

Behind the Published Paper: Pilot Work and Methodological Visibility in HCI

AMNA LIAQAT, Cooper Union for the Advancement of Science and Art, USA

BELU TICONA, George Mason University, USA

NABILA CHOWDHURY, University of Toronto, Canada

ANTONIS ANASTASOPOULOS, George Mason University, USA

HCI papers show us what researchers ultimately did, but often obscure how consequential methodological decisions took shape. We examine pilot studies as a window into this hidden preparatory work. Through a qualitative meta-analysis of 119 CHI papers and interviews with five experienced HCI researchers, we find that “pilot study” describes not a single practice, but a family of activities for reducing uncertainty across research traditions. Our interviews further reveal a gap between research practice and the published record: consequential preparatory work often shapes a study without being identified or reported as a pilot. We argue that the central problem is not definitional inconsistency, but methodological traceability: whether readers can understand how consequential research decisions were reached. We translate this framing into recommendations for authors, reviewers, and the HCI community, arguing that making methodological reasoning more legible can also help newcomers navigate HCI’s often tacit methodological norms.

CCS Concepts: • **Human-centered computing** → **HCI theory, concepts and models**; *Empirical studies in HCI*.

ACM Reference Format:

Amna Liaqat, Belu Ticona, Nabila Chowdhury, and Antonis Anastasopoulos. 2018. Behind the Published Paper: Pilot Work and Methodological Visibility in HCI. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 21 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

“Should we report our pilot study in the paper?” HCI researchers routinely confront this question when deciding how and when to describe preliminary work in a manuscript. Should formative interviews be reported as a pilot? Does iterative prototype testing count as pilot work? Is a feasibility assessment fundamentally different from a pilot study, or simply another name for the same activity? These questions arise because HCI researchers routinely conduct preparatory work before larger empirical investigations [39]. Researchers refine interview protocols, iterate on prototypes, calibrate technical systems, assess recruitment strategies, and identify unforeseen challenges *before* investing in larger empirical studies. Such preliminary work improves study design, reduces methodological risk, and shapes the evidence ultimately presented in published research [1, 13]. Yet, there is little shared understanding regarding which activities constitute a pilot study, what that label is intended to communicate, and which, if any, preparatory activities should be reported.

Across disciplines, methodological researchers have observed that commonly used methodological terms often evolve to encompass diverse research practices, leading to inconsistent terminology and conceptual ambiguity [15, 18]. HCI

Authors’ Contact Information: Amna Liaqat, amna.liaqat@cooper.edu, webmaster@marysville-ohio.com, Cooper Union for the Advancement of Science and Art, New York, USA; Belu Ticona, George Mason University, Fairfax, USA, mticonao@gmu.edu; Nabila Chowdhury, University of Toronto, Toronto, Canada; Antonis Anastasopoulos, George Mason University, Fairfax, USA, antonis@gmu.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

Manuscript submitted to ACM

Manuscript submitted to ACM

has itself developed a substantial tradition of examining its own research practices, with recent work characterizing such inquiry as *meta-research*: the systematic study of how HCI research is conducted, evaluated, documented, and disseminated [40]. This work has exposed variation in methodological concepts and norms including sample size [8], reliability [37], literature reviews [51], and even foundational concepts such as interaction [23]. Yet discussions of pilot studies have largely focused on whether and how pilot work should be reported [1, 39], but not on what constitutes a pilot. As we demonstrate, the term in HCI has instead become an umbrella label for preparatory activities that differ substantially in their goals, methods, and expected outcomes.

This ambiguity has practical consequences for the HCI community. When fundamentally different preparatory activities are described using the same methodological label, authors receive limited guidance regarding what pilot work should be reported, reviewers develop different expectations for evaluating preliminary research, and readers are left to infer what was actually done from limited methodological descriptions. More broadly, these questions concern how methodological norms become established and communicated within a research community. Pilot studies provide a useful case through which to examine these processes because the same label appears across substantially different HCI research traditions.

To address this gap, we examine pilot work through two complementary perspectives. First, we analyze how pilot studies are represented in full CHI papers published between 2015 and 2025. Through iterative *human* qualitative coding, we characterize the motivations, procedures, outcomes, and downstream effects authors attribute to pilot work. Second, we conducted semi-structured interviews with five experienced HCI researchers spanning computational social science, interactive systems and VR, wearable and accessible computing, and social computing. These interviews allow us to examine a dimension that published papers alone cannot reveal: how researchers decide what preparatory work to conduct, what experts consider a pilot, and which parts of that process become visible in the final paper.

Across these analyses, we find that *pilot study* does not denote a single methodological activity. The term is used for a family of practices that address different uncertainties at different points in a research process, including refining study protocols, testing technical infrastructure, iterating prototypes, probing feasibility, and informing design decisions. Our interviews further show that the published pilot is not necessarily the pilot work: researchers routinely described informal testing, debugging, calibration, and iterative refinement that shaped the eventual study but was not reported as a distinct pilot. Reporting decisions were shaped not only by methodological relevance, but also by judgments about contribution, narrative coherence, space, and anticipated reviewer scrutiny.

These findings shift the problem from definition to communication. We do not argue that HCI needs a single universal definition of a pilot study, nor that every preparatory activity should be documented in detail. Instead, we argue that readers need enough methodological clarity to understand consequential decisions in the research process. We contribute (1) an empirical, manual characterization of the diverse functions and reporting practices associated with pilot work in CHI papers; (2) an interview-based account of how experienced HCI researchers conceptualize, conduct, and selectively report preparatory work; and (3) a methodological traceability framework for communicating consequential preparatory work by making clear what was uncertain, what the preparatory work revealed, and what decision followed.

2 Related Work

2.1 Meta-Research on HCI Research Practice

A growing body of HCI meta-research has increasingly turned its attention toward methodological concepts themselves, asking not only how research methods should be conducted but also how methodological concepts are interpreted, operationalized, and communicated across the field. Rather than assuming that widely used methodological labels correspond to shared research practices, recent literature has shown that many methodological concepts encompass substantial diversity in both interpretation and application.

For example, Stefanidi et al. [51] found that HCI researchers use the term *literature review* to describe a wide range of review methodologies, arguing that the field lacks a comprehensive terminology system for characterizing literature review practices. McDonald et al. [37] demonstrated that concepts surrounding reliability and inter-rater reliability are interpreted differently across qualitative research traditions, leading to inconsistent expectations among authors and reviewers. Caine [8] likewise showed that commonly held assumptions regarding sample size often conflict with the methodological goals of HCI studies, arguing that methodological decisions should be evaluated within their broader research context rather than according to universal expectations. Collectively, these studies suggest that methodological labels often mask substantial diversity in the research practices they are used to describe.

These conceptual differences become particularly visible when methodological labels serve as the primary means through which researchers communicate what they did. Bowman et al. [7] found considerable variation in how thematic analysis is conducted and reported in healthcare HCI research, demonstrating that researchers use the same methodological label to describe diverse analytic practices. Similarly, Singh et al. [48] observed substantial variation in how positionality is incorporated into CHI publications, arguing that positionality statements are frequently treated as markers of methodological rigor without clearly articulating how they informed the research process. Rogers et al. [45] likewise found considerable inconsistency in how systematic reviews are defined and reported across CHI publications, motivating best practices to improve methodological transparency and reporting quality. These studies show that methodological diversity becomes difficult to interpret when shared labels no longer reliably communicate the methodological intent underlying research practices.

Our work extends this meta-research tradition by examining pilot studies as both a methodological concept and a window into preparatory research practice. Rather than asking only whether pilot studies are reported adequately, we examine what researchers use pilot work to accomplish, how they understand its boundaries, and which parts of that work become visible in publication.

2.2 Methodological Terminology as Scientific Infrastructure

These observations align with a broader body of methodological scholarship arguing that shared terminology plays a central role in scientific communication. As research practices diversify, methodological concepts often evolve beyond their original definitions, resulting in variation in how researchers interpret and apply commonly used terminology. Rather than treating these differences as merely semantic, recent work has argued that inconsistent methodological terminology hinders scientific communication, complicates methodological learning, and limits the accumulation of methodological knowledge.

For example, Gentles and Vilches [15] argue that shared terminology is essential for communicating qualitative research methods and call for greater conceptual clarity in sampling terminology. Similarly, Guetterman et al. [18] identify inconsistent terminology as a persistent challenge in mixed methods research, arguing that continued methodological

development depends on maintaining a shared vocabulary for describing research practices. Collectively, this work positions methodological terminology as an important object of methodological inquiry rather than simply a means of describing research.

