Crowdsourced Think-Aloud Studies

Zach Cutler
zcutler@sci.utah.edu
University of Utah
Salt Lake City, Utah, USA

Lane Harrison
ltharrison@wpi.edu
Worcester Polytechnic Institute
Worcester, Massachusetts, USA

Carolina Nobre
cnobre@cs.toronto.edu
University of Toronto
Toronto, Canada

Alexander Lex
alex@sci.utah.edu
University of Utah
Salt Lake City, Utah, USA

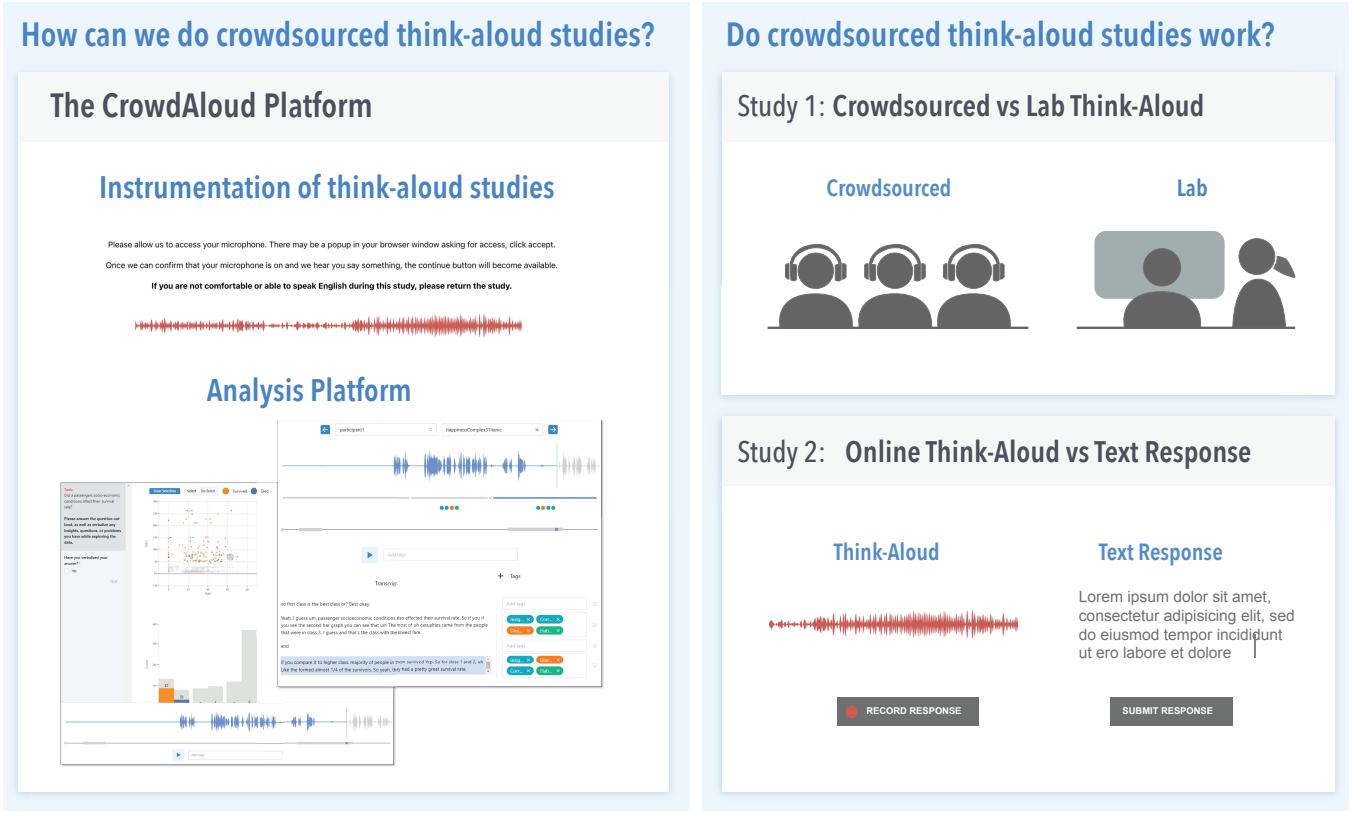

Figure 1: The two contributions of our work: the CrowdAloud platform that enables the collection of crowdsourced think-aloud data (left), and two studies that show that crowdsourced think-aloud studies work.

Abstract

The think-aloud (TA) protocol is a useful method for evaluating user interfaces, including data visualizations. However, TA studies are time-consuming to conduct and hence often have a small number of participants. Crowdsourcing TA studies would help alleviate these problems, but the technical overhead and the unknown quality of results have restricted TA to synchronous studies. To address this gap we introduce CrowdAloud, a system for creating and analyzing asynchronous, crowdsourced TA studies. CrowdAloud captures audio and provenance (log) data as participants interact with a stimulus. Participant audio is automatically transcribed and visualized together with events data and a full recreation of the state of the stimulus as seen by participants. To gauge the value of crowdsourced TA studies, we conducted two experiments: one to compare lab-based and crowdsourced TA studies, and one to compare crowdsourced TA studies with crowdsourced text prompts. Our results suggest that crowdsourcing is a viable approach for conducting TA studies at scale.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3714305

CCS Concepts

• **Human-centered computing → Empirical studies in visualization**; **Visualization design and evaluation methods**; **Empirical studies in HCI**; **HCI design and evaluation methods**.

Keywords

Think-Aloud Study, Talk-Aloud Study, Visualization

ACM Reference Format:

Zach Cutler, Lane Harrison, Carolina Nobre, and Alexander Lex. 2025. Crowdsourced Think-Aloud Studies. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3706598.3714305

1 Introduction

The think-aloud (TA) protocol—asking users to verbalize their thought process while using an interface—is a frequently used usability evaluation method both in industry and in the HCI and Visualization research communities [13, 68]. The visualization research community has adopted TA for usability studies not only to evaluate visualization tools or techniques but also to study how users of visualizations form insights.

Most TA studies are done *synchronously*, either in-person (e.g., as part of a lab study) or remotely via video conferencing. Synchronous user studies require participants and researchers to be present in real-time while the study takes place [16]. Synchronous user studies are thus time-intensive in the data collection phase.

In contrast to synchronous studies, *asynchronous* studies do not require an experimenter to be present when a participant completes a study. Asynchronous studies are often run as crowdsourced experiments, where users complete tasks online [45]. The asynchronous nature of crowdsourcing, combined with a potentially large and easy to recruit participant pool enables experimenters to scale to larger participant numbers at a lower cost than what is achievable in synchronous studies.

However, the think-aloud protocol is difficult to execute asynchronously. Only one of the 67 TA studies we analyzed (see Section 2) was done asynchronously. Setting up a system to record and save audio automatically is a technical challenge. Remote screen capture is an intrusive process requiring participants to grant browser permissions. In practice, **these difficulties restrict think-aloud studies to a synchronous setup, limiting economically achievable sample sizes and participant pools**. It is also worth noting that the overhead associated with recruiting synchronous participants limits the type of study that practically can be run with the TA protocol. For example, short, high participant count studies may be desirable in some cases but do not make economic or logistical sense synchronously.

Nevertheless, if these technical hurdles could be overcome, asynchronous studies could reduce the effort it takes to conduct TA studies and open up new use cases for crowdsourced studies.

TA studies, however, are also time-intensive in the analysis phase. After a study is complete, experimenters need to transcribe the audio or use an (automatic) transcription service then fix mistakes, and review the transcription in the context of the screen capture. Finally, for a more rigorous analysis of TA data, researchers frequently use qualitative coding methods, again manually reviewing the screen capture and audio while coding the transcript. This is a time-consuming process, which does not scale well.

To address these problems we make two primary contributions, illustrated in Figure 1:

First, we introduce **CrowdAloud, a method and system for running crowdsourced TA studies on the web** with minimal effort for study designers. CrowdAloud records, saves, and transcribes audio automatically. We utilize detailed application provenance data to simulate screen capture without having to record video. Additionally, CrowdAloud introduces an analysis interface, enabling study creators to conduct qualitative analysis utilizing the linked audio and provenance data. CrowdAloud thus overcomes key technical difficulties making it possible to run crowdsourced TA studies, and provides a streamlined integrated analysis process, which makes analyzing large-scale TA data feasible.

Our second contribution is **two validation studies for the utility and feasibility of asynchronous think-aloud studies**. The first study (a) identifies the differences between in-person synchronous and crowdsourced asynchronous TA studies; the second study (b) evaluates how an asynchronous TA study compares to the asynchronous elicitation of comments in a text field.

Based on our results, we introduce preliminary guidance on how to conduct crowdsourced TA studies.

We recognize the growing calls within the visualization community to embrace a plurality of epistemologies [2], and that large-scale studies are not desirable for all research. TA evaluation itself is rooted in positivist perceptions of usability testing but also can fit into other epistemologies, such as more interpretivist schools of thought [73]. CrowdAloud is designed to be suitable for research based on a wide range of epistemologies.

2 Think-Aloud Studies in Visualization

To characterize current practices in conducting TA studies in visualization, we surveyed research papers that describe TA studies.

The goals of our survey were to understand: (a) the ways in which TA studies are currently conducted (*how*), (b) what the goals of the studies are (*why*), (c) what kinds of participants are recruited for TA studies (*who*), and (d) the data collected in TA studies (*what*). Answers to these questions shape our approach to supporting asynchronous crowdsourced TA studies. With regards to *how*, driving questions included whether synchronous TA studies are conducted primarily in-person, or remotely? Additionally, are there examples of asynchronous studies that we can learn from? With regards to *why*, to what extent are visualization TA studies designed for usability testing, or to tackle broader insight formulation questions? The survey also examines *who* is participating in these studies, because asynchronous studies present opportunities for recruiting a diverse set of participants. Are current studies mostly run with expert users (e.g., domain scientists), or novices (such as students)? Finally, the survey collects information on *what* kind of data visualization-focused TA studies collect and how it is analyzed.

