most prevalent set of codes but instead those that are salient to the research question or inquiry.

Qualitative analysis tools like NVivo, Dedoose, Atlas.ti, and others can also encourage different approaches to qualitative research through both the kinds of functionality they support and the variable ease with which different types of features can be learned and used. Moreover, in CSCW and HCI, still more methodological variation has been introduced by the need to adapt approaches from other fields and contexts to the pace and structure of technology development and

publication expectations [45,70]. In general, even researchers who use traditional forms of qualitative methods and have a strong understanding of data sources and analysis may have a less firm grasp of how and whether reliability and agreement fit in [3,19,30,53].

The next sections present the concept of reliability and its role in qualitative research, then critique its limitations and potential harms.

2.2 Reliability in Qualitative Research

Reliability and validity are features of empirical research that date back to early scientific practice. The concept of reliability broadly describes the extent to which results are reproducible, for example, from one test to another or between two judges of behavior [29]. Whereas reliability describes the degree to which a measure produces the same answer, validity is the degree to which a measure gives the correct answer. When measuring human behaviors, beliefs, or interactions, correspondence between an instrument (for example, an IQ test) and the construct it measures (intelligence) is always an approximation. As a result, social science has relied heavily on reliability, as "perfect validity" is neither theoretically possible, nor necessarily desirable [53].

Kirk and Miller illustrated different conceptions of reliability in qualitative research: quixotic reliability (when a single method produces the same measure), diachronic reliability (the "stability of an observation through time"), and synchronic reliability (the "similarity of observations within the same time period") [53]. Krippendorff proposed three types of reliability: replicability of results across coders (i.e., IRR), stability or consistency of a single coder's use of codes over time, and accuracy of an established coding scheme compared with others [55]. These definitions underscore the complexity of reliability, and corresponding struggles in social science disciplines to translate and adapt an evolving concept into their own fields [30].

IRR is a statistical measurement designed to establish agreement between two or more researchers coding qualitative data. Calculating IRR does not generate data used in results, but instead provides an artifact and a claim about the process of achieving researcher consensus [43]. Common statistical calculations used in IRR include Cohen's kappa (κ), Krippendorff's alpha (α), Fleiss' kappa (κ), or correlational measures such as Pearson's r. This paper is not focused on the strengths and weaknesses of each approach, which can be reviewed elsewhere (e.g., [42]), though there is general consensus that simply calculating percent agreement between coders is not acceptable because it does not account for the possibility of agreement by chance. Scholars have argued for the use of IRR where "diverse confirmatory instances in qualitative research lend weight to findings" [3]. Krippendorff argues that "a research procedure is reliable when it responds to the same phenomena in the same way regardless of the circumstances of its implementation" [55]. In this light, measurements are not only a critical research tool to guard against systematic bias [65] but can also reveal weaknesses in coding definitions, overlaps in meaning [64], or the challenges of arriving at consensus given the nature of the data [43].

Approaches to establishing and communicating reliability vary and many disciplines have wrestled with how and whether to establish reliability or to use IRR [3,19]. Scholars in some subdisciplines of psychology, sociology, and communication may expect a formal codebook, the use of multiple coders, and formal measures of agreement using IRR [19,30,63]. In contrast, other scholars may rely on a variety of techniques (e.g., interviews, surveys, etc.) and forms of documentation (e.g., photographs, video, etc.) in addition to their field notes, but rarely rely on IRR [61]. Further, debates persist around whether reliability should be understood as a matter of agreement (e.g., over substantial discussion among researchers) or disagreement (e.g., where differences are discussed, and possibly arbitrated, after a round of separate coding) [16] and

whether these agreements or disagreements need to be measured (i.e., IRR). Indeed, the process of reaching or failing to reach agreement may be more important than its measurement. These tensions may be explained by the subtle differences between agreement (where two or more coders reconcile differences through discussion) versus establishing reliability (where two or more coders independently apply the same code to a unit of text).

Some scholars insist that both agreement and reliability are required [19]. However, other techniques for communicating reliability include member checking (e.g., confirming/reviewing results with participants or other members of the community), re-interviewing participants, triangulation of data with secondary data sources (e.g., interviews plus field notes, photographs, etc.), making research process transparent [3], and communicating positionality of the researchers. Many qualitative researchers reject validity and reliability altogether in favor of concepts like dependability, confirmability, credibility, and transferability that resonate with interpretivist accounts [41,62].

2.3 Critiques and Limitations of Reliability: to IRR is Human

In the social sciences, quantitative researchers have sometimes made "the mistake of evaluating qualitative research reports using the standards of quantitative research," expecting IRR regardless of the nature of the qualitative research [30]. As a result, reporting statistical measures may be alluring for qualitative researchers who believe that reviewers who are unfamiliar with their methods will respond to IRR as a signal of reliability; however, for many methods, reliability measures and IRR don't make sense. They may even be outright harmful. These potential harms can take place during the data collection or data analysis process.

Data may not be easily segmented into units of analysis, and how researchers segment data can alter interpretations [56]. Some data, such as raw field notes, may be only interpretable by the researcher who conducted the observations, or they may be interpreted differently by different people depending on researchers' backgrounds and the focus of the analysis [31]. In phenomenological research, the focus of study is on first-person experiences and requires that researchers "bracket" their own interpretations [83]. Scholars have generally argued that "words may have multiple meanings, may be open to interpretation, and may only be understood in the context of other words, which in a way makes them harder to work with than numbers" [19].

Further, aspiring to achieve reliability might reduce sensitivity to complex concepts and nuances in data. In fact, we would argue that in any qualitative research, rigid expectations of reliability should be scrutinized for potential marginalization or minimization of perspectives. For example, feminist HCI emphasizes that knowledge is socially constructed and cannot be disentangled from power and identity [6]. Bardzell points out that the privileging of alternative epistemologies taken up by feminist theory implies a new domain of research—the "marginal" perspective [6]. In practice, this may mean shifting towards participatory design and other community-focused methods such as member checking or collaboration with participant researchers [68], who bring distinct perspectives and interpretations to the research. These new approaches are essential because, as Bellini et al. argues, "[d]espite our best intentions, we may fail to consider the extent of oppressive systems or our complicity within them" [8].

Researchers may also be overly reliant on quantitative outputs—e.g., kappa or alpha values—as evidence of reliability. In fact, most IRR statistical calculations yield numerical values that are subject to interpretation and rely on acceptable levels of agreement that have been normatively accepted without strong empirical evidence. Landis and Koch's 1977 scale (0.0-0.2 Slight; 0.21-0.4 Fair; 0.41-0.6 Moderate; 0.61-0.8 Substantial; 0.81-1 Almost Perfect) has been cited roughly 46,000

times on Google Scholar, but they acknowledge that their scale is "clearly arbitrary" and provided merely as "useful benchmarks" [57]. Krippendorff's inspection of the tradeoffs between statistical techniques establishes that "to assure that the data under consideration are at least similarly interpretable by two or more scholars… it is customary to require $\alpha \geq .800$. Where tentative conclusions are still acceptable, $\alpha \geq .667$ is the lowest conceivable limit" [55]. Although statistical measures can help confirm that interpretations are consistent between coders, they are not a substitute for interpretation and making meaning from the data.

In this paper, we argue that approaches to establishing and measuring reliability should be aligned with epistemological traditions the researcher draws on. An epistemology is a theory of knowledge describing a set of assumptions about what is possible to know and how we communicate that knowledge. Wittingly or unwittingly, researchers invoke such assumptions when they make and describe methodological choices. Burrell and Morgan organized assumptions underlying social science research along two dimensions: assumptions about the nature of social science and assumptions about the nature of societies [17]. They emphasized the need for researchers to be consistent in their assumptions in order to do cogent social science. By understanding their own assumptions, CSCW and HCI researchers can make sound and justifiable decisions about how (and whether) to establish and report reliability.

The wide range of disciplinary traditions infused in CSCW and HCI research can introduce uncertainty about when and how to analyze qualitative data, but also provides a rich dataset for understanding expert practice. In the next section, we present findings from an analysis of three years of published work in CSCW and CHI to describe current norms around communicating reliability and methods in qualitative research.