This perspective aligns with broader scholarship on scientific knowledge organization. Bowker and Star [6] argue that classification systems constitute infrastructure that shapes what becomes visible, comparable, and communicable within scientific communities. Likewise, Star and Griesemer [50] describe shared conceptual artifacts as enabling coordination among communities with different goals by providing sufficiently stable meanings while remaining flexible enough to accommodate local practices. These perspectives suggest that shared conceptual systems play an important role in enabling scientific communication and cumulative knowledge development. Importantly, this scholarship does not argue against methodological diversity itself. Rather, it argues that scientific communities require conceptual systems capable of communicating that diversity clearly. Consequently, examining how methodological terminology is interpreted and applied in practice provides an empirical foundation for understanding how research communities communicate methodological knowledge and establish shared methodological norms.

2.3 Pilot Studies Are Underspecified in HCI

In several disciplines beyond HCI, pilot studies are widely recognized as an important component of the research process, providing opportunities to refine study procedures, assess feasibility, identify methodological challenges, and improve the design of subsequent investigations. Consequently, a substantial body of methodological literature has focused on clarifying the role of pilot studies and developing recommendations for their design and reporting. For example, Arain et al. [1] reviewed the use of pilot studies in clinical research and argued that pilot work is frequently misunderstood or misused, motivating clearer guidance regarding its purpose and appropriate application. Similarly, Eldridge et al. [13] developed a conceptual framework distinguishing pilot and feasibility studies, proposing standardized terminology to improve consistency in methodological reporting. These studies seek to establish clearer definitions of pilot studies and provide guidance for how they should be conducted and communicated.

Within HCI, comparatively little attention has been paid to pilot studies as a methodological concept. Oppenlaender et al. [39] examined pilot study reporting practices in crowdsourcing research, finding considerable variation in how pilot studies are documented and proposing recommendations for improving reporting quality. This work represents an important step toward improving transparency surrounding pilot research within HCI. However, like prior work across other disciplines, it primarily focuses on *how* pilot studies should be reported, implicitly assuming that *pilot study* refers to a coherent methodological concept.

This assumption has received little empirical scrutiny. Existing guidance generally presumes that pilot studies constitute a single methodological concept for which common reporting expectations can be established. Yet, as prior HCI meta-research has shown for methodological concepts such as literature reviews, reliability, and sample size, widely adopted methodological labels often encompass diverse research practices whose differences are not apparent from terminology alone [8, 37, 51]. Whether *pilot study* similarly functions as a single methodological concept, or instead serves as an umbrella label for multiple forms of preparatory research, remains unknown.

Addressing this question is important for both methodological communication and reporting. If pilot studies serve fundamentally different methodological functions, then uniform reporting expectations may obscure rather than clarify the role that preparatory research plays within empirical HCI. This observation motivates the central question of our work: *How do HCI researchers use the term pilot study in practice, and what methodological functions does it communicate?*

3 Methodology

We used a two-part qualitative meta-research design to examine both the published representation of pilot work and the research practices that precede publication. The first component manually analyzes pilot studies explicitly reported in CHI papers, allowing us to characterize how authors describe their purposes, procedures, outcomes, and effects on subsequent research. The second component consists of semi-structured interviews with experienced HCI researchers, allowing us to examine how preparatory work is understood and conducted in practice, including work that may never appear in the final paper. We treat these sources as complementary rather than interchangeable: the corpus captures what becomes visible in the published record, while the interviews help explain how and why researchers make particular activities visible.

3.1 Corpus Creation

We focus our document analysis on full CHI papers published between 2015 and 2025. CHI provides a large and methodologically diverse body of HCI research, making it a useful site for examining how a widely used methodological term is operationalized across different research traditions. Our goal is not to treat CHI as representative of every HCI venue, but to characterize pilot reporting within a major HCI publication venue and then contextualize those patterns through interviews in which participants reflected on their broader HCI research experience.

Pilot work can appear at very different levels of visibility: as a brief sentence embedded in a methods section, as a short paragraph, or as a dedicated pilot or preliminary-study section. Because the present manual analysis requires sufficiently detailed descriptions to code methodological purpose and downstream impact, we restrict the primary corpus to papers with a dedicated heading that identifies pilot or closely related preparatory work. This choice prioritizes analytic depth over prevalence estimation. It also means that our corpus does not capture the full amount of pilot work occurring in CHI, a limitation we return to later.

Our processing pipeline consisted of the following steps:

- (1) **Step 1: Data Collection.** We queried from the ACM Digital Library any paper that contains any term from a curated-list of pilot-related keywords. Previous studies [39] have shown that pilots can have multiple ways to be reported, therefore we conducted a sensitivity study to understand how much to expand the keyword list (see appendix). We filtered ‘full papers’ only, excluding other submission types such as ‘extended abstract’, since our study aimed to understand preparatory work conducted before the main study is finalized HCI publications (rather than work where the pilot is the main study, or where the publication describes an early work-in-progress). This yielded 1098 papers in total.
- (2) **Step 2: Text-Extraction.** We extracted the paper content using a layout-aware PDF parsing pipeline that uses MinerU [54] as a core-engine to identify heading labeled-blocks. We qualitatively evaluated this pipeline against other tools, such as PyMuPDF [49], on a set of 12 randomly selected papers. We observed that the MinerU pipeline handled hierarchical headings better or the same as PyMuPDF five out of 12 times. We reconstructed the papers joining the identified headings, discarding the content before the abstract and after the conclusions/before the appendix, keeping only the paper content organized by headings.
- (3) **Step 3: Headings-Selection.** We analyzed the heading distribution to identify how pilots are referred in publications. We defined three filtering criteria, papers with: a) ‘pilot study’ or ‘pilot studies’ as headings, b) headings containing ‘pilot’ related words, c) headings containing referring to preparatory work beyond the wording ‘pilot’ (e.g. formative study, ‘feasibility study’, ‘preliminary work’). We focused our study on papers that

fulfill criteria a) and b), manually choosing only those referring to a ‘pilot’ related work. Our final corpus for manual coding consisted of 119 full CHI articles.

3.2 Coding Procedure

3.2.1 Initial codebook development. Three coders, the first three authors, independently coded 10 randomly selected papers. We approached this first round inductively, attending to what authors said motivated the pilot, how the activity was conducted, what was reported as an outcome, and how the pilot affected subsequent research. The coders then met to compare interpretations, discuss disagreements, and consolidate recurring distinctions into an initial codebook. This process produced 27 codes. For each code, we developed a working definition and inclusion criteria to support consistent application in subsequent rounds.

3.2.2 Codebook Stabilization. We conducted two additional rounds of independent coding, following an aggregated sampling criteria, selecting 2 paper per year from 2015 to 2025. After each round, the three coders met to resolve disagreements and refine the codebook by clarifying definitions, revising inclusion criteria, and merging or splitting codes where needed. At the end of this process, 30 articles had been independently annotated and discussed in depth, and the codebook contained 35 codes that cover the motivation, procedure, reporting, and downstream consequences of the pilot work.

3.2.3 Reliability and corpus annotation. We conducted stratified sampling by publication to select 20 additional papers for reliability assessment. We excluded one paper because it did not meet our inclusion criteria (pilot study was the main study), leaving 19 papers: 5 for initial calibration and 14 for final inter-annotator assesment. We evaluated reliability for six main coding decisions: four binary variables recording whether authors explicitly described participant recruitment, pilot protocol, pilot results, and distinct pilot impact unreported in the pilot section; and two multi-label variables covering the motivation and impact of the pilot. After the early calibration stage, the authors met and analyzed the disagreements finding that a final refinement of the codebook was needed for the multi-label variables. A single code was too restrictive since pilots can genuinely serve more than one function, and forcing coders to choose one option between two supported interpretations could create apparent disagreement. Alternatively, allowing three or more codes would allow overlapping interpretation and weaken the paper’s classification. Thus, we decided to limit the labels up to two codes, agreeing on selecting different labels only when required to capture different aspects of the motivation and impact of each study. After finalized the codebook, two authors independently annotated the 14 papers left, agreeing on 70 out of 84 decisions (83,3%). Because our variables differed in structure, we assessed binary codes using exact agreement and nominal Krippendorff’s α , and the multilabel Motivation and Impact codes using exact-set agreement, mean Jaccard similarity, and Jaccard-distance α . Motivation achieved 64.3% exact agreement (Jaccard = .774; α = .702), while Impact achieved 78.6% (Jaccard = .857; α = .733). Exact agreement for the binary variables ranged from 78.6% to 100% (see Appendix for detailed results).