2.1 Methodology

To identify papers, we use the Vispubs database of visualization papers from VIS, Eurovis, and CHI [53], and restrict our search to papers published since 2014 at IEEE VIS. Searches of 'think-aloud'

	Participant #	Modality	Purpose	Participant Type	Transcribed	Provenance	Coded
Ren et al. [78]	1	IP	US	SP	Y	N	N
Chen et al. [22]	2	IP	US	EX	Y	N	N
Crnovrsanin et al. [25]	2	IP	US	EX	Y	N	N
Shi et al. [82]	2	IP	US	EX	Y	N	N
Sperrle et al. [84]	2	IP	US	EX	Y	N	N
Abdelaal et al. [1]	3	IP		EX	N	N	N
Nguyen et al. [67]	3	IP	IN SM	EX	Y	Y	N
Sun et al. [86]	3	VC	US	EX	Y	N	N
Park et al. [76]	3	IP	US	EX	Y	N	N
Hoque et al. [46]	3	IP	US	EX	Y	N	Y
Mota et al. [65]	4	IP	US	EX	Y	N	N
Xenopoulos et al. [98]	4	IP	US	EX	N	N	N
Lekschas et al. [56]	5	IP	US	EX	Y	N	N
Zgraggen et al. [103]	5	IP	US	EX	Y	N	N
Chouddhry et al. [24]	5	IP	US	EX	Y	N	N
Ono et al. [75]	6	IP		EX	Y	N	N
Horak et al. [47]	6	VC	US	EX	Y	N	N
Deng et al. [30]	6	IP	US	NV	Y	N	N
Cashman et al. [18]	6	VC	US IN	SP	Y	N	N
Jang et al. [50]	6	IP	US	EX SP	Y	N	N
Yalcin et al.[100]	6	IP	IN	NV	Y	N	N
Zhu-tian et al. [105]	7	IP	US	EX	Y	N	Y
Troidl et al. [89]	7	IP	US	EX	N	N	N
Lu et al. [61]	7	IP	US	NV	N	N	N
Langner et al. [54]	7	IP	US	NV	Y	N	N
Ceneda et al [21]	8	IP		EX	N	N	N
Wang et al. [93]	9	IP	IN	EX	Y	N	N
Guo et al. [43]	9	IP	SM IN	SP	Y	Y	N
Wang et al. [94]	10	IP	IN US	EX	Y	Y	N
Huang et al. [48]	10	IP	US	EX	Y	N	N
Wang et al. [92]	10	VC	US	EX	N	N	N
Han et al. [44]	10	IP	IN US	EX SP	Y	Y	Y
Guo et al [42]	10	IP	IN SM	NV	N	Y	N
Nowak et al [71]	10	AS	IN SM	EX	Y	Y	Y
Ghai et al. [41]	10	VC	US	NV	N	N	N
Wang et al. [95]	10	IP	IN US	SP	Y	N	N
Zhao et al. [104]	10	IP	US	SP	Y	N	N
Ren et al. [79]	11	IP	US	EX	Y	N	N
Kale et al. [51]	12	VC	IN	EX	Y	N	N
Cavallo et al. [20]	12	IP	US IN	EX	Y	N	Y
Srinivasan et al. [85]	12	IP	US	EX SP NV	Y	N	N
Wen et al [97]	12	IP	US	NV	Y	N	N
Tong et al. [88]	12	IP	US	NV	Y	N	N
Behrisch et al. [8]	12	IP		SP	N	N	N
Fujiwara et al. [36]	12	IP	US	SP	Y	N	N
Lin et al. [57]	12	IP	US	SP	N	N	N
Chen et al. [23]	12	IP	US	SP	Y	Y	N
Lee et al. [55]	13	IP	IN SM	NV	Y	N	Y
Badam et al. [5]	14	IP	IN	SP	Y	Y	Y
Dhanoa et al [31]	14	VC	IN	NV	Y	N	N
Epperson et al. [33]	16	IP	IN	EX	Y	Y	N
Batch et al. [6]	16	IP	US	EX	Y	Y	N
Gaba et al. [37]	16	VC	IN	EX NV	Y	N	N
Blascheck et al. [10]	16	IP	IN US	NV	Y	Y	N
Wang et al. [96]	16	VC	US	NV	Y	N	N
Liu et al. [60]	16	IP	IN US	SP	Y	Y	Y
Preston et al. [77]	17	IP	IN	NV	Y	N	Y
Liu et al. [59]	18	IP	US	EX	Y	N	N
Gao et al. [40]	18	IP	IN	SP	Y	N	N
Bertucci et al. [9]	20	IP	US	NV	Y	N	N
Lee et al. [55]	24	IP	IN	NV	Y	Y	N
Narechania et al. [66]	24	VC	US IN	SP	Y	Y	N
Blascheck et al. [11]	24	IP	IN	EX SP NV	Y	Y	N
Battle et al. [7]	30	IP	IN	NV	Y	N	N
Dasgupta et al. [29]	34	IP	US	EX	Y	N	N
Block et al. [12]	41	IP	US IN	NV	Y	Y	N
Ma et al. [62]	56	IP	US IN	NV	Y	N	Y

Table 1: Analysis of think-aloud modalities in visualization research papers. We see that most studies are in-person and only one is asynchronous; that most studies analyze usability, but a sizeable number also evaluate insights. Modality can be IP in-person (lab), VC video conference, or AS asynchronous; the purpose can either be US usability evaluation, IN analysis of insights, or SM sensemaking (empty cells indicate studies conducted for general feedback without a specific purpose); participants can be EX experts, SP skilled, or NV novices. For all other dimensions, we record binary Y yes and N no.

and various synonyms in titles and abstracts provided a limited number of results. We therefore develop the following screening method: First, collect any paper that has 'user study', 'think aloud', or 'talk aloud' in the abstract or title. Then, examine each paper manually to see if it actually uses a think-aloud protocol. We used a similar method with IEEE Xplore to find papers we may have missed, as well as added any visualization papers that we later discovered had described a TA study. With this process, we identified 67 visualization papers that described a TA study.

We seeded a code-book based on our research questions described above, and extended categories as we identified relevant topics during coding. Codes were discussed between all authors; the final coding was conducted by the first author. Using this codebook, we reviewed and coded 67 papers, listed in Table 1.

2.2 Findings

We below report on the insights of our analysis. For an overview, refer to Table 1.

How are TA studies conducted? Notably, 66 of 67 TA studies were performed synchronously (either in-person IP, or via video-conferencing VC); only one study was conducted asynchronously AS, indicating that there is an opportunity for our work. 56 took place in-person and 10 over video conferencing. One exception is a study by Nowak et al. [71] that allowed experts to take their study at any time over a month-long period and asked them to

self-capture audio and screen capture data. For analysis, interactions were hand-coded, and audio was used to create a codebook of *sensemaking processes*. Their findings and successful study indicate promising prospects for asynchronous TA but also have many differences from the type of asynchronous studies we conduct in this paper. Their study utilized experts and asked those experts to self-capture, a delicate process that does not readily transfer to a crowdsourcing environment.

What are the goals of TA studies? (Why?) We found several instances of papers that report on TA studies conducted for *usability testing* US or that *evaluated insights* IN, although frequently papers did both. A majority of papers report on TA studies used only for usability testing (36/67). In these cases, the TA studies were typically a small part of the paper used to evaluate a proposed novel tool or method. Most of the visualization studies we reviewed are for more complicated systems than what is often evaluated in classical usability testing, which would be focused primarily on user experience problems. Visualization researchers use usability testing for a wide range of use cases, including to validate the design of visualizations and to validate the general utility of methods and tools. For the purposes of this paper, however, any of these investigations that do not investigate how participants develop insights are classified as usability testing. Some papers in our dataset used TA only for high-level feedback, in which case we do not explicitly classify a purpose.

Twenty seven papers explicitly evaluated insights, though the employed evaluation methodology varied widely. Some papers [e.g., 12, 60, 66, 100, 102] conduct a qualitative analysis on the TA data, segmenting audio portions and creating and applying codebooks. Others [11, 42, 71] perform a similar coding process but on interactions instead of audio data. Yet another group [e.g., 42, 43, 55, 67, 71] explicitly attempted to evaluate *sensemaking* SM, a similar challenge as insights. Many papers had multiple goals, often primarily focused on usability testing but also reporting insights without a complete qualitative analysis. As a whole, we find that the visualization community is using TA studies for diverse goals.

Who is participating in TA studies? To assess the participant pool for our survey, we coded participants into three categories. *Experts* EX are professionals with years of experience or academics in a particular research area. *Skilled Participants* SP are those that have specialized expertise, such as master's or undergraduate students who have taken a specific course giving them the knowledge they need for the study. Finally, *Novices* NV are participants who do not require any prior knowledge in a field to participate in the study. There is some overlap, as a few studies had participants from multiple categories.

The participant pool in the TA studies we surveyed leaned heavily towards users with some expertise, with 36 of the 67 papers recruiting experts and another 17 employing skilled participants. Twenty one studies used novices, who were almost entirely students, with the exception of Ma et al. [62], who recruited from a local museum, and Yalçın et al. [100] who recruited from public message boards. Not surprisingly, studies with larger participant numbers predominantly recruited novices or skilled participants; not experts. Crowdsourcing, which this paper focuses on, most commonly uses novice users. However, platforms such as Prolific enable targeted recruitment of skilled participants, and approaches for recruiting experts also exist [80]. Also, asynchronous studies like that of Nowak et al [71] may be distributed to experts, indicating broad applicability of asynchronous TA studies.

Participant count varied quite a bit in surveyed papers but tended to be low, with a median of 10 and a mean of 11.8. Participant count was especially low in studies that utilized experts, with a median of 7 and an average of 8.8. We hope that the ease of asynchronous studies, as well as the wider possible participant pools, can help raise participant counts for TA studies.

What kind of data do TA studies collect? All studies recorded audio, and most also recorded participants' screens. Rarely, studies also recorded video for use cases such as eye tracking, e.g., [11]. Most studies (57/67) transcribed the audio, but only a few took the effort to fully code audio segments (10/67); the others used the transcripts for quotes. Interestingly, of the 10 papers that coded audio segments, 6 utilized novices, while only 20 used novices in total. This indicates that low-participant, expert studies tend to be used to evaluate systems, whereas higher number studies, with novices and rigorous analysis approaches tend to investigate broader questions, such as insight formation.

As far as logging interactions is concerned, only a minority of studies (15/67) use any provenance tracking at all. Of the studies that did, most were simply logging system-level interactions (mouse and keyboard events) as opposed to *application events*, which is provenance data specific to actions in an application (brushes, selections, zooms, etc.) [60]. There were a few notable exceptions that did use application events and included them in their analysis in interesting ways. Liu et al. recorded both mouse events and application events and reported how the provenance differed between study conditions. Batch et al. [6] utilized provenance tracking to fully reproduce participant sessions after the fact, allowing them to better analyze the TA data. In our work, we similarly utilize application events to reproduce participant sessions and report on participant interaction count.

3 Related Work

In addition to the analysis of how think-aloud studies are conducted in the visualization research community, we also review the literature on think-aloud methodology more broadly, on crowdsourcing in visualization studies, and on analysis tools for user study data.

3.1 Think-Aloud Study Methodology

Studies that use the think-aloud protocol, first introduced by Ericsson and Simon [34], ask participants to verbalize their thoughts out loud while completing a study. TA is considered to be one of the most effective usability evaluation methods [13, 68], and TA testing is widely used in industry [68]. In a survey, 77% of researchers who conduct user studies indicated using TA in their studies at least 'often' [64]. Despite the clear benefits, conducting a TA study is a large time and effort investment, and if conducted in-person can limit participant pools. There has been research on synchronous remote TA studies, conducted over video conferencing software, which shows similar effectiveness to in-person [87] but only mitigates some of the overhead associated with in-lab TA studies.

Logging in Think-Aloud Studies. Many think-aloud studies incorporate some sort of provenance tracking in their study. Fifteen of the 62 papers we investigated include some sort of provenance. A few papers, such as the work by Liu et al. [60] and a few others [12, 66], do in-depth analysis of both think-aloud and provenance data. In these studies, however, the two are considered separate, with provenance primarily being a quantitative measure to be reported and think-aloud a qualitative one to be coded. We believe the separation of the two is to the detriment of both, as what a participant is *doing* and *saying* is closely related, and jointly analyzing both may improve understanding of the process.