3 STUDY DESIGN

This research draws from and triangulates multiple sources of data. We conducted informal discussions throughout the data collection, analysis, and writing process with CSCW and HCI researchers, ranging from senior scholars to new graduate students, and across the expanse of disciplines typically found at CSCW and CHI. We reviewed scholarship on qualitative methods, including foundational texts on grounded theory, ethnography, phenomenological research, and feminism, as well as textbooks and recent methods-related works in CSCW, HCI, and adjacent fields. The empirical data presented here comes from a systematic review of research papers from the 2016-2018 CSCW (Computer-Supported Cooperative Work and Social Computing) and CHI (Human Factors in Computing) conferences. Our inclusion criterion for papers was the use of qualitative methods for the primary analysis. Our dataset thus contained a wide range of methods, including ethnographies, interview studies, diary studies, user studies, and design-based research.

Our research team has conducted numerous qualitative research studies using multiple sources of data and a variety of approaches to reliability, and has reviewed hundreds of studies in CSCW, CHI, and related venues. This project was born out of our desire to advance rigor, consistency, and disciplinary sensitivity in our own work and the CSCW and HCI community more broadly.

3.1 Dataset

We conducted a qualitative analysis of full paper and note proceedings (hereafter referred to as "papers") from CSCW and CHI 2016-2018. We used the ACM DL to collect metadata for all 617 CSCW papers and 1811 CHI papers from 2016-2018 and selected a random subsample of 250 CSCW and 400 CHI papers for the first pass of coding to identify qualitative papers. We chose

three years to obtain a robust sample, while prioritizing recent practices in the community. We used Cochran's sample size formula to calculate minimum sample sizes of 237 and 317, respectively, based on a 95% confidence level and a 5% margin of error [90]. Some researchers have published multiple papers that appear in our dataset; we did not adjust for these dependencies, which may introduce a bias towards highly productive individuals and institutions who may conduct research (and train large numbers of students) in unique ways.

We conducted a content analysis of methods sections in those 650 papers to determine eligibility for inclusion in our dataset. We coded papers as *include* if they used qualitative methods as a primary method of analysis. These were often exclusively qualitative papers (e.g., an interview study, ethnographic work, or content analysis) and were sometimes mixed methods studies with qualitative methods playing an integral role in the study (e.g., survey study + interview study). Papers coded as *partial* included qualitative methods as a secondary form of analysis, such as a systems paper that included a user study at the end that qualitatively asked users about their reactions to the system or described qualitative research that shaped the design of their system. Papers coded as *exclude* did not contain any qualitative methods. Note that we were concerned with methods, not data; papers that used quantitative methods to analyze qualitative data (for example using machine learning techniques to analyze texts) were coded as *exclude*. Papers coded as *partial* or *exclude* were not included in the final dataset.

The research team reviewed and discussed criteria for inclusion before beginning the coding process. Because the inclusion criteria could be interpreted differently by different coders, and because we wished to divide the dataset between two different coders, we calculated IRR for the application of inclusion criteria. Two of the co-authors coded the first 30 papers in our random sample of CHI papers and had 100% agreement between them. This suggested that our discussion had yielded a shared understanding of the criteria for inclusion, so both coauthors proceeded to code an additional 70 papers, which produced minor disagreement. We calculated IRR between the two coders for those 70 papers. Both Cohen's kappa and Krippendorff's alpha are appropriate for use when there are two coders coding the same dataset and the data are nominal. Both of these methods yielded an acceptable level of agreement (Cohen's kappa unweighted=0.849, $p<.05$; Krippendorff's alpha=0.85). There were no disagreements between inclusion and exclusion; the seven disagreements were primarily between *partial* and *exclude*. We discussed and resolved those disagreements and divided the remaining dataset for coding between the same two coders. Four additional papers were subsequently recoded from include to *partial* (3) or *exclude* (1) during this pass. The final dataset comprised 140 CHI papers (35%) and 121 CSCW papers (48%) coded as *include* for a total of 261 qualitative papers.

3.2 Content Analysis

We coded our dataset for the presence of language that described how researchers approach reliability, who is involved, the use of IRR, and other methodological and epistemological considerations. The CHI dataset represents a broader research community, into which CSCW fits as a specialized, partially overlapping conference. Beginning with the CHI dataset, two authors each took part of the corpus and extracted methods-related text from every paper that described qualitative data collection and analysis procedures. They also took notes during the data capture pass. After reading methods sections—or if there were no methods (or similarly titled) sections, areas where methods were discussed throughout the paper—the whole research team discussed the data and notes to identify and refine important codes in the data. During these discussions, we generated a set of terms that indicated the presence or absence of certain practices—e.g.,

met/meet, discuss, agree, refine, consensus, member (e.g., "member checking" "team members"), revise, collaborate/collaboration (e.g., "collaboratively identified themes"), triangulate/ triangulation, and compare/comparison—and then examined texts with those key words to confirm use of language of agreement.

We coded the following items: specific data collection activity or methods (e.g., interviews, observations, diary study, etc.), number of researchers/coders involved in analysis, use of IRR, specific IRR statistical method (e.g., Cohen's kappa), mention of methodology (e.g., grounded theory, ethnography), methods citations, use of language to signal agreement about codes when IRR or multiple coders were not used, and tools used (e.g., post-its, NVivo, ATLAS.ti, etc.). We coded these items only when they were unambiguously present in the text (i.e., we did not try to infer whether researchers had done something or not, we only coded when the paper described a practice). Many methods sections inevitably omitted details of the research process that will therefore not show up in our data. Because coding involved a simple binary identification—was a measurement reported in the paper or not—we chose not to calculate IRR per our guidelines below. Two researchers coded the full dataset of texts for presence of each measurement.

In the next section, we present our results as a descriptive quantitative analyses of reported methodological choices across recent qualitative CSCW and CHI papers, including excerpts from individual papers with citations to illustrate concepts. In cases where the analysis could be perceived as critical, we lightly disguise the excerpt and do not include a citation to the paper. The intent of this work is to identify areas of excellence in our collective approach to choosing and reporting qualitative research methods and to support improvement of our collective technique, but not to call out individual scholars and their work.

4 LOCAL PRACTICES IN CSCW AND CHI

4.1 Approaches to Reliability

Papers used a variety of techniques to communicate agreement. Some described how multiple coders continually met to develop, discuss, and refine codes (e.g., "After coding the first interview transcript, the two coders met to discuss disagreements in coding and to refine codebook definitions" [60] or "During weekly team meetings, we iteratively discussed and revised these codes until we reached consensus" [47]). Another paper described how they "collaboratively analyzed and discussed the data material in our project team as well as with another qualitative analysis group" [82]. In some cases, two or more researchers separately analyzed a subset of the data and then met to discuss and refine emergent codes. Others described one researcher doing the coding and then meeting with coauthors to discuss the data and refine codes or themes over one or several sessions. There were in-depth qualitative analyses that forwent agreement language altogether. A small portion of papers used external reviewers "to establish validity" [95] in one case and in another, as coders to "evaluate the final prototype" [11].

In 251 (96.2% or all but 10) of the papers, the authors described how they analyzed the data. We present most of our descriptive results (e.g., frequencies) in reference to the dataset involving these 251 papers only. Many used open coding processes in which researchers developed and revised codes as they reviewed the data. A small number of papers justified this choice explicitly, as in "an inductive thematic analysis was then conducted on the data set... This method was

Described data analysis	Total papers 251		CHI papers 134		CSCW papers 117	
Described a method of agreement	139	55.4%	65	48.5%	74	63.2%
Cited a methods text	180	71.7%	89	66.4%	91	77.8%
Used IRR	32	12.7%	22	16.4%	10	8.5%
Cohen's kappa	17	53.1%	13	63.6%	4	40.0%
Other specified	9	28.1%	6	27.3%	3	30.0%
Not specified	6	18.8%	3	13.6%	3	30.0%
Did not specify number of researchers (e.g., "we")	130	51.8%	70	52.2%	60	51.3%
Specified number of researchers	121	48.2%	64	47.8%	57	48.7%
1 researcher	25	20.7%	11	17.2%	14	24.6%
2 researchers	65	53.7%	38	59.4%	27	47.4%
3 or more researchers	31	25.6%	15	23.4%	16	28.1%

Table 1: Approaches to describing reliability among 251 qualitative CSCW and CHI papers in our dataset

deemed appropriate as […] is an emerging and poorly understood topic" [66]. Most papers used a process we would describe as inductive, even if not stated explicitly, however some specified a deductive approach. For example, Fiesler, Morrison, and Bruckman stated that they "began with an inductive approach but then grouped themes deductively under the framework of feminist HCI once we identified its relevance" [32]. Four other CHI papers and five CSCW papers used deductive approaches as part of a deductive thematic approach [36,86] or mixed deductive and inductive approach [13,22,49,52,58,69,92].