3.2.4 Corpus Annotation. After the codebook was finalized, we then applied it to the 69 remaining papers in the corpus. The same two authors evenly annotated half of them each independently. Additionally, one of these coders revisited the first batch of 30 papers coded and mapped the older codes to the final version of codebook.

The analyses reported below focus on patterns that were stable across the coded corpus, particularly the relationship between why pilot work was undertaken, what it changed, and how those consequences were communicated.

Table 1. Expert interview participants and their primary research expertise.

ID	Research Expertise
P1	Computational social science
P2	VR/AR
P3	Accessibility and wearable computing
P4	Social computing
P5	Domestic technologies and tangible computing

3.3 Expert Interviews

To complement the document analysis, we conducted semi-structured interviews with five experienced HCI researchers. We purposefully sought researchers with substantial publication and reviewing experience whose work represented different methodological traditions within HCI. Participants' primary areas included computational social science, interactive systems and virtual reality, wearable and accessible computing, and social computing. We refer to participants as P1–P5. As our participants are all prominent members of the HCI community, we report only their area of expertise in Table 1 for confidentiality. However, all participants publish extensively at HCI venues, hold professorship positions, are leading research labs, have served on organizing committees, frequently review papers, and are engaged in additional service work for the community like serving on ethics and research grant review boards.

Interviews lasted approximately 30–45 minutes and asked participants to reflect on concrete examples of preparatory research from their own work. We asked how they distinguish pilot studies from related activities such as formative studies or iterative testing; what they use pilots to learn; how they decide when a study is ready to proceed; whether and how they report preparatory work; and what they expect when reviewing papers that report pilots. In the latter portion of each interview, we showed examples from the corpus and discussed emerging dimensions of the codebook. We deliberately grounded the interviews in participants' experiences rather than asking them to produce a single abstract definition of a pilot study.

We analyzed the interview transcripts iteratively alongside the document analysis. The research team reviewed each transcript, identified recurring explanations for the diversity and visibility patterns observed in the corpus, and compared areas of convergence and disagreement across participants. We treated the interviews as contextualizing evidence rather than as a representative sample of HCI researchers: their role was to illuminate reasoning and research work that cannot be recovered from published papers alone.

4 Findings

Our analyses reveal a distinction between *pilot studies as published objects* and *pilot work as research practice*. We first characterize what explicitly reported pilot studies make visible in the published record. We then draw on interviews to examine how researchers understand, conduct, and selectively report preparatory work in practice, incorporating corpus evidence where it provides corresponding or contrasting insights. Together, these perspectives reveal both what published pilot reports make visible and what they leave behind.

4.1 Pilot Studies as Published Objects

Over the past decade, HCI researchers have shown growing interest in how methodological work is named and reported, including pilot studies [39, 51]. Papers giving pilot-related work a dedicated heading grew 2.7-fold, broadly tracking

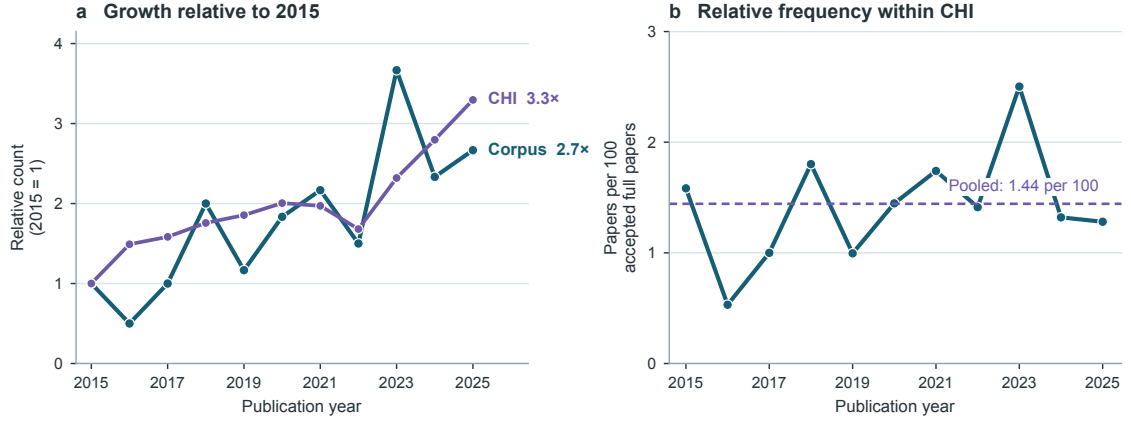


Fig. 1. Growth and relative frequency of CHI papers in our corpus, 2015–2025. (a) Counts indexed to 2015 compare papers with a dedicated heading for pilot or related preparatory work with all accepted CHI full papers. (b) The annual number of these dedicated-heading papers per 100 accepted CHI full papers; the dashed line marks the pooled rate across the period.

CHI’s 3.3-fold growth (Figure 1). Yet these papers remained uncommon, at 1.4 per 100 accepted full papers, with no sustained annual trend. This rate captures only explicitly labeled pilot work and not how often pilots were conducted. As the following sections show, pilots reported under these headings also served different purposes.

Using Wobbrock and Kientz’s taxonomy [57], we also classified the papers’ primary contributions. Empirical studies and interface artifacts or techniques were the most common, while the remaining papers included case studies, methodological contributions, datasets, and replication (Figure 2).

4.1.1 Pilot Reports Make Procedures and Results Visible but Leave Participant Selection Unclear. Pilot reports most often made the procedure, results, and numeric participant count visible, whereas recruitment and the relationship between pilot and main-study participants were less consistently documented (Figure 2). These results show that pilot reports were uneven in what they documented. Coders could often reconstruct what occurred in the pilot and what the authors learned, but not how participants were recruited or whether their participation and data were distinct from the main study. When reported, in most cases coders were able to identify differences between the pilot and main study participant’s although authors do not provide an explicit reason (e.g. [17, 24, 34, 65]). More interestingly, some pilots resulted in the researchers narrowing down their target participants for the main study (e.g. [38, 47]). In other cases, pilots were run with lower risk participants to check whether a study was safe before running the main study with a vulnerable population (e.g. [58]). For example, Loke et al. [35] mention that “A pilot (...) was conducted with the same clinicians (...) [who] identified aspects of the activity that needed improvement before conducting the activity with a child and parent”. Some work even reflected on different results between the pilot study and main study, despite using the same type of participants: While et al. [56] state “we conducted four pilot studies involving volunteer older adults (...) to assess the study design and procedure, which were not tailored to older adults.”, later relating this observation to the main study results: “The discrepancies between the pilots and the main study underscore the importance of adapting the study design with real-time feedback and participant needs”.

In other cases, it was preferred to recruit participants with the same target profile by proximal/convenient sampling, valuing the high-quality feedback from HCI-familiarized participants (e.g., lab peers, researchers) on an early prototype,