Role of the Researcher in TA Studies. In *traditional TA*, as described by Ericsson and Simon [34], researchers are encouraged not to intervene much, as questions would likely result in verbalizations that require more cognitive processing than typically required to complete the task and may affect results. Instead, the researcher's role is to simply remind the participant to continue talking. However, more recently, some researchers have advocated for *interactive* TA, which encourages questions and interaction from the researcher and seeks out high-level verbalizations [73]. Research has found that interactive TA results in more mouse clicks, higher completion rates, and longer task times [4, 74].

By analogy, interactive TA is similar to semistructured interviews, where the role and presence of the researcher are critical for following up on interesting points and asking targeted questions. Traditional TA (and by extension, asynchronous TA, which we study in this paper), is more similar to surveys. Hence, we consider asynchronous approaches to be a methodological mismatch with interactive TA.

3.2 Crowdsourcing Visualization Studies

Crowdsourced user studies have been shown to have a variety of benefits, including reduced costs and wider participant pools [45]. Crowdsourced studies frequently are used for perceptual tests but are also used for evaluating visualization techniques [70]. Studies that evaluate techniques (e.g., [58, 70]) analyze their crowdsourced data in a variety of ways. Frequently, crowdsourced studies rely on quantitative measures, such as Likert scales, to garner feedback. Some studies utilize event data to investigate the insight development process [49, 70], whereas others conduct qualitative analysis on text responses [27, 99, 101].

3.2.1 Asynchronous vs. Synchronous. Many works have evaluated the differences in responses between synchronous and asynchronous studies. Some of these studies have focused on quantitative results, showing similar results between study types [19, 26]. Gadiraju et al. [38] explicitly discuss the pros and cons of crowdsourcing and laboratory experiments and specifically highlight the issue of increased need for quality control in crowdsourced experiments, a widely known issue with crowdsourcing data [14].

The work probably most closely related to ours directly compares asynchronous and synchronous TA studies in a usability context (i.e., identifying defects) [39]. They use screen capture and audio recordings to record crowdsourced participants and analyze their results using traditional review methods. Compared to a synchronous remote (video conference) arm of the study, they find promising results, showing little difference in usability feedback between crowdsourced and video-conference participants. However, their study was focused entirely on usability, in contrast to our interest in insight evaluation, and they call for more work to investigate crowdsourced TA studies in broader contexts. Their concept of usability testing, with the main purpose being usability problems identification, is slightly different from how visualization researchers utilize usability testing, as discussed in section 2.

3.3 Analysis Tools for Think Aloud and Other Study Data

There are many commercial tools with the purpose of conducting qualitative analysis. One of the largest commercial tools is UserTesting.com [90], which is a commercial crowdsourcing platform aimed specifically at usability testing. UserTesting has many features for analyzing usability, such as tagging, note-taking and session playback, and enables the creation and running of asynchronous TA studies. However, UserTesting uses screen capture instead of provenance for replay and thus lacks application provenance data, which limits potential analysis features, such as connecting specific application events to an analysis replay. UserTesting is also a tool designed specifically for usability testing.

On the research side, a few tools have been designed specifically for analyzing TA data. CoUX [83] is a tool for the collaborative analysis of usability tests. CoUX lets analysts review screen captures or videos of TA studies and define and tag problems that arise during the analysis. VisTA is another tool for analyzing TA usability tests, utilizing machine learning capabilities to help analyze TA data [35]. Both of these tools are focused on usability testing, and make little or no use of application provenance data. They are also for analyzing TA data, not conducting TA user studies.

VA2 [10] is a tool for evaluating TA studies that, similar to CrowdAloud, combines TA data with application provenance data. VA2 is designed for analyzing studies focused on insights, and has a host of features aimed at investigating the insight development process. The work by Nobre et al. [69] introduces a tool for analyzing user studies based on application provenance data but does not include speech data. Unlike CrowdAloud, neither of these tools allows users to conduct full qualitative analyses in which analysts develop codebooks and tag text segments. Also, both are analysis tools and are not for conducting TA user studies.

Sensepath [67] is a tool for understanding the sensemaking process, similar to our goals of understanding insight development, and has an analysis and capture process very similar to CrowdAloud. Sensepath utilizes a browser extension to capture provenance data during a qualitative analysis, and combines that provenance data with screen capture and think-aloud data in later analysis. However, Sensepath is not designed for asynchronous capture; the screen capture and audio recordings must be conducted separately from their tool. Sensepath also does not support qualitative coding on audio sessions, as it does not transcribe audio segments or let analysts tag audio segments.

Overall, numerous tools are aimed at conducting qualitative analyses, and many of them have a focus on TA studies. However, these tools either enable qualitative analyses but specifically for

usability testing, or focus on insights but do not allow for qualitative analysis of audio segments.

4 The CrowdAloud Platform

In this section, we introduce our technical solution—CrowdAloud—to lessen the burden of running and analyzing asynchronous think-aloud studies before we evaluate the feasibility of such studies in the next sections. CrowdAloud serves two purposes. First, it enables the easy **instrumentation of think-aloud studies**, handling audio recording and storage in asynchronous settings. Second, it provides an **analysis platform**, making audio and provenance data easily accessible to support qualitative analysis. CrowdAloud is built on top of reVISit, a visualization-specific user study platform [32]. We discuss the relationship between reVISit and CrowdAloud in detail in Section 4.3.

4.1 Instrumenting an Asynchronous Think-Aloud Study

To make a CrowdAloud study, the first step is to add a single line of code into the reVISit configuration:

```
| recordStudyAudio: true
```

Adding this flag activates the CrowdAloud recording features; no further actions are necessary.

ReVISit studies are composed of a series of pages, such as a consent page, a survey page, and trial pages. Audio, responses, and interaction logs are stored for each of these pages. Experiments can define which pages record and which do not record audio via the configuration; for example, audio recording may be turned off for consent or survey components. ReVISit utilizes Google Firebase for storage, and CrowdAloud stores audio in the same database as other study data. CrowdAloud uses Firebase extensions to transcribe audio automatically.

To ensure participants are aware that they are being recorded, CrowdAloud alerts participants at the beginning and during the study to the fact that their microphone is active. When a participant visits a CrowdAloud study page, they are asked to provide access to their microphone via the browser dialog, which is also used, for example, by video conferencing software. During the study, a waveform visualization is visible in the header, indicating to participants that audio is being recorded, as shown in Figure 2 at the top. This waveform also serves as a subtle intervention that may help participants remember to speak.

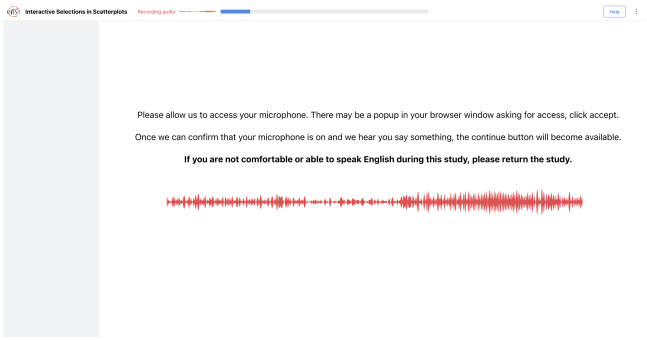

Figure 2: The audio-check interface for a CrowdAloud study. Participants cannot advance past this task until the browser picks up a minimum decibel level from their microphone. Also, notice the waveform visualization and the "Recording Audio" text in the header, which is visible on all pages that record audio. [↗ Mic Check Interface]

Optionally, experimenters can add a mic-check task (shown in Figure 2), which validates that participants have a functioning microphone and that the browser detects audible input. The mic-check is designed to avoid scenarios where people without functioning microphones, or without enabled browser permissions, accidentally take the study without having their speech captured.

Provenance Tracking. The TA protocol typically also involves screen recording. In the current version of CrowdAloud, we utilize detailed interaction provenance and provide video-level replay based on interaction logs instead of actual screen recording. When analyzing interactive stimuli (such as interactive visualizations), it is usually desirable to track *application events* (see Section 2), or full analysis provenance. ReVISit provides tools and documentation that instruct experimenters on how to instrument their studies with provenance. Designers of experiments can choose at what level of fidelity they want to record events. For example, recording all intermediate events during a brushing operation results in relatively large log files (yet still much smaller than video) but enables video-like reproduction. We discuss our approach to replay in the next section.

4.2 Analyzing a Study

Once data collection is complete, the next step in an experiment is to analyze the data. In this section, we describe the CrowdAloud analysis interface that leverages all the data modalities collected and provides an integrated analysis interface. However, it is important to note that the data can also be exported and analyzed with other tools at any time.

A unique characteristic of asynchronous TA studies is that the experimenter is not present during the experiment. As they did not see what the participant did, reviewing what was spoken and what was done is even more important than in synchronous settings.

Traditionally, TA studies enable reviews via screen and/or video recording. Yet although screen and video recording are acceptable in a lab setting, they can be intrusive in a crowdsourced context, as participants use their personal devices and are in private spaces. A second issue with screen or video recordings is that analyzing them takes a lot of time. Participants often do not interact with the system, e.g., while they step away from the experiment or while they are reading instructions. Hence, having data on when "something happens" has the potential to speed up analysis tasks. Nevertheless, lining up audio and video data and even automatically analyzing audio cues that can indicate usability problems is a viable option [35, 83].

Instead of video, CrowdAloud utilizes **application provenance data**, as described in the previous section, to create a replay of the interface. Each provenance node contains the entire state of the application required to reproduce the stimulus. Quickly stepping through states results in a replay very similar to a video, as shown in Figure 3. In contrast to previous work, which similarly utilizes