Among those that discussed their approach to analysis, just under half (121 papers, 48.2%) specify the number of people who analyzed qualitative datasets, with the majority of those specifying 2 researchers or authors (65 papers, 53.7%). Those that don't specify a number typically refer to "we" in their description of the analysis, or sometimes "the team" or "the group": "The entire team then reviewed and analyzed the data…" [14]. Some described a first or single author doing an initial pass and then the "research team" going on to code.

In total, 139 (55.4%) papers provided language that indicated some form of agreement was reached, primarily among multiple researchers. In addition to some of the examples given, Yarosh et al. stated "more than 750 open codes were then discussed among all four investigators to resolve any disagreements" [93].

In all, 10 papers described triangulation of some kind (data not shown). Finn and Oreglia described "triangulation" of their notes with other researchers and also use of other documents to "substantiate (or not)" some of their interpretations [33]. Another paper specified using various types of triangulation, including "data triangulation" (use of multiple studies), "triangulation of sources of evidence" (questionnaires, interviews, and other artifacts) as well as "analysis triangulation" (between two researchers) to achieve "validity and trustworthiness" in place of "generalisability and reliability" [67]. Seven of those 10 described triangulation with different sources of data [26,80,95] such as user logs [39], data provided by observers/researchers [5], and quantitative data gathered during their research [1,94]. One study that described their process of

Used IRR	Specified no. of coders	Described seeking agreement	# of papers	% of papers
✓	✓	✓	20	8.0%
✓	✓	x	10	4.0%
✓	x	✓	1	0.4%
x	✓	✓	75	29.9%
✓	x	x	1	0.4%
x	✓	x	16	6.4%
x	x	✓	43	17.1%
x	x	x	85	33.9%
			251	100%

Table 2: Combinations of approaches to reporting on reliability.

agreement specified that triangulation with their multiple sources (in this case, doctors, patients, NGOs, etc.) was not always appropriate because they maintained "contrary viewpoints" [46].

A relatively small subset, 32 papers (12.7%), described using IRR. Of those 32, the majority (26 papers, 81.3%), specified a statistical method and one paper reported using more than one method. The six papers that did not specify a method reported IRR as a simple percentage. Cohen's kappa was the most commonly used measure of IRR (17 out of 32 papers, 53.1%) with Fleiss' kappa (3 papers, 9%) , Krippendorff's alpha (3 papers, 9%), and Scott's Pi or Kappa (2 papers, 6%) used less often. Papers rarely described why a statistical method was chosen, although in one case, Fleiss and Krippendorff were both corresponding to different types of data being analyzed (nominal or categorical data used Fleiss; ordinal data used Krippendorff).

Eleven papers described agreement as "substantial," "very good," "sufficient," "satisfactory," "moderate," and "fair to good." Some papers conducted a second round of coding and IRR if agreement was too low after the first round; others kept the low IRR and provided a rationale for it, such as "the first pass inter-rater reliability test achieved a Kappa score of 0.61 as there was some confusion about redundant codes" [89] and "The coding comparison revealed a Cohen's Kappa figure of 0.54… reflect a codebook which was meaningful (objective) but not too restrictive (allowed subjectivity)" [60].

A few papers addressed non-use of IRR. For example, Paavilainen et al. stated that "As the informants wrote clear, brief sentences making the data easy to interpret and no major disagreements on coding arise, inter-rater reliability test was not needed" [75]. Jun et al. clarified their decision not to use IRR saying that "Inter-rater reliability was not calculated as it is rarely done with semi-structured interview data due to the possibility of applying the same code to different sections of the interview" [48]. Yet another paper pointed out that they did not perform IRR because while multiple researchers reviewed the codes, only one conducted the coding, but also added that "multiple researchers reviewed the codes and themes and there were critical and detailed discussions at all stages of the analysis" [91].

	Total papers (N=251)		Used IRR		Sought agreement		Specified # of coders	
Interviews	179	71.3%	16	8.9%	101	56.1%	82	45.6%
Observations	76	30.3%	5	6.6%	44	57.9%	30	39.5%
Scraping/Searching Existing Content	44	17.5%	8	18.2%	32	72.7%	31	70.5%
Design Sessions/Probes/Workshops	29	11.6%	4	13.8%	9	31.0%	11	37.9%
Free Response Survey	15	6.0%	2	13.3%	6	40.0%	7	46.7%
Diary Studies	14	5.6%	3	21.4%	6	42.9%	6	42.9%
Focus Groups	11	4.4%	1	9.1%	7	63.6%	5	45.5%

Table 3: Approaches to reporting on reliability by data collection method

Though not the focus of our analysis, we observed some differences between the CHI and CSCW datasets: 16.4% of CHI papers used IRR and 8.5% of CSCW papers did. At the same time, more CSCW papers specified their method of agreement (i.e., describing how they "discussed," "disagreed," "collaborate," "refined," etc. their codes or themes) (63.2%) compared with CHI (48.5%). We report on these differences between the conferences in Table 1.

4.2 Summary of Reliability Approaches

Table 2 summarizes combinations of approaches to reporting reliability and/or agreement in the 251 papers. Of these, 166 (66.1% or the first seven rows in Table 2) used at least one of the approaches to reporting reliability that we identified in the CSCW and CHI literature (i.e., use of IRR, number of coders, description of agreement process). 20 (8.0%) of papers articulated three forms of agreement (IRR, number of coders, description of agreement process), 86 (34.3%) articulated two of the three forms of agreement, and 60 (23.9%) articulated one of the three forms of agreement. Notably, the one paper that used IRR but did not describe either of the other two dimensions of agreement referred only to "we" and although they provided a detailed description of their process, did not use language that explicitly described their method of agreement.

4.3 Approaches to Reliability by Data Collection Method

Of the 251 papers, 179 (71.3%) described conducting interviews (see Table 3). Authors described conducting observations in 76 (30.3%) of the papers, and when used, observations were often accompanied by interview methods (62 of the 76 observation studies used both interviews and observations). Scraping/Searching existing content were third most common, reported in 44 papers (17.5%). Addition to using one of the methods described in Table 3, the most common other methods (n = 34) were recorded video (n = 12) and photographs (n = 10). There was little overlap among types of data collection excluding interviews.

Table 3 also describes use of measures of reliability by data collection activity. One variance of note is that scraping or searching of existing content, such as tweets, was more likely to use at

	249 papers named a method		Used IRR*		Indicated agreement		Specified no. of coders		Cited methods text	
Code for "themes"	161	64.7%	14	8.7%	93	57.8%	79	49.1%	116	72.0%
Other**	127	51.0%	14	11.0%	70	55.1%	62	48.8%	101	79.5%
Iterative	94	37.8%	6	6.4%	54	57.4%	41	43.6%	74	78.7%
Open coding	85	34.1%	9	10.6%	60	70.6%	46	54.1%	64	75.3%
Inductive	78	31.3%	11	14.1%	53	67.9%	37	47.4%	66	84.6%
Thematic analysis	65	26.1%	4	6.2%	36	55.4%	35	53.8%	54	83.1%
Grounded theory	44	17.7%	6	13.6%	31	70.5%	18	40.9%	40	90.9%
Axial	27	10.8%	5	18.5%	23	85.2%	17	63.0%	19	70.4%
Codebook	27	10.8%	10	37.0%	24	88.9%	24	88.9%	21	77.8%
Affinity diagram	27	10.8%	3	11.1%	21	77.8%	14	51.9%	21	77.8%
Ethnographic	13	5.2%	0	0.0%	6	46.2%	3	23.1%	12	92.3%

Table 4: How agreement was reported in the 249 papers that named specific methods

***Percentages in columns to the right are out of corresponding base "n" in this column.**
****Examples of "Other": coding, categorizing or classifying (24), memos (22), clustering/clusters (19), deductive (13), constant comparative (12), content analysis (9), feminist HCI (3), etc.**

least one measure of reliability. For example, Starbird et al. analyzed tweets for predetermined categories which lends itself to multiple coders for scale, and measures of agreement. They stated that "the first dimension, which we designed to identify crowd corrections, consists of five mutually exclusive categories: Affirm, Deny, Neutral, Unrelated, and Uncodable" [84] and subsequently calculated IRR to test agreement between coders for these categories.