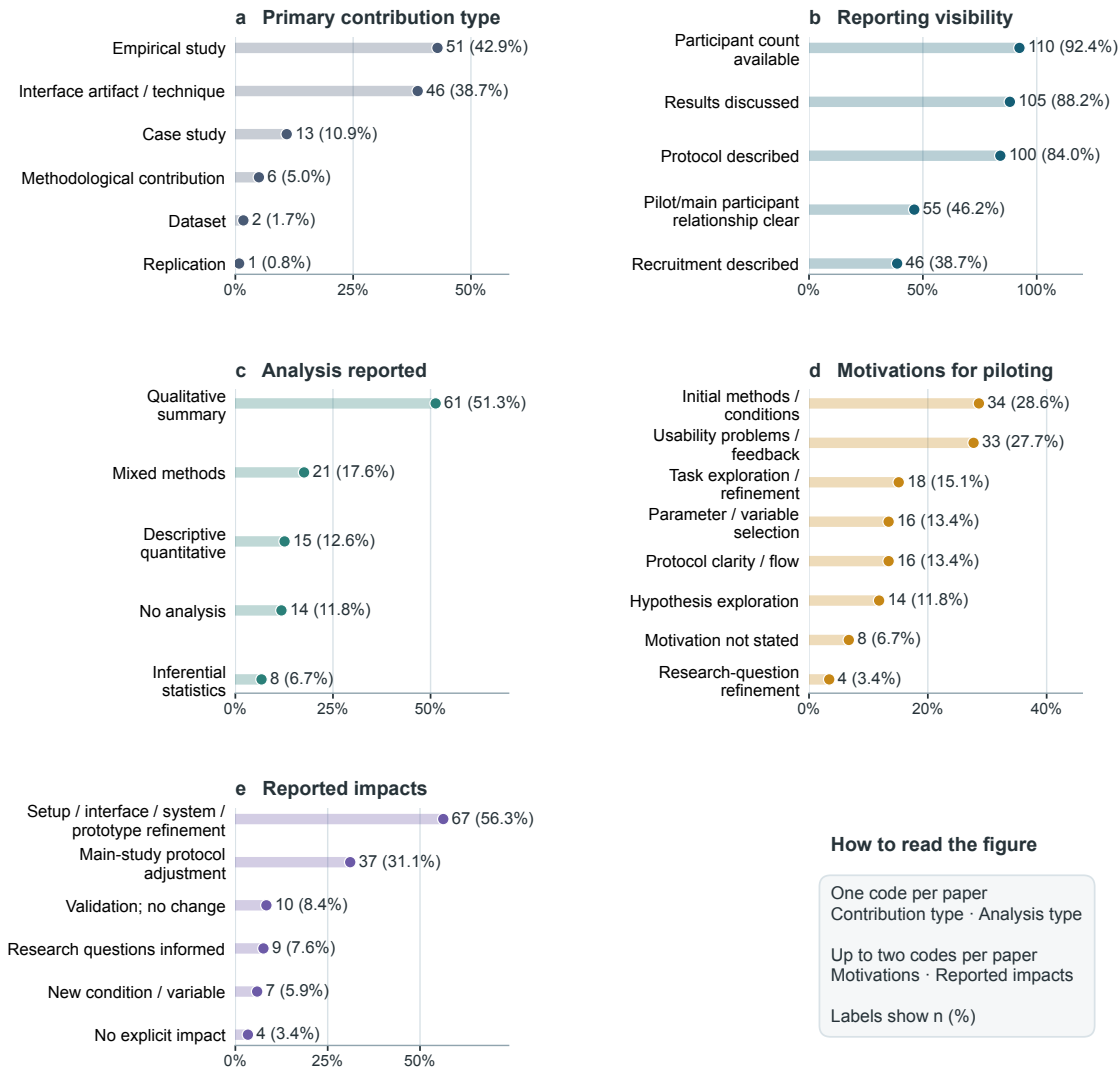


Fig. 2. Corpus summary ($N = 119$); labels show n (%). Empirical studies and interface artifacts or techniques were the most common contribution types (a). Participant counts, results, and protocols were usually reported, while recruitment and the pilot/main-study participant relationship were less often clear (b). Qualitative summaries were the most common form of analysis (c). Initial methodological insight and usability feedback were the most frequent motivations (d), and refining a setup, interface, system, or prototype the most frequent impact, followed by adjusting the main-study protocol (e). Each paper received one primary contribution and analysis code; motivations and impacts allow up to two codes per paper and therefore need not sum to 100%.

or study design. For example, in [28], where the target profile was older adults, and a retired HCI researcher and an older adult with no HCI background were recruited. Even though we do not report specific demographics stats as part of our analysis, qualitatively we observed that pilots tend to report gender and age mainly, including additional characteristics (e.g., language proficiency, nationality, work or educational background).

4.1.2 Why Pilot Studies Are Conducted: Addressing Different Forms of Uncertainty. The two most common reasons to conduct a pilot were to understand the study methods and experimental conditions, and to gather user feedback about usability problems. Pilots in the first group probed different parts of the setup: whether an experimental manipulation or condition produced its intended effect (e.g., [10, 21]); whether the apparatus and measures were reliable enough to collect usable data (e.g., [27, 32, 33]); and whether the overall method was feasible before committing to it (e.g., [11, 26]). Where the goal was user feedback, pilots differed instead in *when* in a tool’s development that feedback was sought. In the earliest case, authors used a basic tool, object, or channel in an exploratory study to observe interaction and discover requirements in a specific setting (e.g., [2, 28, 53]). To illustrate, Holloway et al. [22] gathered feedback on how blind users worked with 3D-printed models, and Du et al. [12] sought to understand how participants track moving objects. Once a system of their own existed, authors instead showed participants an early prototype they had built (e.g., [20, 52, 64]). Least often, the focus of the pilot was comparing a tool or prototype against a baseline or alternative tools to assess whether the intended design goals had been met (e.g., [47, 62, 63]). User feedback was therefore treated as informative across the whole arc of development, not only once a system was complete enough to evaluate. Most papers stated a motivation explicitly, but some remained vague, and in others we found a mismatch between the motivation the authors stated and the uncertainty the pilot actually addressed.

Because pilots were most often run to understand study methods or obtain user feedback, the evidence they produced leaned toward observation rather than formal hypothesis testing. Qualitative summaries (observations and participant quotations) were by far the most common form of reported analysis, followed by mixed-methods and then descriptive quantitative summaries. A smaller set of papers reported no analysis, and inferential statistics were rarest of all (Figure 2). These differences caution against treating pilot studies uniformly as smaller versions of a later experiment: the reported evidence was frequently oriented toward learning what to revise, retain, or investigate next.

4.1.3 From Pilot Evidence to Study Decisions: Concrete Changes or Validation. Pilot studies most often changed the main study in one of two ways: by refining the system, interface, or experimental setup (56.3%, $n = 67$) and, less frequently, by adjusting the study protocol (31.1%, $n = 37$). Within the refinement group, three targets recurred, and a paper could name more than one. About half of these papers changed the study or system configuration, settling concrete values and settings before the main run ($n = 35$; e.g., [16, 29, 31]); almost as many revised the interface or the interaction itself, changing what participants saw or how they acted on it ($n = 32$; e.g., [3, 25, 41]); and a smaller group revised design principles or goals rather than a specific setting or control ($n = 10$; e.g., [59, 60]). These changes were usually stated rather than left implicit, with only 3.4% of papers ($n = 4$) reporting no explicit impact. Goguey et al. [16], for instance, reported that “we therefore used these values of 250 ms and 750 ms for the main experiment,” while Kopeinik et al. [30] gave the connection its own “Implication for the Main Study” section, letting readers follow the pilot through to the design it justified. A further 16.0% of papers ($n = 19$) attributed a change to the pilot somewhere other than the pilot section itself. In Cauchard et al. [9], for example, the pilot section reports no findings, “we only report the results from the main study”, yet a safety instruction elsewhere is described as “added after the pilot when some participants answered while driving”. Likewise, Yang et al. [61] fixed image dimensions, batch size, and training hyperparameters “based on this pilot study” without reporting what that study showed. In these cases the decision is attributed to preparatory evidence the paper never presents, so readers cannot judge what it rests on.

4.2 Pilot Work in Research Practice

The corpus characterizes what authors make visible when they explicitly report pilot work. Our interviews provide a complementary view of how such work is understood and conducted in research practice, including preparatory activities that may never appear as pilots in publication. In this section, we organize the interview findings around four themes and draw on the corpus where it provides corresponding or contrasting evidence.

4.2.1 Pilot Studies Are a Family of Decision-Making Practices. Participants did not describe pilot studies as a single methodological activity. Instead, their accounts echoed the functional diversity visible in our corpus while revealing why that diversity arises: researchers use preparatory work to resolve different uncertainties within different methodological traditions. The interviews reinforced this functional diversity and showed that even experienced researchers draw the boundaries of “pilot study” differently. P2, whose work centers on interactive systems, initially distinguished formative studies, pilot studies, and formal evaluations temporally: formative work identifies needs before a system is built, whereas pilot work tests an existing prototype or protocol before a formal study. Yet P2 also described iterative usability testing of early prototypes as pilot work, even when this testing occurred without an explicit research question. P4, also a systems-oriented HCI researcher, articulated a similar distinction between formative interviews that establish user needs and pilot testing conducted after a system exists, while noting that interview protocols themselves may also be piloted.

Other participants organized the space differently. P1 described one meaning of pilot in HCI systems work as informal stress-testing of a developing system, but described a second meaning in preregistered experimental research: the sequence of studies conducted before preregistration to establish feasibility, manipulation strength, outcomes, and protocol details. P3 described pilots that tested protocol timing, device placement, battery life, sensor sampling, accessibility of question formats, and interaction modalities. For P3, whether these activities counted as pilot work depended on the context and the uncertainty being addressed.

P5 similarly described encountering substantially different uses of the term across HCI. In her own qualitative work, a pilot typically involves only a few participants whose data are used to refine a protocol rather than contribute to the eventual analysis. Yet she had also encountered papers describing substantially larger early-stage studies as pilots: “I’ve heard Pilot referred to in a lot of different ways.” These differences had changed her own understanding over time; although she initially learned piloting as a small-scale rehearsal of a study protocol, she had “since learned that it can be a lot broader.”