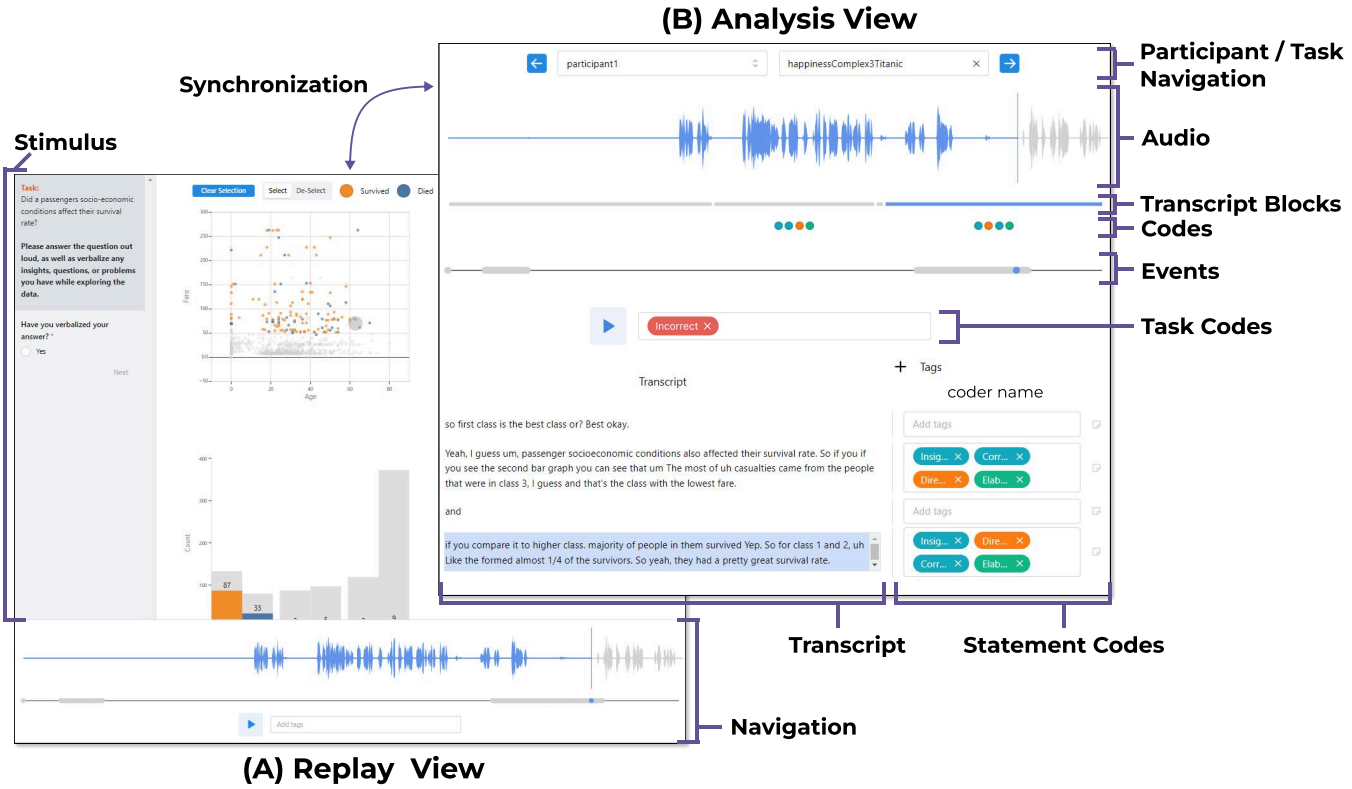

Figure 3: The Replay (A) and Analysis (B) Views that are used to review, transcribe, and code individual trials. (A) The Replay View [↗ Replay View] shows the state of the application that the participant saw at a selected time. Alternatively, it can show a video-like replay of the interaction and the audio. The integrated navigation bar enables experimenters to quickly jump to time points where events were logged or where the participant spoke. The Replay View is fully synced with the Analysis View. (B) The Analysis View [↗ Analysis View] shows the transcript and enables experimenters to edit it when necessary. Experimenters can assign statement codes to each block of text. The Codes, Events, and Blocks of Transcripts are synchronized with the audio and the replay in the Replay View. Experimenters can also assign Task Codes that apply to the whole task (as opposed to only a statement).

provenance for replay [69], the CrowdAloud replay is also linked with audio and transcript data.

Many visualization studies, however, also conduct detailed qualitative analyses of transcripts and the replay (see Section 2). To support this, CrowdAloud provides an **Analysis View**, which has the tools required to conduct a qualitative analysis while listening back to the audio and viewing transcripts.

The Replay and Analysis views, shown in Figure 3 are intended to be used simultaneously in separate browser windows. The views communicate with each other to stay in sync so that when, for example, a sentence in the transcript is selected, the replay view shows the associated state. By bringing together these typically separate modalities (audio/video, transcripts, events), CrowdAloud aims to reduce the analysis burden associated with TA studies.

4.2.1 Replay View. The replay view (see Figure 3) shows the application stimulus and state at a time point for a selected participant and the task as it appeared to the participant. Below the stimulus is a navigation bar with a waveform visualization and a visualization of events, so that analysts can see when users interacted and/or spoke. Analysts can select any time point, play back the audio, and view the interactions.

4.2.2 Analysis View. The analysis view enables experimenters to correct transcriptions, break (or merge) transcripts into semantically coherent blocks of text, code and annotate text blocks, and develop code-books—consistent with other qualitative analysis tools [63, 90]. To be consistent with our goal of creating a tool that serves the wide variety of TA studies and analyses that are conducted, the Analysis View also supports multiple coders, and distinguishes between *participant codes* (codes applying to a participant), *task codes* (codes applying to a whole task, for example "incorrect" to indicate that a task was incorrectly answered), and *statement codes* (codes applying to a specific transcribed statement).

What distinguishes the CrowdAloud analysis view from other qualitative analysis tools is the integration of all modalities: audio/transcripts, provenance events, codes, and replay view. Above the transcription portion is a timeline view that contains a waveform visualization showing the audio signal. Below the audio visualization, blocks of transcripts are encoded with a line. An active

Figure 4: The Participants View shows basic attributes for all participants who completed a task. The attributes shown are the completion time, the number of interactions, and the number of words spoken by each participant. It also shows the task and statement codes for each participant. The table can be ranked and filtered. In this example, it is filtered to only show participants that were coded with one of three statement codes; and ranked by word count. [↗ Analysis Interface]

segment is highlighted in both the timeline view and the transcription. Below, provenance events and codes are placed on the timeline at the position at which they occurred. The event that is shown in the replay view is highlighted.

4.2.3 Participant View. The Replay and Analysis views enable researchers to analyze individual participant sessions but do not provide an overview of all participants for a task. We provide such a higher level perspective in the Participant View, shown in Figure 4. The Participant View shows data on each participant and task, including time spent, interaction count, and word count. This data can be useful to get a sense of where participants spent their time in a study, and where they exerted effort via interaction/talking. The view also shows the codes assigned to the task for the participants. The participant view also supports filtering by tag, for example, in order to show only participants with a certain code for the task, so that experimenters can quickly get a sense of how frequently a tag was used or review any trials associated with that tag. Finally, the table view can be used to assign both participant tags and task tags.

By using these views, experimenters can efficiently read, code, and review the interactions on individual tasks with CrowdAloud and then export codes and quotes for reporting or further analysis.

4.3 Relationship to reVISit

The goal of this work is to support researchers in conducting asynchronous TA studies and decrease the effort associated with conducting such studies. We intend to support a wide range of studies that the visualization research community currently runs—from usability studies to insight-elicitation studies; from simple studies with novices on crowdsourcing platforms to asynchronous evaluation of bespoke tools by experts. With this in mind, we make use of an existing visualization study platform, reVISit [32], following recent calls for *care-ful* tool development that encourage leveraging "existing technological and social infrastructure that are already supported by a community" [3]. reVISit already has extensive features (and the associated documentation) required for conducting a user study, such as the ability to include a diverse set of stimuli, design sophisticated study sequences, and track provenance data [28].

However, there are many features that reVISit does not yet have that we contribute with CrowdAloud, both to reVISit's study instrumentation and to reVISit's built-in analysis platform. ReVISit does not currently support audio recording while conducting a study, (automatic) transcription of the audio, and storing the transcripts with the rest of the study data. The bulk of CrowdAloud's contributions to reVISit, however, are analysis features. Most prominent among them are the Analysis view, which enables audio playback, editing

of transcripts, developing codebooks, and coding transcripts; and the Replay View, which can be used to review interactions. Replay capability was demonstrated in a previous paper [69], but was not implemented in reVISit.

CrowdAloud is now fully integrated with reVISit. The code is available at https://github.com/reVISit-studies/.

5 Study 1: Asynchronous vs Synchronous TA

In this second part of our work we shift our focus from the technical feasibility of asynchronous (crowdsourced) think-aloud studies to the larger questions of utility: Do asynchronous think-aloud studies work? What are the differences between synchronous and asynchronous studies? For which kinds of questions are they appropriate? How do they compare to alternative methods commonly used in crowdsourced studies, specifically text responses?

To investigate the feasibility of crowdsourced TA studies, we conducted a user study (Study 1) to compare results between an in-person lab study (shown in blue) and a crowdsourced think-aloud study (shown in red). We show quotes from lab participants "*[LAB-TA] like this*" and quotes from crowdsourced TA participants "*[CS-TA] like so*". Study 1 focused on the following research questions:

- **RQ1**: Do crowdsourced participants talk as much as in-lab participants?
- **RQ2**: What differences, if any, are there between crowdsourced and in-lab responses for investigating usability?
- **RQ3**: Are crowdsourced responses useful for studies intended to investigate insights?
- **RQ4:** Are there logistical challenges in recruiting and recording crowdsourced participants?
- **RQ5:** What is the difference in terms of effort for an experimenter?
- **RQ6:** Do crowdsourced participants find TA studies enjoyable or uncomfortable?

Our foremost concern about crowdsourced studies was that participants would not talk sufficiently during the study (RQ1). In contrast to synchronous TA studies, where an experimenter is present, crowdsourced participants have no such external pressure. We also aimed to investigate how the differences in study environment might affect speech. For example, even if crowdsourced participants do speak frequently, do they stay on topic, or do they lose focus without an experimenter present? Crowdsourced participants may also find themselves in a physical location where they cannot freely speak and might find it awkward to talk into a "void". In the design phase of this project, we considered complex interventions to remind participants to speak [39], e.g., by audibly prompting them if no signal is detected, or by showing animated messages. However, we first wanted to test the baseline case in order to determine if such interventions were necessary, as we do here. In both of our studies, the only intervention is the aforementioned small visualization in the experiment interface title bar (see subsection 4.1). Given the high rate of speech in the baseline condition, we did not further pursue interventions to increase speech.

Research questions 2 and 3 are about the suitability of TA studies for different kinds of questions experimenters may want to answer. As discussed in section 2, both usability evaluations and understanding insight formation are frequent subjects of visualization TA studies. In our study, we evaluate whether asynchronous TA is suitable for both types of tasks and what the potential differences are. Although there is preliminary evidence for the former being successful [39], using crowdsourced TA studies to investigate insight formation and chart understanding has not been studied.

Questions 4 and 5 are about logistics: is it hard to find crowdsourcing participants if they have to enable a microphone? And is the increased instrumentation effort of a crowdsourced study worth it? Question 6 then asks about participants' experience: are crowdsourced TA studies a burden or a welcome departure from other monotonous tasks?

5.1 Procedure

Our study was composed of two components: an in-lab, synchronous study and a crowdsourced, asynchronous study. Both studies were reviewed by the local IRB and were deemed exempt from full review. We obtained consent from participants in both studies.

Before conducting the study, we conducted two pilots. The first was an in-person pilot with two students from our lab. The pilots were conducted in the same room we used for the in-lab portion of the study. We did not make any changes after the pilot. We also conducted a crowdsourced pilot with five participants using the crowdsourcing platform Prolific to ensure CrowdAloud was capturing provenance and audio data properly. Our return rate for this pilot was very high, and we refined the information we provided, e.g., how to give browser permissions to access the microphone. We do not include data from either pilot in our study.

For the in-person synchronous study, we conducted a lab study with 11 students from the computer science department at our research institution. This group completed the study in a dedicated user study room with a provided computer and an experimenter present. As discussed in section 2, this kind of participant pool (16/18 studies with novice participants used students) and sample size (a median of 10 participants) was the most common way visualization TA studies recruited novice participants. In-lab participants were rewarded with $20 Amazon gift cards.

For the asynchronous study, we conducted a crowdsourced study with 41 participants on the crowdsourcing platform Prolific. We recruited more participants in the online study because we could do so with less effort and at a lower cost. All crowdsourced participants indicated they were proficient in English, and we recruited only from individuals residing in the USA or UK. We limited participants to countries where English is the native language (US, UK) because we wanted to minimize issues with accents that may cause automatic transcriptions to fail. Online participants were paid $5 for an average of $15.68 per hour.