4.4 Methods and Methodologies

Almost all papers (249) specified the method(s) used in the paper. The most commonly invoked descriptive term was iterative (94 papers, 37.8%) followed by open coding (85 papers, 34.1%) and other inductive approaches (see Table 4). Most papers (161, 64.7%) described using coding to identify themes, but fewer than half specified method of analysis (e.g., "thematic analysis") (data not shown). Almost one in five papers referenced grounded theory (Table 4). However, most of these papers used language that suggested a variation of grounded theory, such as "we coded and organized our data, based on a grounded theory approach" [9], "Data analysis was approached through a constructivist grounded theory approach" [59] or "We performed a thematic analysis of our data … adopting an inductive and grounded approach" [34]. In our analysis, we counted use of the terms "grounded theory," "grounded approach," and "grounded data analysis" in the context of describing the methods such as, "we analyzed review data through a grounded approach." In some examples, the term grounded was used in a way that did *not* clearly invoke grounded theory as a method and was not included in our dataset, for example, "our analysis was well-grounded in the data." 34 out of 44 papers included the precise term "grounded theory." Although 27 papers (10.8%) mentioned axial coding, only nine of those papers described their

methods as grounded theory. Similarly, open coding was mentioned in 85 papers (34.1%) but only 20 of those papers described their methods as grounded theory (data not shown).

4.5 Justifications and Rationales

In the above sections, we provide a description of *how* qualitative researchers report their methods of analyzing data related to establishing and communicating reliability. Most papers in our dataset described their data analysis methods to some extent, however, few papers presented justifications for those choices. Even the most detailed descriptions of *how* research was performed were often written assuming sufficient prior knowledge on the part of readers to interpret *why* the research was done in this way. Sometimes one step was justified while others were simply described. Although we do not report formal counts of methodological justifications, we infer from our readings that many CSCW and HCI researchers assume most readers have sufficient methodological expertise or training to understand why authors might choose to employ multiple coders or use the statistical measure they used.

In contrast, rationales for study design choices often were included, such why it was important to use deception and why interviewers did not define terms for participants [76], or because a certain step was required by a methodological text [21]. We speculate that authors provide a rationale when they anticipate readers may question their choices, either based on prior review experience or because they are aware of a common disciplinary standard from which their practices depart. Outside of those rationales, presentation style norms and page limits might lead authors to limit discussion of methodological concerns. In the next section, we discuss recommended approaches to agreement and reliability in different forms of qualitative research, including reasons not to seek agreement and rationales for these approaches.

5 MOVING FORWARD: REFLECTIONS ON BEST PRACTICES

In this section, we draw on our results as well as methods literature to reflect on instances of when *not* to seek agreement or IRR, when *to* seek agreement or use IRR, as well as considerations for communicating and justifying methods in written scholarship. Our goal is to assist researchers in aligning their assumptions and choice of methods and in making reasoned justifications for their choices. In the next sections, we traverse a wide range of qualitative research traditions to reflect on when researchers might choose to use multiple coders and seek agreement, when they might also choose to use IRR, and what they might consider in describing these efforts in publications. We also reflect on where those approaches are not needed, or can even be outright rejected.

5.1 Reasons Not to Seek Agreement and/or IRR

5.1.1 When codes are the process not the product. Many papers in our dataset demonstrated that coding takes place over multiple meetings to discuss disagreements and refine codes (descriptive text attached to a unit of analysis), the primary goal of which is not agreement, but to eventually yield concepts and themes (recurrent topics or meanings that represent a phenomena). Even if coders agree on codes, they may interpret the meaning of those codes differently, and those differences may be valuable as described here: "Afterwards, our team examined the codes and initial themes together, and had several group meetings where authors discussed and did a comparative analysis of each of their codes to note similarities and differences" [77]. The coding process may itself involve multiple rounds of coding, meeting, diverging, synthesizing,

highlighting, revising memos, and recording, the output of which is not always agreement, but discovery of emergent themes as described in Klein et al. [54].

5.1.2. Expert researcher. Agreement (formal or informal) is rarely appropriate when a single researcher with unique expertise and experience is conducting the research. Many techniques such as memoing, reflection, and review are designed to support the single researcher process without requiring agreement (e.g., "To answer these questions, I reviewed the data collected during the four years of my fieldwork" [81]), though triangulation may be useful in some cases. This is often the case in ethnographic work where researchers are embedded in a topic or field for long periods. Barkhuus and Rossitto argue that "Subjective analysis is what makes ethnographic methods relevant and powerful… it is impossible to take the preliminary experiences out of the ethnographer" [7]. The ethnographer is often placed in a privileged position by virtue of acceptance of the ethnographers' subjectivity with access to an "ideal of correspondence" between an event and its rendering [4]. In CSCW and HCI, emphasis on a participatory turn (the "insertion of the ethnographer into the scene") suggests "the self" as "an instrument of knowing" [28]. Dourish underscores the role of ethnographic participation not just as an "unavoidable consequence of going somewhere, but as the fundamental point" and goes on to say that "If we accept a view of ethnographic material as the product of occasions of participative engagement, then we surely need to be able to inquire into the nature of that engagement" [28]. Autoethnographic accounts are a product of often single-researcher participation in a practice of interest and rigorous self-reflection [2] and have been advocated in the HCI literature as a way of developing empathy with users [73].

5.1.3. Social context or social action. In some types of research where social context or social change is integral, agreement among external researchers may not be meaningful. For research that takes a social, ethical, or political stance, agreement may not be required as communities themselves may be charged with producing their own insights that in turn allow them to enact change. For example, participatory action research (PAR) engages participants in "self-investigation" with the hope of producing action [79]. Other approaches include co-design or interventions where the participants themselves took part in interpretation, for example, "with this intervention 'in the wild,' we as investigators have participated in the study at the same time as the participants have been investigators" [44]. One paper in our dataset embraced the potentially idiosyncratic role of researchers interpreting the unfamiliar as follows: "abductive reasoning can be used in situations of uncertainty… abduction is driven by astonishment, mystery, and breakdowns" [71].

5.1.4. Agreement can be harmful. Some scholarly standpoints are philosophically at odds with the concept of agreement. For example, feminist HCI [6,8] researchers seeking to challenge hegemonic categories of available knowledge and to privilege marginal or subaltern perspectives may not adopt an analytical stance a priori. Similarly, intersectional approaches challenge the systems that reproduce inequality, including the power assumptions embedded into the very concept of coding and labeling [12]. Research that is focused on social change, power structures, and empowerment may consider the concept of agreement to perpetuate injustices the research is looking to overcome [23].

5.1.5. Ease of coding. For some data and analyses, multiple coders and IRR are overkill because the data is straightforward. If the coding is binary, then the coding likely only requires one coder. Another example from our dataset states: "three researchers independently coded the responses using constant comparison to iteratively arrive at themes… Nine remaining open-ended questions were simpler elaborations, which were coded independently by one researcher" [51]. Though a

stigma may exist around having a single coder in qualitative research (which may foster the use of "we" in single-authored papers), it is worth reflecting on when one researcher really is just as good as two. We did not calculate IRR for our content analysis of methods sections because flagging the presence of terms or citations is a simple form of coding that required little to no interpretation.