Together, these accounts suggest that the variation in our corpus is not simply a case of researchers applying an otherwise stable methodological category inconsistently. Researchers themselves encounter and inherit different understandings of what pilot work is, shaped by research tradition, method, and the uncertainty being addressed. As P3 observed, a pilot for an interview study in CSCW need not resemble a pilot for testing a new headset in VR. P4 similarly attributed local differences in method reporting to HCI’s combination of design, psychology, computer science, and other disciplinary backgrounds. Rather than identifying one canonical pilot-study procedure, our data characterize pilot work as a family of practices for making research decisions under uncertainty.

4.2.2 Pilot Work Extends Beyond What Becomes Visible in Published Papers. The corpus captures only work that authors chose to identify and describe as a pilot or related preparatory activity. The interviews made clear that this published record represents only a portion of the preparatory work researchers actually conduct. Participants described informal

testing, debugging, technical calibration, protocol rehearsal, early interviews, and repeated system iteration that shaped their studies but was never reported as a discrete pilot.

P1 provided the clearest articulation of this gap, describing “this part of the effort to make [sure] the experiment or the system works” as work that goes “widely under-documented in the paper.” In one systems project, friends and collaborators used an early version repeatedly while the system was being adjusted, but the final paper did not identify this process as a formal pilot. In a more quantitative project, by contrast, the team ran four or five paid pilot studies before preregistration to stabilize a complicated multi-stage protocol. P1 also described months of failed or partial attempts required to make an automated data-collection pipeline work. Although these activities were consequential to the eventual research, they did not appear in the publication under the label *pilot study*.

The same pattern appeared in the other interviews. P2 described loosely testing early prototype versions with users until the system was ready for formal evaluation. P3 described checking camera placement, sensor settings, memory capacity, battery life, and question formats before a main study; such adjustments might be mentioned as part of the method but not presented as pilot results. P4 likewise said that pilot work is often embedded within the method section rather than given its own heading, and that informally conducted pilots may receive little or no mention at all. P5 described a similarly routine process in qualitative research: pilot an interview or study protocol, revise questions or procedures, amend the ethics protocol if necessary, and then proceed with the main study. Yet these iterations rarely appear in the eventual paper. As she put it, “Normally, I don’t report anything from the pilot study, unless it’s something [...] significant or interesting. Normally, the pilot study stays internal.” Thus, methodological changes produced through piloting may become incorporated into the final protocol even when the process through which those decisions were reached is absent from the published account.

This distinction matters for interpreting the document corpus. A missing pilot section does not imply that no preparatory work occurred, and even an explicitly reported pilot may represent only one selected moment in a longer process of iteration. The published pilot is therefore not necessarily equivalent to the pilot work. Published papers expose selected parts of the process through which a study becomes research-ready.

4.2.3 What Pilot Work Communicates Matters More Than What It Is Called. Across both datasets, methodological purpose and consequence were more informative than the label attached to an activity. In the corpus, papers sometimes provided substantial procedural detail while leaving the motivation for the pilot or its effect on the main study difficult to infer. Conversely, relatively brief descriptions could be interpretable when they clearly stated why the pilot was conducted and what it changed. This suggests that reporting quantity and reporting usefulness are not equivalent.

Participants repeatedly made the same distinction. P2 described themselves as pragmatic about terminology: whether an activity was called a formative study, a pilot, or Study 1 mattered less than whether the paper made clear what knowledge it produced and why the study had been conducted. P2 emphasized that, as a reviewer, a recurring problem was not missing procedural detail but a missing study goal: papers would list findings from a formative or preliminary study without explaining what question motivated it in the first place. P3 likewise argued that terminology is shaped by disciplinary and community backgrounds and need not be standardized for its own sake. What mattered was whether authors clarified their intention. For example, testing feasibility, checking setup, exploring a hypothesis, or informing a methodological change.

The interviews also emphasized the connection between preparatory work and subsequent decisions. P3 described the main value of reporting as justifying decisions about study length, data collection, interaction format, and communication channels. P4 similarly noted that when pilot work is mentioned, the most useful information is often how the protocol

was improved as a result. P1, after discussing examples from our corpus, reformulated the issue as the need to tell readers “what aspects of the main study were informed by the pilot.”

P5’s reaction to examples from our corpus made this interpretive problem particularly visible. When shown a paper that briefly reported a pilot without explaining its relationship to the subsequent study, her immediate questions were: “What was the point of the pilot study? What did you learn?” She concluded that the description “raises more questions than it answers.” When asked what an informative pilot report should communicate, she emphasized not exhaustive procedural detail but the relationship between the preparatory work and what followed: “I think that’s the most important part [...] what changed as a result of these findings?” Her response closely mirrors the distinction emerging across our interviews: pilot reporting becomes useful when readers can understand not simply that preparatory work occurred, but what uncertainty it addressed and what subsequent decision it informed.

Taken together, these observations suggest that the communication problem is not solved by requiring more detail in every pilot section. A minimally reported pilot can be interpretable when its purpose and consequence are explicit, while a richly described pilot can remain difficult to interpret when readers cannot determine why it was conducted or what it informed. The central reporting relationship is therefore between an uncertainty, the evidence or experience produced through preparatory work, and the consequential decision that followed.

4.2.4 Reporting Reflects Judgments of Relevance as Well as Publication Incentives. Participants did not describe pilot reporting as a simple commitment to either transparency or omission. Instead, they made selective judgments about what information was scientifically useful, narratively appropriate, and strategically safe to include in a publication. This helps explain why consequential preparatory work may disappear even when researchers value methodological transparency.

One criterion was perceived contribution. P3 described a rough hierarchy: a routine technical check that revealed nothing notable might warrant only a sentence in the methods; preparatory work that led to consequential methodological or design changes deserved fuller explanation; and a pilot that generated novel knowledge useful to other researchers might become a substantive finding in its own right. P2 described a similar pattern, noting that pilots were rarely emphasized unless they yielded key insights or new knowledge, whereas formative studies were more likely to be reported because their findings could directly motivate design requirements. P4 likewise said that pilots are difficult to integrate when the final evaluation is intended to carry the paper’s main contribution.

A second criterion was the anticipated effect of reporting on review. P1 was particularly explicit that authors may “gain very little” from reporting a pilot while opening additional aspects of the research to scrutiny. Informal or messy iteration could make a paper appear disorganized even when it was part of a reasonable research process. P4 independently raised the possibility that reviewers may challenge a pilot when it is reported prominently without an equally clear contribution, and noted that even in venues without strict page limits authors are expected to keep the paper’s length proportional to its contribution.

At the same time, P3 offered an important counterpoint from the reviewer perspective: seeing that authors had piloted a study could signal an “extra level of care” and greater methodological rigor, because the researchers had tested their assumptions before committing to the main study. P2 and P4 also expressed strong support for methodological transparency, suggesting appendices or supplementary materials as places to preserve detail without interrupting the main narrative.

P5 introduced an additional consideration shaping what becomes publishable: the ethical status of pilot data. In her own work, pilot participants often understand their participation as helping the researchers improve a subsequent

study rather than contributing data to the study’s substantive findings. She therefore distinguished using pilot data for methodological refinement from later pooling those data into the research analysis: “where I think it gets kind of [...] murky is if [...] the participants think that their data is being used to improve the study, and then their data is used to arrive at research conclusions.” For P5, reporting was therefore shaped not only by whether pilot findings were interesting, but also by what participants had consented to their data being used for.

These accounts reveal a tension in current publication practice. Authors may experience iterative preparatory work as something that complicates a clean research story or creates new surfaces for criticism, while reviewers may interpret the same work as evidence of care and methodological reflection. Reporting is therefore selective for two intertwined reasons: researchers make epistemic judgments about what readers need to know and strategic judgments about what belongs in a coherent, publishable account. This tension motivates our discussion of methodological visibility below.

5 Looking Ahead: Rethinking Methodological Reporting in HCI

Our findings began with an apparently terminological question: what does HCI mean by *pilot study*? Examining that question revealed a broader issue concerning how methodological knowledge becomes visible in HCI publications. Methodological discussions of pilot studies often present the pilot as a discrete activity that precedes a larger study [1, 13]. Our findings complicate this model. Participants described preparatory work as recurring throughout system development and study design. This embedding of preparatory work within iterative research practice resonates with accounts of design-oriented HCI research in which knowledge develops through processes of making, evaluation, and refinement [66]. HCI scholarship has cautioned against treating methodological diversity itself as evidence of inadequate standards, particularly where different research traditions operate according to different epistemological commitments [14, 19]. Participants themselves drew different boundaries around pilot, formative, feasibility, and iterative work, often in ways that reflected their research traditions.