There was an intentional, environment-specific difference in the experiment between in-person and crowdsourced studies. All participants saw the same interface and instructions, but in-person participants were also walked through the introduction pages by a researcher, as is common in such studies. In-person participants were also encouraged to ask questions at any time. Crowdsourced participants could not rely on individualized instructions and had to rely on the information provided in the study. The full study can be seen at https://vdl.sci.utah.edu/ThinkAloud/ThinkAloud.

Figure 5: Stimulus from Question 2 of the usability tasks. Participants were asked to brush on a scatterplot and then find answers using the information in the linked bar chart and respond using the Study UI elements. [↗ Stimulus]

All of the code required to reproduce the study can be found at https://github.com/visdesignlab/ThinkAloud.

5.2 Study Stimulus

In order to address both RQ2 (usability) and RQ3 (insights), Study 1 was designed to have two different phases, one targeting *Usability Problems* and the other targeting the *Elicitation of Insight*, each consisting of four questions. The tasks ask participants to answer a data analysis question using an interactive scatterplot, a brush, and a linked bar chart, as shown in Figure 5. After the tasks were completed, participants were asked to complete a poststudy feedback form, for which the audio recording was turned off. All questions, except for the feedback form, encouraged participants to think-aloud throughout the study.

Usability Problems. The first four questions were intended to gauge how frequently participants discuss usability problems. The think-aloud prompt was: *"While answering this question, please verbalize your thoughts, especially about any insights you have or problems you run into."* For example, one question reads *"Consider only movies that had a production budget greater than 100 million and made less than 150 Million in Gross income. Which genres had the most and the least movies in this group of movies?"*, as seen in Figure 5. Participants were expected to brush in the designated range and then look at the filtered bar chart to answer the question. These tasks were intended to be straightforward, not requiring a lot of thought, but requiring interaction with an unfamiliar system to induce primarily UI/UX feedback.

Insight Problems. The second set of questions was designed to be more open-ended, using a complex set of visualizations, including potentially unfamiliar visualizations like violin plots, with the purpose of eliciting insights that are not immediately obvious in the data. All charts were interactive and linked. An example can be seen in Figure 6. The think-aloud prompt for this set of questions was *"Please answer the question out loud, as well as verbalize any insights, questions, or problems you have while exploring the data."* The insight questions used the Titanic dataset and asked questions about patterns and relationships in the data, such as *"Did women and children have a better survival rate than men?"*. The final question was entirely open-ended, simply asking *"Explore the dataset for anything interesting to you. What insights do you gain from the data?"*. Participants were expected to answer by speaking. To proceed to the next question, participants had to check a radio box asking *"Have you verbalized your answer?"*.

5.3 Codebook Development and Coding

To evaluate **RQ2** and **RQ3**, we conducted a qualitative analysis of the TA data. Transcripts were edited to fix mistakes in the transcription process, segmented, and coded.

When developing a codebook for the insight questions that we would use to address **RQ3**, we investigated how previous studies have approached the problem. We considered a few methods, such as ranking insights on a Likert scale [81], or less subjective user-activity metrics such as insights per minute [102]. Some of the tasks in our study were quite open-ended, yet the data was not complex enough to justify a 1-5 scale. Insights per minute or total insights as a measurement is helpful at times and is used in our analysis, but we wanted to evaluate both the quantity and quality of insights.

Initially, we seeded our codebook with very different insight characteristics, which were specific to our tasks and study. After beginning to code, we recognized that our goal in analyzing the TA data was not learning what insights participants gained from our system but that we instead should focus on how useful the insights would be to researchers who would conduct similar studies.

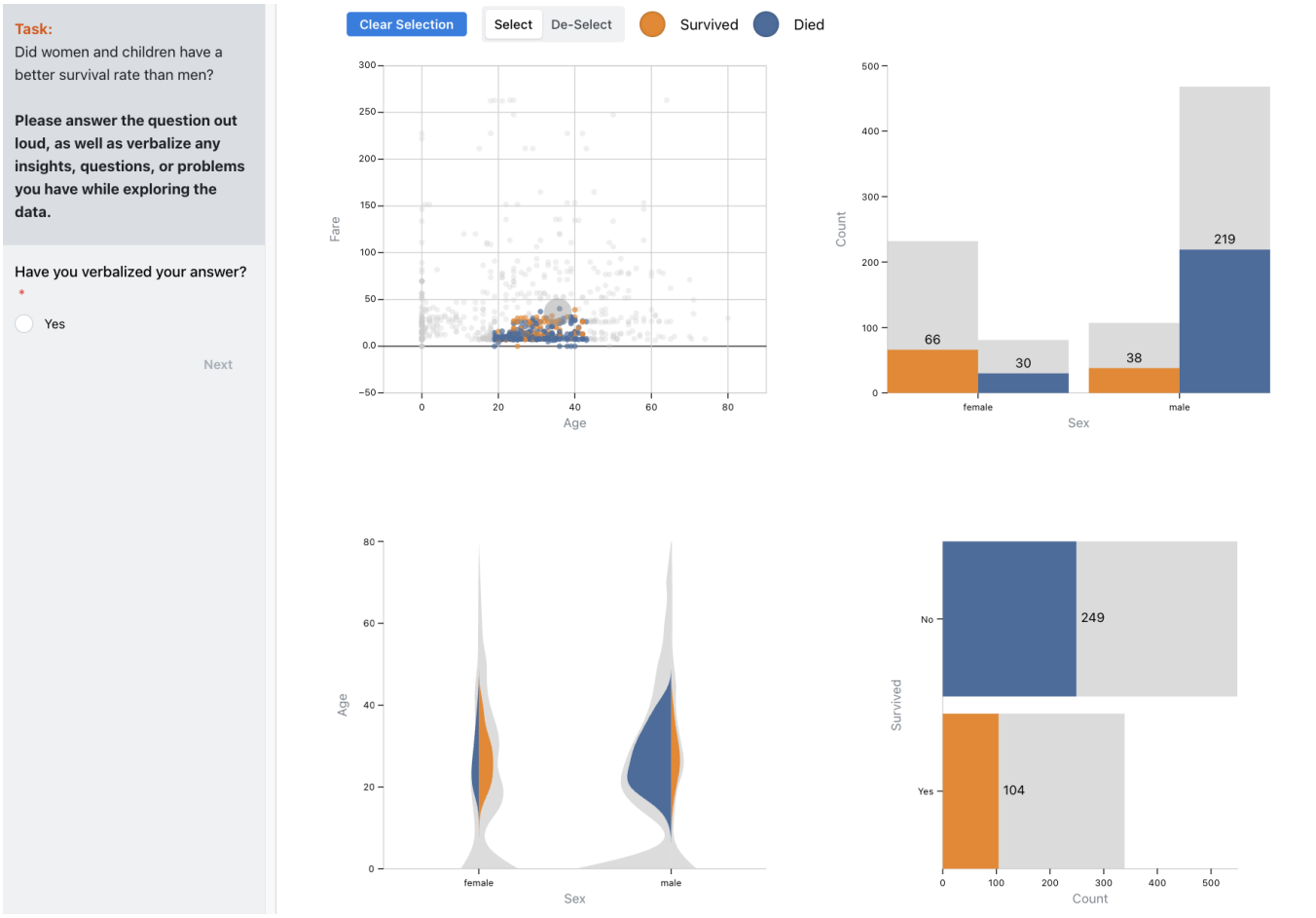

Figure 6: Stimulus from Question 2 of the insight tasks. Participants could brush on the scatterplot or click on any bar/violin to select across all plots. The task is open-ended, there is no answer to provide. [↗ Stimulus]

Hence, we iterated on our codebook multiple times by coding and discussing with all co-authors. The codebook was reviewed and revised based on feedback from all authors. The final coding was conducted solely by the first author. Any ambiguous or unclear cases found by the first author were discussed among all authors until consensus was reached.

Taking inspiration from Saraiya et al.'s work [81], we decided to give each insight the following **Insight Characteristics**

- `Correct vs. Incorrect`: Whether or not the stated insight was correct.
- `Directed vs. Undirected`: Most tasks asked specific questions and almost always produced directed insights related to the asked question. However, participants frequently expressed undirected insights that they found and explored on their own. For our final open-ended question, we marked insights that were explored in previous questions as directed.
- `Hypothesis`: Participants occasionally came up with hypotheses for *why* they saw certain trends in the data. A hypothesis is frequently an indication that an insight is *knowledge building*, i.e., that it connects and extends a user's existing knowledge [7].
- `Elaborated`: The goal of the elaborated tag is to capture whether or not participants explained their thought processes while coming to their insights.

To address **RQ2**, we wanted to tag all text segments that dealt with usability, but also recognized more nuance was required to properly answer the question. First, we wanted to know the sentiment of a usability comment to investigate whether negative comments were more likely from crowdsourced participants who were not in a room with an experimenter who they might perceive they were critiquing. However, not all usability comments have a sentiment. For example, some of the most helpful usability comments are actually questions that indicate that an interface is confusing to the participant. With this in mind, we coded usability comments with three separate codes, `UI/UX Comments` that came with a `positive` or `negative` tag, `UI/UX Questions`, and `UI/UX Suggestions`, for situations where participants made concrete suggestions on how to improve the interface.

5.4 Results

Of the 11 participants who took the in-lab study, we gathered complete transcripts from all but one. Due to internet problems during

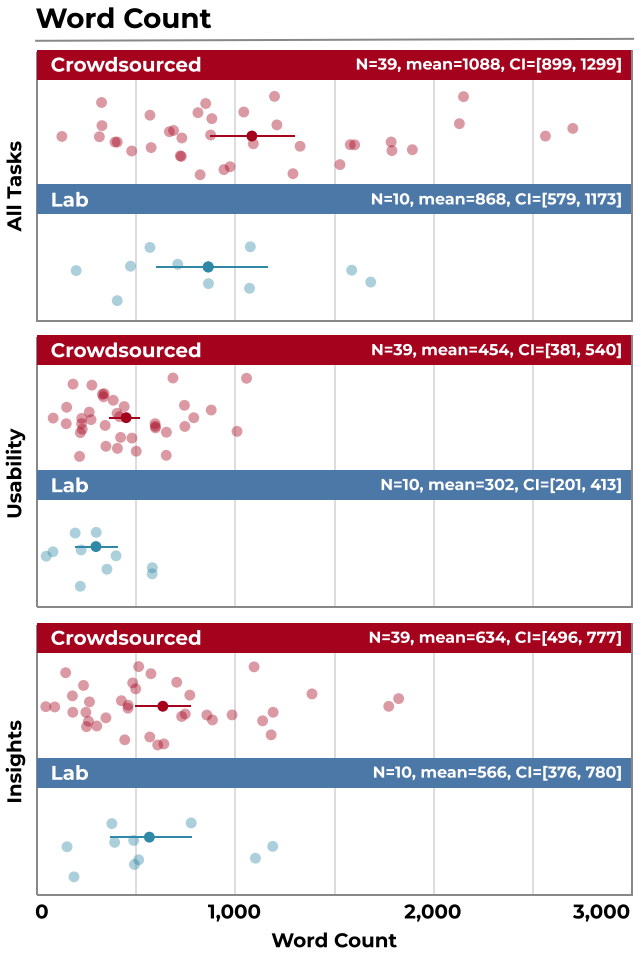

Figure 7: Asynchronous, crowdsourced participants spoke slightly more across all task categories.