5.1.6. Grounded theory. We pay special attention to the curious case of grounded theory because although it does not require IRR, we encountered several papers that used grounded theory methods (e.g., "grounded theory" or "grounded approach") in conjunction with IRR. Grounded theory often (though not always) engages in a standardized format of analysis that requires multiple iterations of interactions with the data, followed by analysis, and then more theoretical sampling and analysis, and finally the development of theory [20,24,72]. There are many possible steps in the coding process: open coding to identify topics of interest, axial coding to identify relationships among the codes so that they can be organized into clusters of more complex and themes, and selective coding where the researcher focuses more narrowly on topics of interest [20,72].

Grounded theory rarely, if ever, requires IRR. Grinter has pointed out that IRR is not specified in any of the foundational grounded theory texts and explained that, for the grounded theorist, codes are "merely" an interim product that support the development of theory, not a final result that requires testing [40]. From initial sampling to theoretical sampling, the researchers' aim is to refine the theory and test it "to broaden the range of situations and attributes over which the theory makes good predictions or descriptions" [72]. The procedures used in grounded theory, when done in this way, perform the critical, reflexive work that informal discussions of agreement might achieve. That said, some papers in our dataset still readily report use of grounded theory methods (e.g., open coding, axial coding, constant comparison) while also performing IRR. The differences in language choice of using grounded theory versus grounded approach also indicates variance in CSCW researchers' alignments to grounded theory principles; future work could explore whether those language choices are intentional or meaningful.

5.2 Reasons to Seek Agreement

Agreement of any kind plays a large role in qualitative research. Recall that 2/3 of papers in our dataset used some form of agreement and just over 1/2 used agreement without IRR (i.e., either described seeking agreement or specified number of coders).

5.2.1 Capacity to code more data. When larger datasets are desirable (e.g., coding tweets for frequency of a topic), seeking agreement provides researchers with the capacity to code more data by spreading coding across multiple researchers. In order to ensure that interpretations of data are aligned, coders can use informal discussion or formal statistical methods to develop and test their agreement. Agreement is particularly useful when datasets have readily discernable units of analysis, such as tweets or posts.

5.2.2 Confirmation of measurements. Krippendorff asserts that "reliability is the degree to which members of a designated community concur on the readings, interpretations, responses to, or uses of given texts or data" [55]. When multiple people come to agreement on codes and themes in a dataset, agreement signals confirmation that the measurements are consistent. For example, in grounded theory, Strauss argues that each data collector should be engaged in analysis of their own data as well as with others (together and individually) from start to finish [85].

5.2.3 Identifying points of disagreement. Identifying where a group of researchers disagrees about interpretations can be helpful to illustrate points of tension around codes, and the process

of seeking agreement may be useful regardless of whether agreement is eventually demonstrated. Disagreement can also prompt more reflection that ultimately results in a stronger codebook and coding process. Many papers in our dataset described this process as refining or discussing codes, for example in this inductive approach: "the two coders met to discuss disagreements in coding and to refine codebook definitions" [60]. When appropriate, authors may consider describing not only the process by which disagreements were addressed, but also the data that sparked disagreement.

5.3 Reasons to Use IRR

If multiple coders participate in data analysis, there are a number of situations in which it is appropriate to use and report formal measures of IRR.

5.3.1. Ensure consistency between coders coding different subsets. As described in 5.2.1, large datasets often require that multiple coders attend to different subsets of the data. In those cases, IRR can be useful to gauge consistency of application of codes. It may also be useful for increasing researchers' confidence that coders are implementing the codes as intended. Once agreement is established on a small portion of the data, coders can then code separate portions of the data independently. We used this approach in our application of inclusion criteria for a random sample of CSCW and CHI papers as described in section 3.1.

5.3.2. Developing codes. IRR can be used in the process of developing codes themselves; that is, to support the process of interpretation instead of to justify and/or demonstrate the strength of codes and interpretation. Here, again, discovering where researchers disagree could be more interesting than demonstrating that they agree. IRR can reveal something about the nature of the data itself: how easy or challenging was it to synthesize and how easy or hard it was to attach meaning to units of analysis [43].

5.3.3. Identify predetermined codes. IRR can be used to evaluate whether data are being interpreted in the same way relative to the concepts being measured. For example, a study that aims to label tweets using a pre-determined set of codes would require that researchers identify those codes in the same way. As a general rule with multiple coders, the more interpretation that is required to code the data, the more important it might be to use IRR.

5.3.4. Describe results quantitatively. Some analyses seek to report quantitative or statistical analysis of qualitative data (e.g., that a topic showed up in X% of the dataset). If findings hinge on frequency counts, there is a strong argument that IRR should be used. In cases where topics represented by codes are going to be compared to each other (e.g., X was three times more common than Y), IRR may also be useful.

5.3.5. Enable replicability. Quantitative researchers across disciplines like psychology, medicine, and more recently, HCI, are advocating for open science to counter concerns about p-hacking, bias towards significant results, and other statistical concerns [50]. Often qualitative research is not designed to be replicated, but in some cases, researchers may seek to reuse methods or expand and refine results in subsequent studies. In these cases, IRR forces researchers to develop a well-articulated process of codebook development and coding that can then be made available to other researchers. Inspired by the open science movement, one way that researchers could facilitate replicability is by being concrete about thresholds for agreement at each stage. Papers could suggest, or even preregister, a target agreement value (e.g., 0.7 or above) with justification before conducting the analysis, which would lend credibility to their subsequent interpretation of determined values.

5.3.6. Confirmation bias. Most research methods are subject to confirmation bias. In qualitative research, confirmation bias can occur in the data collection, analysis, or interpretation process where researchers(s) may have preconceived notions about what they will discover, or want to discover, and those biases can inadvertently influence their scholarship. While we have noted some cases where researchers' perspectives are an important part of the scholarly process (which is often acknowledged with a positionality statement in the methods section), in many cases it is important to provide checkpoints to minimize the effects of confirmation bias. Using multiple coders and agreement processes can help reduce individual biases of a particular researcher.

5.4 Considerations When Writing Methods Sections

Although our analysis focused on reliability, through the process of reading hundreds of methods sections, we observed many variations and developed general observations for effectively communicating qualitative research methods.

5.4.1 Scaffolding readers' understanding. An important outcome of a methods section is that readers are prepared to understand the research that was done and the findings that are presented. Providing pragmatic details about the process of analysis, even if it may seem obvious to a seasoned researcher, is valuable for providing other researchers with insight about how to execute (or not) similar research.

5.4.2 Passive voice. CSCW and HCI researchers often use passive voice in methods sections. There are numerous examples from our dataset: "data were combined and analyzed," "open coding was applied to the transcripts," "interviews transcripts were analyzed," "field notes were reviewed," "coding was discussed," "the emergence of themes," etc. Although researchers may have carefully considered reasons for using passive voice, passive voice can obfuscate the role of researchers in performing the analysis. Using passive voice has the potential for, as Bohner writes, "obscuring agency by placing the actor in the background" and "introducing interpretational ambiguity" both of which he argues authors semi-consciously do out of insecurity about the rigor of their methods [10]. Some researchers consciously choose to treat data as though the data itself has agency and, similarly, that theories can have agency in analysis. In those cases, it may be appropriate to communicate that a theory was "discovered" or that themes "emerged" from data rather than that the researcher constructed the theory or developed the themes. However, it is important that researchers are making an intentional choice to assign data and theories that agency.

5.4.3 Performativity. The publication process encourages scholars to perform scholarship aligned with normative expectations. Prior discussions have pointed out instances of how these expectations shape scholarship, such as "implications for design" at CHI [27]. In qualitative research, we see norms emerge and evolve over time, such as new approaches and perspectives moving into the community (e.g., grounded theory and ethnography in earlier decades; intersectionality currently) that go through a period of uncertainty and refinement. However, some norms emerge and can spread without deep discussion or shared reflection. Examples like the widespread adoption of IRR rating scales, or the ambiguous language of "a grounded theory approach" may foster uncertainty about what research activities these terms actually represent. In some cases, papers make processes visible and call into question emergent norms, such as this example in our dataset: "Without pretending to achieve saturation of the data analysis, the highlighted themes helped us to elaborate the four design opportunities…" [87].