This perspective helps explain why the term acquires different meanings across HCI. Different research programs encounter different uncertainties. But what they share is a role in reducing uncertainty before researchers commit to a consequential next step. This functional diversity reflects HCI’s broader methodological and epistemological pluralism, in which different research traditions produce knowledge through substantially different practices [19]. We argue that the central problem is therefore not the absence of a shared definition of pilot study, but whether consequential methodological reasoning remains visible in the published record. Published papers necessarily present a selective account of the research process; the goal cannot be to preserve every iteration, but to make it possible for readers to understand how consequential decisions were reached. We refer to this as *methodological traceability*. We next consider what this concept means in practice for authors, reviewers, and the broader HCI community.

5.1 For Authors: Report Decision Traces, Not Every Iteration

For authors, methodological traceability can be operationalized through three questions: *What was uncertain? What did the preparatory work reveal? What decision followed?* Answering these questions need not require a dedicated “Pilot Study” section. A consequential decision trace may require a paragraph; another may require only a sentence. Conversely, a lengthy account of preparatory work may provide little methodological value if its relationship to the eventual study remains unclear.

This concept also provides a basis for deciding what belongs in the paper and what remains elsewhere. The main paper should preserve the reasoning necessary to understand consequential methodological decisions. Procedural detail, such as extended protocols, intermediate versions, calibration results, or additional analyses, can be preserved

in appendices or supplementary materials when useful for reproducibility or reuse. The aim is therefore not “more reporting” in general, but more *informative* reporting: allocating limited space to the methodological history that clarifies how the final study should be understood.

Importantly, reporting what was learned from preparatory work is distinct from treating preparatory data as research evidence. Our interviews surfaced ethical concerns about reusing data collected from participants who may have understood their role as helping researchers refine a study rather than contributing evidence toward its substantive conclusions. Greater methodological visibility should therefore not imply that pilot data should automatically be pooled with main-study data or reported as findings. Authors should distinguish between making a methodological decision traceable and making claims from the data that produced it.

5.2 For Reviewers: Evaluate the Basis of Decisions, Not the Presence of Pilots

Our findings also suggest an opportunity for reviewers to shift from asking for specific methods to asking for evidence that support their methodological decisions. The presence of a pilot should not function as a proxy for rigor, nor should its absence imply that a study was insufficiently prepared. Instead of asking, “*Was this piloted?*”, reviewers can ask: “*What supports the consequential methodological decisions on which this study depends?*”

The answer need not be a pilot. A decision may be supported by prior literature, established methodological practice, formative research, technical testing, domain expertise, earlier studies, or preparatory empirical work. Conversely, merely stating that a pilot was conducted should not resolve a methodological concern if the paper does not explain what the pilot established. This reframes review from checking for the presence of a methodological activity to evaluating whether consequential choices are sufficiently justified and interpretable.

This shift may also reduce a tension surfaced in our interviews: authors can perceive reporting pilot work as creating additional surfaces for reviewer scrutiny, even as reviewers may interpret such work as evidence of methodological care. Review practices that focus on the reasoning supported by preparatory work, rather than treating deviations and revisions as weaknesses to defend, could make it easier for authors to communicate how studies actually developed. This recommendation also connects to broader calls for meta-research on how HCI’s peer-review practices and evolving expectations shape the research that authors communicate [40].

5.3 For the HCI Community: Build Infrastructure for Methodological Traceability

Methodological traceability should not depend entirely on individual authors knowing what reviewers expect. HCI venues could make these expectations explicit through lightweight reporting infrastructure. Author guidelines, submission forms, methodological checklists, or reviewer guidance could include a prompt such as:

If preparatory empirical work materially shaped the reported study, briefly describe what was uncertain, what the preparatory work revealed, and what consequential decision followed.

Such a prompt deliberately does not require authors to conduct a pilot, use the term “pilot study,” or follow a particular research workflow. It instead specifies the communicative outcome that matters when consequential preparatory work has occurred. This distinction is important for a field encompassing experimental, qualitative, design-oriented, systems, accessibility, and other research traditions. The HCI community could explore initiatives from other research communities towards methodological transparency [36, 42, 44], such as using checklists in submission platforms [4, 55], or other publishable artifacts [5, 46].

Future reporting infrastructure could go further. Structured supplementary materials could provide optional spaces for intermediate protocols, parameter choices, design iterations, or other methodological decision traces that are difficult to accommodate in the main paper. Research artifacts and repositories could preserve links between final methodological choices and the preparatory evidence that informed them. Future work could investigate whether such mechanisms actually help readers interpret and reuse research, which kinds of methodological decisions readers most need to trace, and when additional visibility becomes burdensome rather than useful.

Making these expectations explicit may be particularly important for researchers entering HCI from neighboring disciplines. HCI’s publication practices themselves contain community-specific conventions that researchers must learn [43]. One participant described pilot-study practice as an example of the “local norms” researchers often acquire informally. When expectations about what methodological reasoning should be reported remain tacit, newcomers must infer them by reverse-engineering published papers, mentorship, and review feedback. Making methodological traceability an explicit communicative principle would not eliminate the diversity of HCI research practice. Rather, it could make that diversity easier to navigate without requiring researchers to already know the field’s unwritten rules.

6 Limitations

This study has several limitations that bound our claims. First, our document corpus is restricted to full CHI papers published between 2015 and 2025. We therefore make claims about reporting patterns in this CHI corpus rather than estimating the prevalence or distribution of pilot practices across HCI as a whole. Our interviews broaden the interpretive scope by asking experienced researchers to reflect on their work across HCI venues and traditions, but five interviews cannot establish how common these perspectives are across the field.

Second, our primary corpus intentionally focuses on papers with dedicated headings for pilot or closely related preparatory work. This enabled detailed manual coding, but it biases the corpus toward comparatively visible and elaborated pilots. Our interviews suggest that informal and embedded pilot work may be especially likely to disappear from the published record. We therefore cannot estimate how much preparatory work goes unreported, and our claims about invisible work are grounded primarily in interview accounts rather than corpus prevalence.

Third, methodological categories inevitably simplify activities that may serve several functions at once. A single pilot can simultaneously reveal usability problems, test a protocol, calibrate technical parameters, and reshape a research question. We use the codebook as an analytic and communicative scaffold rather than as a claim that pilot activities naturally fall into mutually exclusive kinds.

Finally, our analysis captures what authors and interview participants chose to describe. It cannot reconstruct every decision made during a research project or determine retrospectively which omitted pieces of preparatory work would have changed a reader’s interpretation. Future work could examine a broader range of HCI venues, compare published papers with research artifacts such as preregistrations and supplementary materials, and study author–reviewer expectations directly to further develop community-informed reporting guidance.

7 Conclusion

Pilot studies offer a window into a broader feature of HCI research: studies do not arrive fully formed. They take shape through interactions with participants, prototypes, technical systems, methodological traditions, and earlier decisions. Our analysis shows that this preparatory work serves diverse functions and does not map neatly onto a single definition of “pilot study.” Rather than resolving this diversity through stricter terminology, we argue that the more consequential question is what methodological reasoning remains visible in the published record.

We propose *methodological traceability* as a lightweight principle for making that reasoning legible. When preparatory work materially shapes subsequent research, readers should be able to understand what was uncertain, what the preparatory work revealed, and what decision followed. This does not require documenting every iteration or adding a dedicated pilot section. It asks instead that consequential methodological decisions remain connected to the evidence and experience through which they were reached. As HCI continues to encompass diverse research traditions and welcome researchers into the community, making these decision traces more visible can help published papers communicate not only what research was ultimately conducted, but some of the reasoning through which it came to take that form.