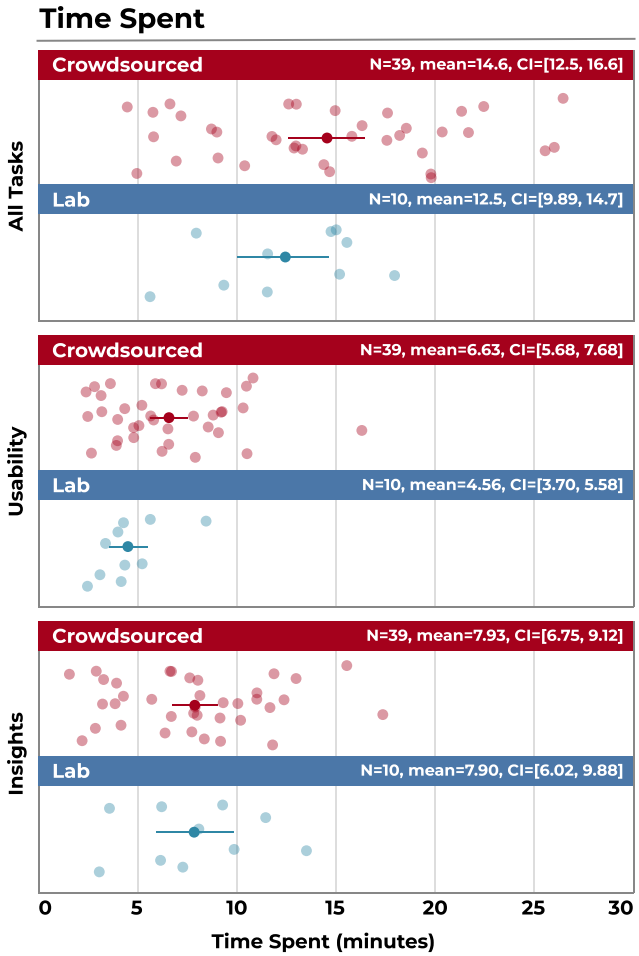

Figure 8: Time spent is lower in the lab overall, driven by lower times on the usability task. The time spent on the insight tasks is about the same.

the study, we lost data from two questions for this participant. As a result, this participant was removed from the analysis. Comments made by the researcher were removed from transcripts, although comments directed at the researcher from participants were left in.

Of the 41 online participants who completed our study, we gathered 39 complete transcripts. Two participants did not produce usable data. One participant did not speak at all, and one participant spoke throughout but had extremely poor audio quality, which could not be transcribed automatically or by hand.

There was a difference in correctness within the usability task responses, as in-person participants had 75% correct answers, compared to 65% for crowdsourced participants. Insight tasks only asked participants to think-aloud, so they did not have quantitative response data.

We also note that the quantitative analysis of codes is not always in line with the spirit of qualitative research, as coding can be imprecise and noisy. However, for the goals of this particular study—comparing user study methods rather than actually analyzing insights and usability problems—quantitative analysis enables useful comparisons between the methods under consideration.

RQ1: Will asynchronous participants talk as much as synchronous participants?

We found that (slightly) more words were spoken by asynchronous crowdsourced participants, (mean=1088, 95% confidence interval=[899, 1299]) than in-person (mean=868, 95% CI=[579, 1173]) participants (see Figure 7). Yet, the 95% confidence intervals overlap substantially, suggesting statistically similar results. This trend holds for both usability and insight questions, with a slightly larger gap during usability questions. Crowdsourced participants also took slightly longer on average, as seen in Figure 8, but with overlapping confidence intervals.

Interestingly, interaction count had the opposite trend as word count, as in-lab participants interacted more (mean=39.9, 95% CI=[20.3, 59.7]) than crowdsourced participants (mean=25.5, 95% CI=[17.3, 37.1]), as seen in Figure 9, though with overlapping confidence intervals. We speculate that interaction count correlates with effort, which is valuable to know since crowdsourced studies that often suffer from low-effort participants [17]. However, as mentioned in Section 5.1, crowdsourced participants did not receive a guided introduction (which for in-lab participants also introduced

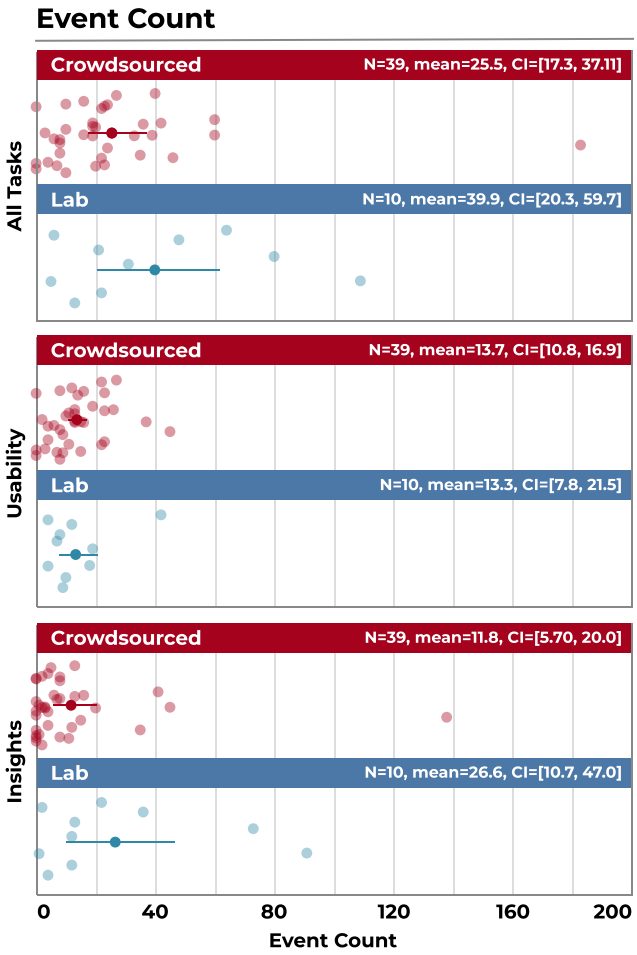

Figure 9: Crowdsourced and synchronous participants interacted at roughly equal rates for usability tasks, but crowdsourced participants interacted less in insight tasks. We observe a skewed distribution for insight tasks, where some participants interact heavily, but others do not. Many crowdsourced participants did not interact on the insight tasks.

brushing). Hence, it is possible that crowdsourced participants who did not fully read the instructions only discovered brushing during the study.

RQ2: What, if any, differences are there between asynchronous and synchronous responses for investigating usability?

As seen in Figure 10, usability comments were fairly rare in our study, for both in-person (mean=2.5, 95% CI=[0.25, 4.75]) and crowdsourced (mean=1.95, 95% CI=[1.16, 2.73]) participants. So, although our in-person participants do give slightly more UI/UX feedback, this is heavily influenced by one outlier in the in-person data who gave 11 of the total 25 usability comments.

Apart from the rate, there was a noticeable difference in the sentiment between asynchronous and synchronous participants. Of the 41 total UI/UX comments that were negative, 38 came from asynchronous participants, with only 3 from synchronous participants. This result suggests that while in-person participants are hesitant

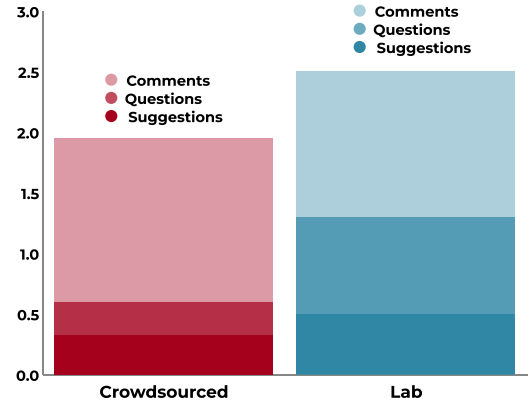

Figure 10: Average frequency of Usability-Related Comments, Questions, and Suggestions for the whole study. Crowdsourced participants raised fewer usability issues in total, and in particular asked fewer UI/UX questions.

to express a negative comment in front of an experimenter, online participants had few such concerns. However, of the 11 UI/UX comments that were positive, only 5 came from asynchronous participants and 6 from synchronous participants.

RQ3: What, if any, differences are there between asynchronous and synchronous responses for investigating insights?

Total insight counts were comparable, with 11.36 insights per crowdsourced participant and 9 insights per in-person participant across all tasks, as seen in Figure 11. Insight quality was also comparable, as seen in Figure 12. The percentage of insights that were correct, undirected, and contained a hypothesis were all within 2% of each other for crowdsourced and in-person. Crowdsourced participants did elaborate slightly more often on their insights, at 58% vs 41%.

RQ4: Are there logistical challenges in recruiting and recording crowdsourced participants?

In our previous experience conducting studies on Prolific, participant recruiting, even at large scales, is easy if the pay exceeds about $15 an hour. This was not quite the case for our TA study: it took around 30 minutes to find participants each time we started a new batch (we opened the study to 10 participants at a time over the course of a few hours). Part of the delay was the high rate of returned studies—we had more people return our study (46) than eventually complete it (41). This is much higher than the return rates we observed for other, more traditional studies.

We speculate that some participants were potentially uncomfortable with speaking out loud due to the study's novelty, were possibly in an environment where they could not speak out loud, or were having trouble granting browser permissions during the mic-check task. Overall, the crowdsourced condition still appears efficient (e.g., many in-lab participants will self-screen and choose not to participate in a study), but factors like these should be explored in follow-up studies.

RQ5: What is the difference in terms of effort for an experimenter?

To find students for our in-person study, we posted a recruitment message through university slack channels. There were 16 total slots for our study, of which 14 had people signed up. However, 3

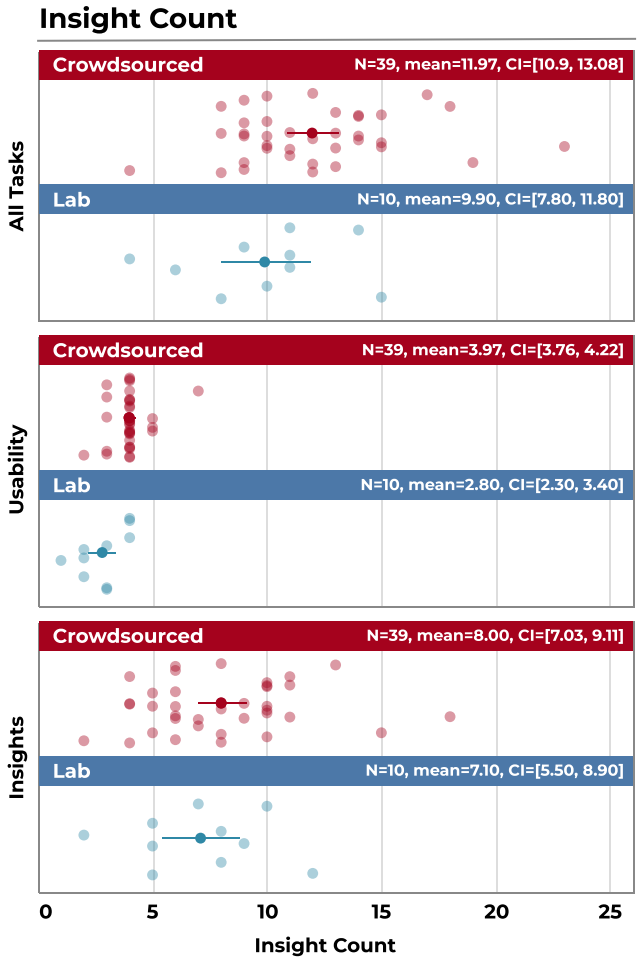

Figure 11: Insight counts were similar for participant types, with slightly more offered by crowdsourced participants.