Indeed, our review of methods sections suggests that some authors may assume there are expected standard practices for doing qualitative research that need not be elaborated, and they

may also use these perceived standards to guide decisions about when to include justifications for methodological choices. This may be to our collective detriment. Descriptions in our datasets such as "in keeping with standard qualitative analysis techniques" and "followed standard guidelines to code our themes" may be attempts to signal conformity to an accepted method; yet, our research makes clear the diversity of approaches and epistemologies. There is no standard qualitative method in CSCW or HCI. We recommend clear and detailed descriptions of analytical procedures, even when researchers suspect that most people share the same practices. For Glaser and Strauss, theory is process: "an ever-developing entity" rather than a "perfected product" [38]. When conducting and reporting qualitative research, process and rationale should be treated as a critical part of the research.

5.5 Logistics and Equity

Research is constrained and enabled not only by scholars' creativity, insight, and skill, but by practical aspects of their work environment. Access to resources can influence research design and execution.

5.5.1 Time and resources. Research is also constrained by access to resources, including collaborators, research assistants, financial support, and hardware and software tools. In most studies, there may be an "ideal" way to conduct research and then there is the practical way, the latter of which is what takes place. A challenge for the community moving forward is how to balance ideals of scholarship with constraints and inequalities that make some of the recommendations above impractical to implement consistently. For example, researchers may determine that multiple coders are needed for a particular study design, but multiple coders may not readily materialize. Although institutions or individuals with access to more resources can usually access multiple coders (e.g., by hiring and training students), not every researcher will have access to multiple coders. The guidelines in this paper may also help researchers consider feasibility of different research designs in low-resource conditions.

5.5.2 Expertise. Universities and research labs with few or only one CSCW and HCI researcher cannot have expertise in all related methods. At universities where there may not be faculty with training in qualitative methods, but who have students who want to do this kind of work, the broader CSCW and HCI research community has an opportunity to help train those students through the publication process. This may take place through direct feedback during peer review, but can also happen when researchers provide detailed accounts of methods. In this spirit, we hope that the reflections and recommendations in this paper provide a resource for the CSCW, HCI, and adjacent research communities to continue to reflect on their collective practices.

6 CONCLUSIONS

Qualitative research is a rich and powerful way to understand the world around us. Yet, we have demonstrated that in CSCW and HCI, there is little consensus about how to approach reliability in qualitative research. As a result, authors often struggle to communicate methodological choices with confidence and reviewers may communicate confusing expectations. As researchers, we might expect that epistemological stances dictate methodological choices; however, in practice, other forces like perceived norms and reviewer variance may play a role. Our descriptive analysis of recent papers finds that most qualitative CSCW and HCI papers code data as part of their research process and most provide some description of the method they use to do so. Far fewer report the number of coders involved in the process and even fewer use IRR. Justifications for

these choices are sparse. We argue that diverse epistemological standpoints in CSCW and HCI support these diverse uses, but that the community could be clearer about communicating the rationale for methodological choices. Further, there is variability in how terms and concepts are used, and although implemented in similar ways, they can be used differently and toward different ends. It is precisely these differences that necessitate thorough descriptions of methods and analytical process. Through our analysis and discussion, we examine a wide range of research cases, more experiences and situations than perhaps any one researcher is likely to have encountered in their own practice. Finally, we discuss considerations for best practices in writing methods and provide a set of recommendations for when researchers should seek agreement in qualitative research and when it is not needed, or even potentially harmful.

In sum: there is no one correct way to approach reliability in qualitative research. Different epistemologies invite different frames; we hope that this work serves as a generative starting point for researchers to reflect on their epistemological goals and how they produce knowledge.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation grants CNS-1703736 and CHS-1552503. Thank you to Kentaro Toyama, Susan Wyche, and Denise Agosto for their valuable feedback on early drafts of this paper.

REFERENCES

[1] Ali Abdolrahmani, William Easley, Michele Williams, Stacy Branham, and Amy Hurst. 2017. Embracing Errors: Examining How Context of Use Impacts Blind Individuals' Acceptance of Navigation Aid Errors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 4158–4169. DOI:https://doi.org/10.1145/3025453.3025528

[2] Tony E Adams, Carolyn Ellis, and Stacy Holman Jones. 2017. Autoethnography. *The international encyclopedia of communication research methods* (2017), 1–11.

[3] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. 1997. The Place of Inter-Rater Reliability in Qualitative Research: An Empirical Study. *Sociology* 31, 3 (August 1997), 597–606.

[4] Paul Anthony Atkinson, Amanda Coffey, Sara Delamont, John Lofland, and Lyn H. Lofland (Eds.). 2001. *Handbook of Ethnography* (1 edition ed.). SAGE Publications Ltd, London ; Thousand Oaks, Calif.

[5] Mara Balestrini, Paul Marshall, Raymundo Cornejo, Monica Tentori, Jon Bird, and Yvonne Rogers. 2016. Jokebox: Coordinating Shared Encounters in Public Spaces. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16), 38–49. DOI:https://doi.org/10.1145/2818048.2835203

[6] Shaowen Bardzell. 2010. Feminist HCI: Taking Stock and Outlining an Agenda for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), 1301–1310. DOI:https://doi.org/10.1145/1753326.1753521

[7] Louise Barkhuus and Chiara Rossitto. 2016. Acting with Technology: Rehearsing for Mixed-Media Live Performances. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 864–875. DOI:https://doi.org/10.1145/2858036.2858344

[8] Rosanna Bellini, Angelika Strohmayer, Ebtisam Alabdulqader, Alex A. Ahmed, Katta Spiel, Shaowen Bardzell, and Madeline Balaam. 2018. Feminist HCI: Taking Stock, Moving Forward, and Engaging Community. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI EA '18), SIG02:1–SIG02:4. DOI:https://doi.org/10.1145/3170427.3185370

[9] Kirsten Boehner and Carl DiSalvo. 2016. Data, Design and Civics: An Exploratory Study of Civic Tech. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 2970–2981. DOI:https://doi.org/10.1145/2858036.2858326

[10] Gerd Bohner. 2001. Writing about rape: Use of the passive voice and other distancing text features as an expression of perceived responsibility of the victim. *British Journal of Social Psychology* 40, 4 (December 2001), 515–529. DOI:https://doi.org/10.1348/014466601164957

[11] Adrien Bousseau, Theophanis Tsandilas, Lora Oehlberg, and Wendy E. Mackay. 2016. How Novices Sketch and Prototype Hand-Fabricated Objects. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 397–408. DOI:https://doi.org/10.1145/2858036.2858159

[12] Geoffrey C. Bowker and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences* (Revised edition ed.). The MIT Press, Cambridge, Massachusetts London, England.

[13] LouAnne E. Boyd, Alejandro Rangel, Helen Tomimbang, Andrea Conejo-Toledo, Kanika Patel, Monica Tentori, and Gillian R. Hayes. 2016. SayWAT: Augmenting Face-to-Face Conversations for Adults with Autism. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 4872–4883. DOI:https://doi.org/10.1145/2858036.2858215

[14] LouAnne E. Boyd, Kyle Rector, Halley Profita, Abigale J. Stangl, Annuska Zolyomi, Shaun K. Kane, and Gillian R. Hayes. 2017. Understanding the Role Fluidity of Stakeholders During Assistive Technology Research "In the Wild." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 6147–6158. DOI:https://doi.org/10.1145/3025453.3025493

[15] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[16] Alan Bryman and Robert G. Burgess. 1994. *Analyzing Qualitative Data*. Routledge, London.

[17] G. Burrell and G. Morgan. 1979. *Sociological Paradigms and Organizational Analysis: Elements of the Sociology of Corporate Life*. Ashgate Publishing.

[18] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 981–992. DOI:https://doi.org/10.1145/2858036.2858498

[19] John L. Campbell, Charles Quincy, Jordan Osserman, and Ove K. Pedersen. 2013. Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research* 42, 3 (2013), 294–320.

[20] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis* (1 edition ed.). SAGE Publications Ltd, London ; Thousand Oaks, Calif.