References

- [1] Muhammad Arain, Michael J. Campbell, Christopher L. Cooper, and Gillian A. Lancaster. 2010. What Is a Pilot or Feasibility Study? A Review of Current Practice and Editorial Policy. *BMC Medical Research Methodology* 10, 67 (2010). doi:10.1186/1471-2288-10-67
- [2] Ahmed Sabbir Arif, Sunjun Kim, Wolfgang Stuerzlinger, Geehyuk Lee, and Ali Mazalek. 2016. Evaluation of a Smart-Restorable Backspace Technique to Facilitate Text Entry Error Correction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 5151–5162. doi:10.1145/2858036.2858407
- [3] Calvin S. Bao, Siyao Li, Sarah G. Flores, Michael Correll, and Leilani Battle. 2022. Recommendations for Visualization Recommendations: Exploring Preferences and Priorities in Public Health. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, 17 pages. doi:10.1145/3491102.3501891
- [4] Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-Repeatable Experiments and Non-Reproducible Results: The Reproducibility Crisis in Human Evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 3676–3687. doi:10.18653/v1/2023.findings-acl.226
- [5] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. doi:10.1162/tacl_a_00041
- [6] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. MIT Press. doi:10.7551/mitpress/6352.001.0001
- [7] Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. 2023. Using Thematic Analysis in Healthcare HCI at CHI: A Scoping Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23). Association for Computing Machinery, Article 491, 18 pages. doi:10.1145/3544548.3581203
- [8] 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). Association for Computing Machinery, 981–992. doi:10.1145/2858036.2858498
- [9] Jessica R. Cauchard, Janette L. Cheng, Thomas Pietrzak, and James A. Landay. 2016. ActiVibe: Design and Evaluation of Vibrations for Progress Monitoring. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 3261–3271. doi:10.1145/2858036.2858046
- [10] Tsai-Wei Chen and S. Shyam Sundar. 2018. This App Would Like to Use Your Current Location to Better Serve You: Importance of User Assent and System Transparency in Personalized Mobile Services. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 13 pages. doi:10.1145/3173574.3174111
- [11] Andy Cockburn, Philip Quinn, and Carl Gutwin. 2015. Examining the Peak-End Effects of Subjective Experience. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (CHI '15). Association for Computing Machinery, New York, NY, USA, 357–366. doi:10.1145/2702123.2702139
- [12] Fan Du, Nan Cao, Jian Zhao, and Yu-Ru Lin. 2015. Trajectory Bundling for Animated Transitions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (CHI '15). Association for Computing Machinery, New York, NY, USA, 289–298. doi:10.1145/2702123.2702476
- [13] Sandra M. Eldridge, Gillian A. Lancaster, Marion J. Campbell, Lehana Thabane, Sally Hopewell, Claire L. Coleman, and Christine M. Bond. 2016. Defining Feasibility and Pilot Studies in Preparation for Randomised Controlled Trials: Development of a Conceptual Framework. *PLOS ONE* 11, 3 (2016), e0150205. doi:10.1371/journal.pone.0150205
- [14] William W. Gaver. 2012. What Should We Expect from Research through Design?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 937–946. doi:10.1145/2207676.2208538
- [15] Stephen J. Gentles and Silvia L. Vilches. 2017. Calling for a Shared Understanding of Sampling Terminology in Qualitative Research: Proposed Clarifications Derived From Critical Analysis of a Methods Overview by McCrae and Purssell. *International Journal of Qualitative Methods* 16 (2017), 1–7. doi:10.1177/1609406917725678
- [16] Alix Goguy, Carl Gutwin, Zhe Chen, Pang Suwanaposee, and Andy Cockburn. 2021. Interaction Pace and User Preferences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, 14 pages. doi:10.1145/3411764.3445772
- [17] Lea Theresa Gröber, Matthias Fassl, Abhilash Gupta, and Katharina Krombholz. 2021. Investigating Car Drivers' Information Demand after Safety and Security Critical Incidents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21).

- Association for Computing Machinery, New York, NY, USA, 17 pages. doi:10.1145/3411764.3446862
- [18] Timothy C. Guetterman, Vicki L. Plano Clark, and José F. Molina-Azorin. 2024. Terminology and Mixed Methods Research: A Persistent Challenge. *Journal of Mixed Methods Research* 18, 1 (2024), 9–13. doi:10.1177/15586898231217855
 - [19] Steve Harrison, Phoebe Sengers, and Deborah Tatar. 2011. Making Epistemological Trouble: Third-Paradigm HCI as Successor Science. *Interacting with Computers* 23, 5 (2011), 385–392. doi:10.1016/j.intcom.2011.03.005
 - [20] Sabrina Hauser, Melinda J. Suto, Liisa Holsti, Manon Ranger, and Karon E. MacLean. 2020. Designing and Evaluating Calmer, a Device for Simulating Maternal Skin-to-Skin Holding for Premature Infants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 15 pages. doi:10.1145/3313831.3376539
 - [21] Jay Henderson, Jeff Avery, Laurent Grisoni, and Edward Lank. 2019. Leveraging Distal Vibrotactile Feedback for Target Acquisition. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland, UK) (CHI '19). Association for Computing Machinery, New York, NY, USA, 11 pages. doi:10.1145/3290605.3300715
 - [22] Leona Holloway, Kim Marriott, and Matthew Butler. 2018. Accessible Maps for the Blind: Comparing 3D Printed Models with Tactile Graphics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 13 pages. doi:10.1145/3173574.3173772
 - [23] Kasper Hornbæk, Aske Mottelson, Jarrod Knibbe, and Daniel Vogel. 2019. What Do We Mean by “Interaction”? An Analysis of 35 Years of CHI. *ACM Transactions on Computer-Human Interaction* 26, 4, Article 27 (2019), 30 pages. doi:10.1145/3325285
 - [24] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 12 pages. doi:10.1145/3313831.3376383
 - [25] Mahmood Jasim, Christopher Collins, Ali Sarvghad, and Narges Mahyar. 2022. Supporting Serendipitous Discovery and Balanced Analysis of Online Product Reviews with Interaction-Driven Metrics and Bias-Mitigating Suggestions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, 24 pages. doi:10.1145/3491102.3517649
 - [26] Mirabelle Jones, Christina Neumayer, and Irina Shklovski. 2023. Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, 24 pages. doi:10.1145/3544548.3580885
 - [27] Aisling Kelliher, Setor Zilevu, Thanassis Rikakis, Tamim Ahmed, Yen Truong, and Steven L. Wolf. 2020. Towards Standardized Processes for Physical Therapists to Quantify Patient Rehabilitation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 13 pages. doi:10.1145/3313831.3376706
 - [28] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E. Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. MyMove: Facilitating Older Adults to Collect In-Situ Activity Labels on a Smartwatch with Speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, 21 pages. doi:10.1145/3491102.3517457
 - [29] Rebecca Kleinberger, Jennifer Cunha, Megan McMahon, and Ilyena Hirskey-Douglas. 2024. No More Angry Birds: Investigating Touchscreen Ergonomics to Improve Tablet-Based Enrichment for Parrots. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, 16 pages. doi:10.1145/3613904.3642119
 - [30] Simone Kopeinik, Martina Mara, Linda Ratz, Klara Krieg, Markus Schedl, and Navid Rekabsaz. 2023. Show Me a “Male Nurse”: How Gender Bias is Reflected in the Query Formulation of Search Engine Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, 15 pages. doi:10.1145/3544548.3580863
 - [31] Kuno Kurzahls, Emine Cetinkaya, Yongtao Hu, Wenping Wang, and Daniel Weiskopf. 2017. Close to the Action: Eye-Tracking Evaluation of Speaker-Following Subtitles. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 6559–6568. doi:10.1145/3025453.3025772
 - [32] Leona Lassak, Philipp Markert, Maximilian Golla, Elizabeth Stobert, and Markus Dürmuth. 2024. A Comparative Long-Term Study of Fallback Authentication Schemes. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, 19 pages. doi:10.1145/3613904.3642889
 - [33] Rong-Hao Liang, Bin Yu, Mengru Xue, Jun Hu, and Loe M. G. Feijs. 2018. BioFidget: Biofeedback for Respiration Training Using an Augmented Fidget Spinner. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 12 pages. doi:10.1145/3173574.3174187
 - [34] Wanyu Liu, Olivier Rioul, Joanna McGrenere, Wendy E. Mackay, and Michel Beaudouin-Lafon. 2018. BIGFile: Bayesian Information Gain for Fast File Retrieval. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 13 pages. doi:10.1145/3173574.3173959
 - [35] Lian Loke, Aaron Blishen, Carl Gray, and Naseem Ahmadpour. 2021. Safety, Connection and Reflection: Designing with Therapists for Children with Serious Emotional Behaviour Issues. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan). Association for Computing Machinery, New York, NY, USA, 17 pages. doi:10.1145/3411764.3445178
 - [36] Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What Have We Learned from the Checklist?. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational

- Linguistics, Toronto, Canada, 12789–12811. doi:10.18653/v1/2023.findings-acl.809
- [37] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW, Article 72 (2019), 23 pages. doi:10.1145/3359174
 - [38] Arinobu Nijima, Masato Shindo, and Ryosuke Aoki. 2025. Invisible Light Touch: Standing Balance Improvement by Mid-Air Haptic Feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '25). Association for Computing Machinery, New York, NY, USA, 11 pages. doi:10.1145/3706598.3713396
 - [39] Jonas Oppenlaender, Tahir Abbas, and Ujwal Gadiraju. 2024. The State of Pilot Study Reporting in Crowdsourcing: A Reflection on Best Practices and Guidelines. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1, Article 184 (2024). doi:10.1145/3641023
 - [40] Jonas Oppenlaender, Sylvain Malacria, Xinrui Fang, Niels van Berkel, Fanny Chevalier, Koji Yatani, and Simo Hosio. 2025. Meta-HCI: First Workshop on Meta-Research in HCI. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 791, 8 pages. doi:10.1145/3706599.3706723
 - [41] Han Qiao, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2025. To Use or Not to Use: Impatience and Overreliance When Using Generative AI Productivity Support Tools. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '25). Association for Computing Machinery, New York, NY, USA, 18 pages. doi:10.1145/3706598.3714103
 - [42] Paul Ralph. 2021. ACM SIGSOFT Empirical Standards Released. *ACM SIGSOFT Software Engineering Notes* 46, 1 (2021), 19. doi:10.1145/3437479.3437483
 - [43] Raquel Breejon Robinson, Alberto Alvarez, and Elisa D. Mekler. 2024. How to Write a CHI Paper (Asking for a Friend). In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 558, 8 pages. doi:10.1145/3613905.3644051
 - [44] Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'Just What do You Think You're Doing, Dave?' A Checklist for Responsible Data Use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 4808–4820. doi:10.18653/v1/2021.findings-emnlp.414
 - [45] Katja Rogers, Teresa Hirzle, Sukran Karaosmanoglu, Paula Toledo Palomino, Ekaterina Durmanova, Seiji Isotani, and Lennart E. Nacke. 2024. An Umbrella Review of Reporting Quality in CHI Systematic Reviews: Guiding Questions and Best Practices for HCI. *ACM Transactions on Computer-Human Interaction* 31, 5, Article 57 (2024), 55 pages. doi:10.1145/3685266
 - [46] Mohammad Reza Saleh Sedghpour, Alessandro Vittorio Papadopoulos, Cristian Klein, and Johan Tordsson. 2024. Artifact Evaluation for Distributed Systems: Current Practices and Beyond. *arXiv preprint arXiv:2406.13045* (2024). <https://arxiv.org/abs/2406.13045>
 - [47] Korok Sengupta, Sabin Bhattarai, Sayan Sarcar, I. Scott MacKenzie, and Steffen Staab. 2020. Leveraging Error Correction in Voice-based Text Entry by Talk-and-Gaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 11 pages. doi:10.1145/3313831.3376579
 - [48] Aneesh Singh, Martin Dechant, Dilisha Patel, Ewan Soubutts, Giulia Barbareschi, Amid Ayobi, and Nicola Newhouse. 2025. Exploring Positionality in HCI: Perspectives, Trends, and Challenges. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25'). Association for Computing Machinery. doi:10.1145/3706598.3713280
 - [49] Artifex Software. 2026. PyMuPDF: Python library for PDF processing. <https://github.com/pymupdf/pymupdf>
 - [50] Susan Leigh Star and James R. Griesemer. 1989. Institutional Ecology, 'Translations' and Boundary Objects... *Social Studies of Science* 19, 3 (1989), 387–420. doi:10.1177/030631289019003001
 - [51] Evropi Stefanidi, Marit Bentvelzen, Paweł W. Woźniak, Thomas Kosch, Mikołaj P. Woźniak, Thomas Mildner, Stefan Schneegass, Heiko Müller, and Jasmin Niess. 2023. Literature Reviews in HCI: A Review of Reviews. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. doi:10.1145/3544548.3581332
 - [52] Ayaka Tsutsui, Tatsuki Fushimi, Takahito Murakami, Ryosei Kojima, Kengo Tanaka, and Yoichi Ochiai. 2024. HIFU Embossment of Acrylic Sheets. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, 23 pages. doi:10.1145/3613904.3642890
 - [53] Vinoba Vinayagamoorthy, Maxine Glancy, Christoph Ziegler, and Richard Schäffer. 2019. Personalising the TV Experience using Augmented Reality: An Exploratory Study on Delivering Synchronised Sign Language Interpretation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 12 pages. doi:10.1145/3290605.3300762
 - [54] Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. 2024. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839* (2024).
 - [55] Mirjam Wester, Cassia Valentini-Botinhao, and Gustav Eje Henter. 2015. Are We Using Enough Listeners? No! – An Empirically-Supported Critique of Interspeech 2014 TTS Evaluations. In *Interspeech 2015*. 3476–3480. doi:10.21437/Interspeech.2015-689
 - [56] Zack While, Tanja Blascheck, Yujie Gong, Petra Isenberg, and Ali Sarvghad. 2024. Glanceable Data Visualizations for Older Adults: Establishing Thresholds and Examining Disparities Between Age Groups. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA). Association for Computing Machinery, New York, NY, USA, 17 pages. doi:10.1145/3613904.3642776
 - [57] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *interactions* 23, 3 (2016), 38–44. doi:10.1145/2907069
 - [58] Qin Wu, Chenmei Yu, Yanjun Chen, Jiayu Yao, Xi Wu, Xiaolan Peng, and Teng Han. 2020. Squeeze the Ball: Designing an Interactive Playground towards Aiding Social Activities of Children with Low-Function Autism. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing*

- Systems* (Honolulu, HI, USA). Association for Computing Machinery, New York, NY, USA, Article 759, 14 pages. doi:10.1145/3313831.3376888
- [59] Chengyuan Xu, Kuo-Chin Lien, and Tobias Höllerer. 2023. Comparing Zealous and Restrained AI Recommendations in a Real-World Human-AI Collaboration Task. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, 15 pages. doi:10.1145/3544548.3581282
 - [60] Ke Xu, Shunan Guo, Nan Cao, David Gotz, Aiwen Xu, Huamin Qu, Zhenjie Yao, and Yixin Chen. 2018. ECGLens: Interactive Visual Exploration of Large Scale ECG Data for Arrhythmia Detection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 12 pages. doi:10.1145/3173574.3174237
 - [61] Fumeng Yang, Yuxin Ma, Lane Harrison, James Tompkin, and David H. Laidlaw. 2023. How Can Deep Neural Networks Aid Visualization Perception Research? Three Studies on Correlation Judgments in Scatterplots. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (*CHI '23*). Association for Computing Machinery, New York, NY, USA, 17 pages. doi:10.1145/3544548.3581111
 - [62] Chenyang Zhang, Tiansu Chen, Eric Shaffer, and Elahe Soltanaghai. 2024. FocusFlow: 3D Gaze-Depth Interaction in Virtual Reality Leveraging Active Visual Depth Manipulation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI '24*). Association for Computing Machinery, New York, NY, USA, 18 pages. doi:10.1145/3613904.3642589
 - [63] Mingrui Ray Zhang and Shumin Zhai. 2021. PhraseFlow: Designs and Empirical Studies of Phrase-Level Input. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI '21*). Association for Computing Machinery, New York, NY, USA, 13 pages. doi:10.1145/3411764.3445166
 - [64] Hui Zheng and Vivian Genaro Motti. 2018. Assisting Students with Intellectual and Developmental Disabilities in Inclusive Education with Smartwatches. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 12 pages. doi:10.1145/3173574.3173924
 - [65] Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhan Luo. 2025. Customizing Emotional Support: How Do Individuals Construct and Interact With LLM-Powered Chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (*CHI '25*). Association for Computing Machinery, New York, NY, USA, 20 pages. doi:10.1145/3706598.3713453
 - [66] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through Design as a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 493–502. doi:10.1145/1240624.1240704

A Inter-Annotation Reliability

Reliability sample construction. As mention in the main text, we first independently coded five papers as a calibration exercise and refined the codebook after reviewing our disagreements. Across both stages, we independently coded 19 eligible papers. We report the calibration and post-refinement stages separately because they used different codebook versions. We restricted the reported analysis to the six main coding decisions. We ignored the order of multilabel selections and considered two annotations an exact match only when their complete sets were identical. We pooled papers across publication years because the individual year strata were too small for stable year-specific estimates.

Table 2 provides the complete variable-level reliability results for the post-refinement assessment, including exact agreement and the coefficients applicable to each coding decision.

Table 2. Variable-level inter-annotator reliability in the post-refinement assessment.

Coding decision	Exact agreement	Mean Jaccard	Krippendorff's α
Motivation	64.3%	.774	.702
Recruitment described	78.6%	—	.585
Protocol described	92.9%	—	.640
Results discussed	100%	—	1.000
Impact	78.6%	.857	.733
Outside-section change	85.7%	—	-.038
Descriptive aggregate	83.3%	—	—

Jaccard-based statistics apply only to the multilabel variables. The aggregate exact-agreement score is descriptive and not chance-corrected. The negative coefficient for Outside-section change reflects a strongly imbalanced response distribution despite high observed agreement. Dashes indicate that a statistic was not applicable.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009