Figure 12: Crowdsourced participants have a longer tail of high-response participants, with some participants stating up to 20 insights. The ratio of all characteristics is similar between the conditions, with the exception of Elaborated, for which Crowdsourced participants have more occurrences.

participants did not show up for the study, so we ended with 11 participants. Each slot was 30 minutes, so a researcher was either with a student or waiting for a student for a total of 7 hours. This does not account for time requesting a room that could be used for the study, ensuring the setup with monitors/computer would work beforehand, or managing logistics to pay the participants.

Very little additional effort was required to conduct the study on Prolific. ReVISit already has support for Prolific integration. We did not change any of our recording instrumentation from in-person to Prolific, as both used CrowdAloud to record audio and provenance data. However, it is worth noting that our level of instrumentation (provenance tracking) is desirable for both scenarios, but it is necessary for asynchronous studies.

We found that in-person studies are also more expensive per participant. In-person participants were paid $20 for a study that ended up taking less than 20 minutes. We spent $220 on the 11 participants who completed our in-person study. We could have chosen a lower rate, but we felt that people would be unwilling to schedule an experiment and physically travel to the location for a smaller reward. Even with a pay rate exceeding $60 an hour, we could not fill all of our desired slots and had 3 no-shows.

In comparison, we paid our 41 participants from Prolific $5 for a rate that came out to $15.68 an hour, and with Prolific fees included, we spent a total of $273. Prolific enforces an hourly rate of $8 and recommends $12. An advantage for crowdsourced participants is that they have barely any time overhead (scheduling, traveling) compared to in-lab participants.

RQ6: *Do crowdsourced participants find TA studies enjoyable or uncomfortable?* In a poststudy survey, we asked participants questions on their views regarding think-aloud studies. First, in response to the question *"Do you have any privacy concerns about having your audio recorded?"* no participants reported concerns, and some participants justified their lack of concern, saying, for example, "*[CS-TA] I didn't have any concerns. Really, my voice can't realistically be linked to me as a person*".

Responses varied greatly to the question *"Please provide any comments about the process of answering questions out loud"*. Many participants responded positively, frequently saying TA helped them focus or was natural: "*[CS-TA] Sometimes it is easier thinking and talking out loud to help understand the scenario better*", "*[CS-TA] It was smooth. I liked the visualizations, and the interactions made it easier to communicate the results.*"

However, quite a few participants believed TA took away from their ability to answer the question, saying, "*[CS-TA] I found that I could not really concentrate on the question that well while answering aloud*", "*[CS-TA] I felt it hard to interpret the information and verbalize it*".

Others felt they lacked direction on what to say, "*[CS-TA] Wasn't sure what to say and rambled a bit but enjoyed the task*", "*[CS-TA] I was sometimes confused about what to say*". One participant, who was not a native English speaker, stated that they were worried about their accent, though they also thought TA had benefits, "*[CS-TA] It feels a bit awkward and I am always a bit conscious about my accent. Answering questions out loud seems to slow down the process of answering but also feels like it allows me to explore the questions in depth.*"

In-lab participants mostly had positive comments about the TA process, though one in-lab participant also believed focus was split between TA and tasks, saying, "*[LAB-TA] It often happens that focus is more on thinking about the answer and the thoughts are not spoken*"

5.5 Discussion

We were excited to conclude that crowdsourced participants consistently speak their thoughts out loud and that more invasive interventions to trigger verbalization seem unnecessary.

Usability Feedback. The biggest difference in the usability feedback was the tone of UI/UX comments. Almost all negative comments were provided by crowdsourced participants (38 vs. 3), while positive comments were majority in-person (6 vs. 5). Some of this can be explained by in-person participants phrasing criticism as suggestions or questions, such as one who suggested, "*[LAB-TA] It would be cool to, I don't know, use the scroll wheel or something to expand this circle so we don't have to drag everywhere.*" Some of the positive comments from in-person were encouraging but uncritical comments, such as "*[LAB-TA] Oh, this is a nice tool*" and "*[LAB-TA] Okay, this is a nice chart*".

There were also certain problems that went unmentioned by in-person participants but were frequently highlighted by crowdsourced participants. For example, a few participants mentioned a lag on the selection brush when brushing very quickly, saying, for example, "*[CS-TA] The problem is, I'm not sure if it's a device issue. But um, it's taking time to go to the points. There's sort of a lag.*" The experimenter observed that multiple in-person participants encountered the same problem, but none critiqued it.

However, it is also worth noting that usability comments were not the main focus of participants' speech. Combining both in-person and crowdsourced, for all UI/UX comments, questions, and suggestions there were a total of 85 mentions, compared to 543 total insights spoken. Our prompts asked for both insights and feedback on problems encountered. We speculate that a prompt that more explicitly elicits feedback from participants will lead to more UI/UX comments, and that it might be beneficial to separately study UI/UX and insight formation.

Insight Quality. Insight quality was similar between participant types, as can be seen in Figure 12. *Undirected* comments, which showed people engaging with a question beyond what was asked of them, were very similar with 29% of crowdsourced insights and 27% of in-person insights.

Insights that contain a *hypothesis* (connect previous knowledge with new knowledge) show a similar picture, with almost no difference between crowdsourced (11%) and in-person (10%) participants.

The only difference in insight quality appeared in elaborated insights, where crowdsourced participants more frequently explained what they were thinking (58% to 41%).

As a whole, these results are very encouraging for the prospects of Crowdsourced TA. Insight quantity and quality were remarkably similar between crowdsourced and in-person participants, and we found no reason to believe crowdsourced participants will not give at least equal insight quality in comparison to in-person TA studies. Usability feedback differs in tone, and we echo the results of Gamboa et al. [39] in finding no significant difference between the quantity of usability comments.

5.6 Limitations

Our in-person study had a small sample size and a limited demographic pool of only graduate computer science students. These were intentional parameters to reflect typical studies, which we discussed in section 2, in which the median study had 10 participants and almost all novices were students. However, for our usability analysis especially, a higher sample size may have yielded more definitive results.

6 Study 2: Online Think-Aloud vs Online Text Response

In this second study, we compare two crowdsourced modalities for asynchronously gathering data about insights and insight formation: using a crowdsourced think-aloud protocol (shown as red, quotes "*[CS-TA] like this*"), and using a text box where participants enter their insights (shown as purple, quotes "*[CS-TXT] like so*"). Previous work has shown synchronous TA effective as a method to elicit insights, including in video-conference settings [13, 68, 87]. In contrast, text responses are commonly used to query insights in crowdsourced settings [70, 101]. Researchers need to make a choice between TA and text responses. Due to the increased overhead associated with TA studies, it is important to show TA has a benefit over text responses, which is the purpose of this study.

We focused on insight gathering and chart understanding in this study and wanted to minimize the complexity of the system. Hence, our stimulus is static, and we did not expect UI feedback. We ask the following research questions:

- **RQ7:** Do crowdsourced participants provide more insights when speaking than when responding with text?
- **RQ8:** What, if any, differences are there between spoken and text responses? Do spoken responses let us glean insight into the process of chart understanding and insight formation more than text responses?

6.1 Procedure

Study 2 was reviewed by the local IRB and was deemed exempt from full board review. We obtained consent from participants before conducting the study. The study was deployed on the crowdsourcing platform Prolific where we recruited 20 participants. All

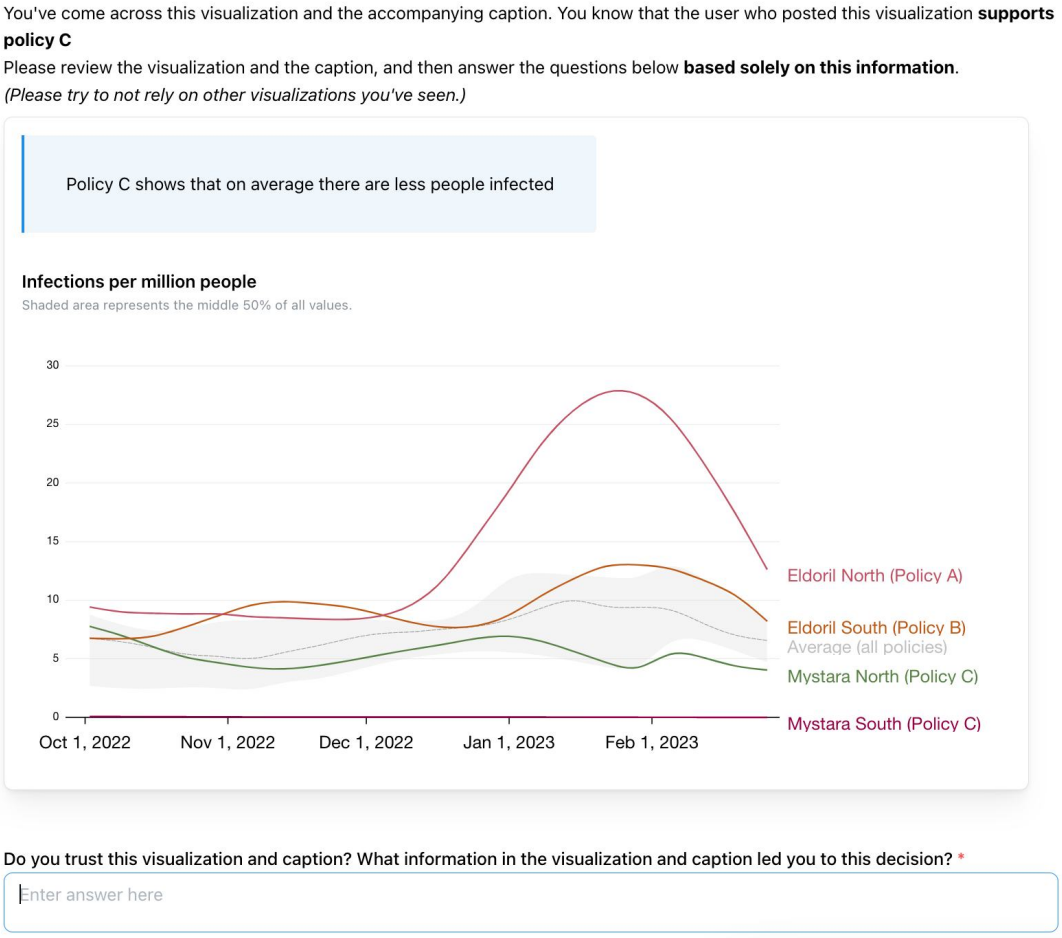

Figure 13: An example of the stimulus used in Study 2. Participants were asked to analyze a chart showing cases of a fictional disease. In half of the tasks they were asked to provide text responses, as shown here, in the other half they were asked to think aloud. [↗ Stimulus]

crowdsourced participants indicated they were proficient in English, and we recruited only from individuals residing in the USA or UK. Online participants were paid $5.50 for a study which averaged just under 20 minutes in length, for $15.99 per hour.