[21] Ana Paula Chaves and Marco Aurelio Gerosa. 2018. Single or Multiple Conversational Agents?: An Interactional Coherence Comparison. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 191:1–191:13. DOI:https://doi.org/10.1145/3173574.3173765

[22] Chia-Fang Chung, Elena Agapie, Jessica Schroeder, Sonali Mishra, James Fogarty, and Sean A. Munson. 2017. When Personal Tracking Becomes Social: Examining the Use of Instagram for Healthy Eating. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 1674–1687. DOI:https://doi.org/10.1145/3025453.3025747

[23] Patricia Hill Collins. 2015. Intersectionality's Definitional Dilemmas. *Annual Review of Sociology* 41, 1 (2015), 1–20.

[24] Juliet Corbin and Anselm Strauss. 2007. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (3rd ed.). SAGE Publications, Inc.

[25] John W. Creswell. 1998. *Qualitative Inquiry and Research Design : Choosing Among Five Traditions*. Sage Publications Inc, Thousand Oaks, Calif.

[26] Dharma Dailey and Kate Starbird. 2017. Social Media Seamsters: Stitching Platforms & Audiences into Local Crisis Infrastructure. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 1277–1289. DOI:https://doi.org/10.1145/2998181.2998290

[27] Paul Dourish. 2006. Implications for design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '06), 541–550. DOI:https://doi.org/10.1145/1124772.1124855

[28] Paul Dourish. 2014. Reading and Interpreting Ethnography. In *Ways of Knowing in HCI* (Judith S Olson and Wendy A. Kellogg (Eds)). Springer-Verlag, New York, 1–23.

[29] Ellen A. Drost. 2011. Validity and Reliability in Social Science Research. *Education Research and Perspectives* 38, 1 (June 2011), 105–123.

[30] Robert Elliott, Constance T. Fischer, and David L. Rennie. 1999. Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology* 38, 3 (September 1999), 215–229.

[31] Robert M. Emerson, Rachel I. Fretz, and Linda L. Shaw. 2011. *Writing Ethnographic Fieldnotes, Second Edition* (Second edition ed.). University of Chicago Press, Chicago.

[32] Casey Fiesler, Shannon Morrison, and Amy S Bruckman. 2016. An archive of their own: a case study of feminist HCI and values in design. 2574–2585.

[33] Megan Finn and Elisa Oreglia. 2016. A Fundamentally Confused Document: Situation Reports and the Work of Producing Humanitarian Information. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16), 1349–1362. DOI:https://doi.org/10.1145/2818048.2820031

[34] Katie Z. Gach, Casey Fiesler, and Jed R. Brubaker. 2017. "Control Your Emotions, Potter": An Analysis of Grief Policing on Facebook in Response to Celebrity Death. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (December 2017), 47:1–47:18. DOI:https://doi.org/10.1145/3134682

[35] C. Geertz. 1983. Thick Description: toward an interpretive theory of culture. In *The Interpretation of Cultures*. Basic Books, New York, 3–32.
[36] Kathrin Gerling, Kieran Hicks, Michael Kalyn, Adam Evans, and Conor Linehan. 2016. Designing Movement-based Play With Young People Using Powered Wheelchairs. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 4447–4458. DOI:https://doi.org/10.1145/2858036.2858070
[37] Lisa Given. 2008. Natural Setting. In *The Sage encyclopedia of qualitative research methods*. SAGE, London.
[38] Barney Glaser and Anselm Strauss. 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine Transaction.
[39] Daniel Gooch, Asimina Vasalou, Laura Benton, and Rilla Khaled. 2016. Using Gamification to Motivate Students with Dyslexia. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 969–980. DOI:https://doi.org/10.1145/2858036.2858231
[40] Beki Grinter. 2010. Inter-Rater Reliability. *Beki's Blog*. Retrieved January 29, 2019 from https://beki70.wordpress.com/2010/09/09/inter-rater-reliability-apply-with-care/
[41] Egon G. Guba. 1981. Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal* 29, (1981), 75–91.
[42] Kevin A. Hallgren. 2012. Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutor Quant Methods Psychol* 8, 1 (2012), 23–34.
[43] David Hammer and Leema K. Berland. 2014. Confusing Claims for Data: A Critique of Common Practices for Presenting Qualitative Research on Learning. *Journal of the Learning Sciences* 23, 1 (January 2014), 37–46.
[44] Hanna Hasselqvist, Mia Hesselgren, and Cristian Bogdan. 2016. Challenging the Car Norm: Opportunities for ICT to Support Sustainable Transportation Practices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 1300–1311. DOI:https://doi.org/10.1145/2858036.2858468
[45] John Hughes, Tom Rodden, and Hans Andersen. 1994. Moving out from the control room: ethnography in system design. *ACM Conference on Computer-Supported Cooperative Work* (1994), 429–439.
[46] Azra Ismail, Naveena Karusala, and Neha Kumar. 2018. Bridging Disconnected Knowledges for Community Health. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 75:1–75:27. DOI:https://doi.org/10.1145/3274344
[47] Jialun "Aaron" Jiang, Casey Fiesler, and Jed R. Brubaker. 2018. "The Perfect One": Understanding Communication Practices and Challenges with Animated GIFs. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 80:1–80:20. DOI:https://doi.org/10.1145/3274349
[48] Eunice Jun, Blue A. Jo, Nigini Oliveira, and Katharina Reinecke. 2018. Digestif: Promoting Science Communication in Online Experiments. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 84:1–84:26. DOI:https://doi.org/10.1145/3274353
[49] Vaishnav Kameswaran, Jatin Gupta, Joyojeet Pal, Sile O'Modhrain, Tiffany C. Veinot, Robin Brewer, Aakanksha Parameshwar, Vidhya Y, and Jacki O'Neill. 2018. "We Can Go Anywhere": Understanding Independence Through a Case Study of Ride-hailing Use by People with Visual Impairments in Metropolitan India. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 85:1–85:24. DOI:https://doi.org/10.1145/3274354
[50] Matthew Kay, Steve Haroz, Shion Guha, Pierre Dragicevic, and Chat Wacharamanotham. 2017. Moving Transparent Statistics Forward at CHI. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17), 534–541. DOI:https://doi.org/10.1145/3027063.3027084
[51] Christina Kelley, Bongshin Lee, and Lauren Wilcox. 2017. Self-tracking for Mental Wellness: Understanding Expert Perspectives and Student Experiences. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 629–641. DOI:https://doi.org/10.1145/3025453.3025750
[52] Ryan Kelly, Daniel Gooch, Bhagyashree Patil, and Leon Watts. 2017. Demanding by Design: Supporting Effortful Communication Practices in Close Personal Relationships. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (CSCW '17), 70–83. DOI:https://doi.org/10.1145/2998181.2998184
[53] Jerome Kirk and Marc L. Miller. 1986. *Reliability and validity in qualitative research*. Sage Publications, Beverly Hills.
[54] Maximilian Klein, Jinhao Zhao, Jiajun Ni, Isaac Johnson, Benjamin Mako Hill, and Haiyi Zhu. 2017. Quality Standards, Service Orientation, and Power in Airbnb and Couchsurfing. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (December 2017), 58:1–58:21. DOI:https://doi.org/10.1145/3134693
[55] Klaus H. Krippendorff. 2003. *Content Analysis: An Introduction to Its Methodology* (2nd ed.). Sage Publications, Inc.
[56] Karen S. Kurasaki. 2000. Intercoder Reliability for Validating Conclusions Drawn from Open-Ended Interview Data. *Field Methods* 12, 3 (August 2000), 179–194.
[57] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. DOI:https://doi.org/10.2307/2529310
[58] Simone Lanette, Phoebe K. Chua, Gillian Hayes, and Melissa Mazmanian. 2018. How Much is "Too Much"?: The Role of a Smartphone Addiction Narrative in Individuals' Experience of Use. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 101:1–101:22. DOI:https://doi.org/10.1145/3274370

[59] Amanda Lazar, Hilaire J. Thompson, Shih-Yin Lin, and George Demiris. 2018. Negotiating Relation Work with Telehealth Home Care Companionship Technologies That Support Aging in Place. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 103:1–103:19. DOI:https://doi.org/10.1145/3274372

[60] Pierre Le Bras, David A. Robb, Thomas S. Methven, Stefano Padilla, and Mike J. Chantler. 2018. Improving User Confidence in Concept Maps: Exploring Data Driven Explanations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 404:1–404:13. DOI:https://doi.org/10.1145/3173574.3173978

[61] Margaret LeCompte and Judith Goetz. 1982. Problems of Reliability and Validity in Ethnographic Research. *Review of Educational Research* 52, 1 (1982), 31–60.

[62] Yvonna S. Lincoln and Egon G. Guba. 1986. But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Program Evaluation* 1986, 30 (1986), 73–84. DOI:https://doi.org/10.1002/ev.1427

[63] Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research* 28, 4 , 587–604. DOI:https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

[64] Catherine MacPhail, Nomhle Khoza, Laurie Abler, and Meghna Ranganathan. 2016. Process guidelines for establishing Intercoder Reliability in qualitative studies. *Qualitative Research* 16, 2 (April 2016), 198–212.

[65] Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 348:1–348:12. DOI:https://doi.org/10.1145/3173574.3173922

[66] Joe Marshall, Conor Linehan, and Adrian Hazzard. 2016. Designing Brutal Multiplayer Video Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 2669–2680. DOI:https://doi.org/10.1145/2858036.2858080

[67] Roberto Martinez-Maldonado, Lucila Carvalho, and Peter Goodyear. 2018. Collaborative Design-in-use: An Instrumental Genesis Lens in Multi-device Environments. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 118:1–118:24. DOI:https://doi.org/10.1145/3274387

[68] Alice Marwick, Claire Fontaine, and danah boyd. 2017. "Nobody Sees It, Nobody Gets Mad": Social Media, Privacy, and Personal Responsibility Among Low-SES Youth. *Social Media + Society* 3, 2 (April 2017), 2056305117710455.

[69] Marius Mikalsen and Eric Monteiro. 2018. Data Handling in Knowledge Infrastructures: A Case Study from Oil Exploration. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 123:1–123:16. DOI:https://doi.org/10.1145/3274392

[70] D. R. Millen. 2000. Rapid ethnography: time deepening strategies for HCI field research. 280–286.

[71] Trine Møller. 2018. Presenting The Accessory Approach: A Start-up's Journey Towards Designing An Engaging Fall Detection Device. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 559:1–559:10. DOI:https://doi.org/10.1145/3173574.3174133

[72] M. J. Muller and S. Kogan. 2010. Grounded theory method in hci and cscw. In *Technical Report*. IBM Watson Research Center.

[73] Aisling Ann O'Kane, Yvonne Rogers, and Ann E Blandford. 2014. Gaining empathy for non-routine mobile device use through autoethnography. 987–990.

[74] Judith S. Olson and Wendy A. Kellogg (Eds.). 2014. *Ways of Knowing in HCI*. Springer-Verlag, New York.

[75] Janne Paavilainen, Hannu Korhonen, Kati Alha, Jaakko Stenros, Elina Koskinen, and Frans Mayra. 2017. The PokéMon GO Experience: A Location-Based Augmented Reality Mobile Game Goes Mainstream. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 2493–2498. DOI:https://doi.org/10.1145/3025453.3025871

[76] Chanda Phelan, Cliff Lampe, and Paul Resnick. 2016. It's Creepy, But It Doesn'T Bother Me. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 5240–5251. DOI:https://doi.org/10.1145/2858036.2858381

[77] Laura R. Pina, Carmen Gonzalez, Carolina Nieto, Wendy Roldan, Edgar Onofre, and Jason C. Yip. 2018. How Latino Children in the U.S. Engage in Collaborative Online Information Problem Solving with Their Families. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 140:1–140:26. DOI:https://doi.org/10.1145/3274409

[78] David Pinelle and Carl Gutwin. 2000. A Review of Groupware Evaluations. In *Proceedings of the 9th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises* (WETICE '00), 86–91.

[79] MD A. Rahman. 2008. Some Trends in the Praxis of Participatory Action Research. In *The SAGE Handbook of Action Research* (P. Reason and H. Bradbury (eds)). Sage, London, 49–62.

[80] Shriti Raj, Mark W. Newman, Joyce M. Lee, and Mark S. Ackerman. 2017. Understanding Individual and Collaborative Problem-Solving with Patient-Generated Data: Challenges and Opportunities. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (December 2017), 88:1–88:18. DOI:https://doi.org/10.1145/3134723

[81] Amon Rapp. 2018. Gamification for Self-Tracking: From World of Warcraft to the Design of Personal Informatics Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 80:1–80:15. DOI:https://doi.org/10.1145/3173574.3173654

[82] Marén Schorch, Lin Wan, David William Randall, and Volker Wulf. 2016. Designing for Those Who Are Overlooked: Insider Perspectives on Care Practices and Cooperative Work of Elderly Informal Caregivers. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16), 787–799. DOI:https://doi.org/10.1145/2818048.2819999

[83] Alfred Schutz. 1967. *The Phenomenology of the Social World*. Northwestern University Press.

[84] Kate Starbird, Emma Spiro, Isabelle Edwards, Kaitlyn Zhou, Jim Maddock, and Sindhuja Narasimhan. 2016. Could This Be True?: I Think So! Expressed Uncertainty in Online Rumoring. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 360–371. DOI:https://doi.org/10.1145/2858036.2858551

[85] Anselm Strauss. 1987. *Qualitative Analysis for Social Scientists*. Cambridge University Press, Cambridge.

[86] Aaron Tabor, Scott Bateman, Erik Scheme, David R. Flatla, and Kathrin Gerling. 2017. Designing Game-Based Myoelectric Prosthesis Training. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 1352–1363. DOI:https://doi.org/10.1145/3025453.3025676

[87] Matthieu Tixier and Myriam Lewkowicz. 2016. "Counting on the Group": Reconciling Online and Offline Social Support Among Older Informal Caregivers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 3545–3558. DOI:https://doi.org/10.1145/2858036.2858477

[88] James R. Wallace, Saba Oji, and Craig Anslow. 2017. Technologies, Methods, and Values: Changes in Empirical Research at CSCW 1990 - 2015. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (December 2017), 106:1–106:18. DOI:https://doi.org/10.1145/3134741

[89] April Y. Wang, Ryan Mitts, Philip J. Guo, and Parmit K. Chilana. 2018. Mismatch of Expectations: How Modern Learning Resources Fail Conversational Programmers. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 511:1–511:13. DOI:https://doi.org/10.1145/3173574.3174085

[90] G. Cochran William. 1977. *Sampling Techniques, 3rd Edition* (3rd edition ed.). John Wiley & Sons, New York.

[91] Lillian Yang and Carman Neustaedter. 2018. Our House: Living Long Distance with a Telepresence Robot. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (November 2018), 190:1–190:18. DOI:https://doi.org/10.1145/3274459

[92] Svetlana Yarosh, Elizabeth Bonsignore, Sarah McRoberts, and Tamara Peyton. 2016. YouthTube: Youth Video Authorship on YouTube and Vine. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (CSCW '16), 1423–1437. DOI:https://doi.org/10.1145/2818048.2819961

[93] Svetlana Yarosh, Sarita Schoenebeck, Shreya Kothaneth, and Elizabeth Bales. 2016. "Best of Both Worlds": Opportunities for Technology in Cross-Cultural Parenting. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 635–647. DOI:https://doi.org/10.1145/2858036.2858210

[94] Svetlana Yarosh and Pamela Zave. 2017. Locked or Not?: Mental Models of IoT Feature Interaction. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 2993–2997. DOI:https://doi.org/10.1145/3025453.3025617

[95] Jason C. Yip, Tamara Clegg, June Ahn, Judith Odili Uchidiuno, Elizabeth Bonsignore, Austin Beck, Daniel Pauw, and Kelly Mills. 2016. The Evolution of Engagements and Social Bonds During Child-Parent Co-design. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 3607–3619.

Received April 2019; revised June 2019; accepted August 2019.