Before conducting our study, we ran a 3 participant pilot using Prolific. The pilot was used to ensure that all data was being saved correctly and that there were no problems with the stimulus. No changes were made between the pilot and the study. We do not include data from the pilot in our study. The full study can be seen at https://vdl.sci.utah.edu/guardrails-think-aloud. All of the code required to reproduce the study can be found at https://github.com/visdesignlab/guardrails-think-aloud.

6.2 Study Stimulus

Our study makes use of the data generated by Lisnic et al. [58] to test the viability of adding *guardrails* to visualizations shared on social media in order to make readers aware of cherry-picking. In their study, they asked participants to create cherry-picked time-series charts and caption them to make a point not supported by the full dataset. They designed and implemented 4 different guardrails to mitigate cherry-picking. Charts could be one of two anonymized datasets based on COVID-19 deaths and stock data. For more details on the dataset, please refer to the original paper [58].

We show participants five different charts from the same data and scenario, one for each guardrail and one with no guardrail, once in the text response condition (shown in Figure 13), and once in the think-aloud condition (a within-subject design). The order of the datasets and which dataset is TA vs text response is controlled for using a Latin square. Each question shows participants a chart and an associated caption and simply asks, 'Do you trust this visualization and caption? What information in the visualization and caption led you to this decision?' For text responses, there is a textbox below the question. For TA responses, there is a checkbox that must be checked indicating 'I have verbalized my response' to advance to the next question.

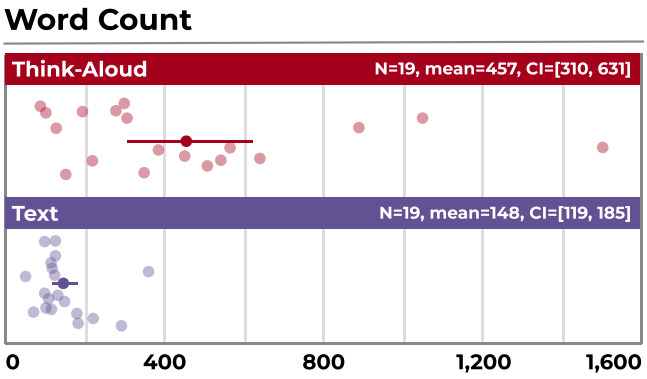

Figure 14: Word counts for think-aloud responses are much higher than text responses on average.

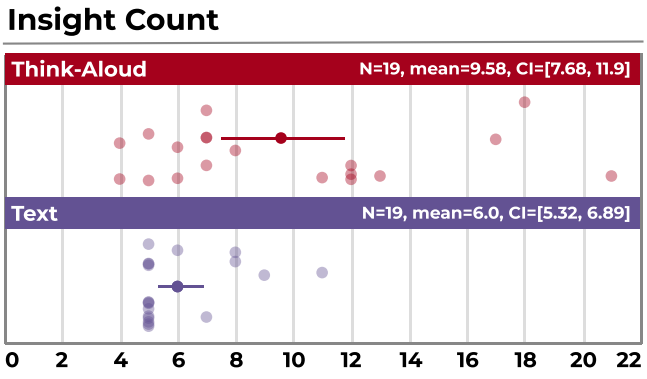

Figure 15: Insight counts are higher for the Think-Aloud condition. As there were 5 tasks in each condition, we see that text responses mostly elicit a single insight per task; the mean number of insights is 6.0. Think-Aloud responses contain about 1.5 times as many insights (mean of 9.58), but also show greater variance.

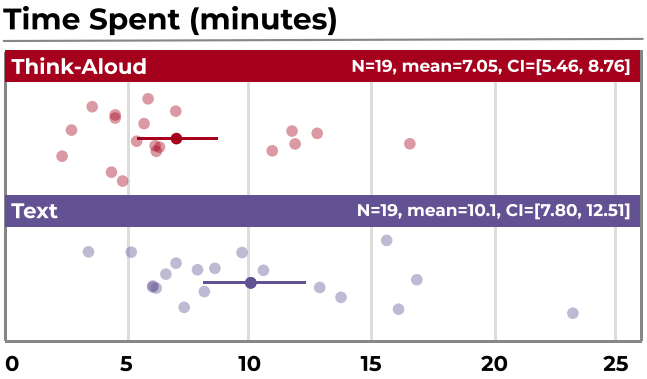

Figure 16: Think-aloud produced more insights, but the time spent on responding was slightly lower on average.

6.3 Codebook Development and Coding

As the nature of the tasks and the questions differed from Study 1 (no directions, no measure of correctness), we only re-used the `Elaborated` code. We also added a code to capture `Reasoning Process`, to investigate whether TA responses reveal more about the chart understanding and insight formation process. An insight was tagged as a reasoning process if it shows that a participant is weighing different considerations against each other while coming to a final choice (as opposed to explaining their thought process when `Elaborating`). For example, one participant indicated they trusted the chart but wanted more information on the source of the data, "*[CS-TXT] Although it does look like there are less infections with policy A, I am not able to verify the source of the data and whether it is accurate so I reserve judgment.*"

6.4 Results

Of the 20 participants who took the study, we gathered 19 complete transcripts used in the analysis. We received no audio data from one participant, although they passed the mic-check task, so we removed this participant. We did not remove any text from the transcripts before analysis.

RQ7: Do asynchronous participants provide more insights with TA than with a textbox response?

As seen in Figure 14, the average word count is much higher in TA responses compared to text responses, which is consistent with our findings of study 1. Figure 15 shows that the extra words also are rich in content: the average number of insights is about 1.5 times as high in the TA condition, totaling 182 insights for TA responses to 114 for text responses.

RQ8: What, if any, differences are there between TA and text responses for investigating insights?

Although they vary in quantity, the quality of insights between TA and text responses in Study 2 was similar. TA insights are `Elaborated` slightly more often (84% to 80%). Insights in which a participant explained their `Reasoning Process` (29% to 25%) also occurred slightly more often during TA responses. However, as there were many more insights in total in TA responses and the rate of elaboration and explaining one's reasoning process was about the same, there are also more examples of these behaviors in TA responses in total. There is a difference in time spent between conditions, with text responses taking longer than TA responses, though there is some overlap in the CIs, as shown in Figure 16.

Many insights in the TA responses provided glimpses into participants' reasoning processes. One participant indicated they trusted the graph but believed there was nuance in the data that was not captured in the graph caption. "*[CS-TA] the car stocks do look good, they do seem to be stable for a while, but I'm still not sure if I trust this caption, I agree that this investment has showed significant growth upwards but it doesn't seem able to maintain highs for any length of time*". Another participant indicated they trusted a graph "*[CS-TA] I would moderately trust this visualization and the caption uh, they're correct in saying that the cases are trending downwards. That's easy to see.*" However, they later backtracked on this trust, saying, "*[CS-TA] What I do find interesting, and gives me a little bit of distrust about this data, about this visualization, is that all of the other policy data lines showed a much lower, uh, level of infections per million people even at their peak.*"

Many text responses were short, focused on only one insight, yet still included details on reasoning processes. For example, one participant weighed their trust in the graph against their trust in the caption, saying, "*[CS-TXT] I wouldn't agree that the airline A has funds that soar but it is evident it has outperformed Car B looking at this graph.*". Another participant disagreed with a graph caption, though indicated they saw some merit to the argument: "*[CS-TXT] Policy A doesn't appear to be very effective although it does have a sharper fall of infections than the other policies.*"

6.5 Discussion

The increased rate of insight and shorter completion times indicate that even for the simple baseline case of a static chart, TA studies are a viable alternative to text responses. However, this has to be weighed against the simplicity and familiarity of text boxes.

We speculate that the study of insights, insight formation, and chart understanding on interactive charts would benefit even more from the TA protocol and associated provenance/screen tracking, as it enables participants to express insights for intermediate states, whereas a text-box is most suited for reflections on a final state.

With regards to the differences in learning about the insight generation or chart understanding process (RQ8), we found that participants in both conditions elaborated or revealed their reasoning process at the same rates. However, qualitatively, we found TA responses to be longer, more detailed, and reflect a changing state of mind as participants were understanding the chart, which might be helpful to researchers who want to investigate chart understanding.

7 Discussion and Design Recommendations

Based on our experience running two crowdsourced TA studies, we present the following recommendations for designing such studies.

7.1 Ethical Considerations

Listening to and storing audio data from crowdsourced participants has ethical implications. Crowdsourced participants may not be in a location in which they were prepared to be recorded, and those around them may be unaware they are being recorded. Two of our participants were interrupted multiple times during their study by a third party, presumably their children, who can be heard on the recording. Of course, in these cases, the third party did not consent to be recorded. In one of our pilots, a participant came close to de-anonymizing themselves (gave their first name and exact location) during their mic-check, causing us to not store the mic-check recording in study 2. Hence, we recommend that researchers ensure that they only record sections of the study they plan to use in analysis and we advocate for the development of protocols and designs that minimize risks.

These incidents were minor, but we should consider what happens when they are not. A participant, for example, may incidentally disclose identifiable or embarrassing information during a study. We are considering a way for participants to remove their own data at any time during the study as a remedy. Another approach to limit unintentional disclosures could be to not always record a participant, but instead to allow them to start and stop the recording at will. More research will need to be done to investigate if such a solution affects results.

7.2 Increased Negativity in Feedback

As discussed in subsection 5.5, there was a noticeable difference in tone between crowdsourced and in-person users for usability feedback, with crowdsourced being much more negative. We believe that the main factor changing the tone of the responses is the presence of an experimenter: social pressure may lead to more positive feedback or self-censorship on negative feedback, while anonymous comments (such as crowdsourced responses) may lead to less filtered feedback. However, crowdworkers may also be more sensitive to usability problems in a study, especially when problems or bugs may hinder their ability to complete a study and thus get paid. Similar problems arising during in-person studies would likely not have the same monetary consequences for participants.

We speculate that this difference in tone between in-person and crowdsourced participants could influence analysis outcomes, especially for studies that are focused on usability. However, it is not clear that a more unfiltered perspective is always undesirable, as it might lead to the identification of issues that are otherwise overlooked. At the same time, crowdsourced participants might hold back on positive feedback that is also useful for evaluation. Ultimately, we believe that this trade-off should be considered when designing a study.

7.3 Focus on Training & Encourage Interaction

Crowdsourced participants in Study 1 interacted noticeably less than their in-person counterparts and had slightly worse results on the tasks with defined answers. Multiple crowdsourced participants mentioned, as shown in subsection 5.4, that it was difficult to focus on the task at hand while also speaking. Another participant was more specific about the same problem, saying, "*[CS-TA] I'm just aware of pauses and uums and aahs while trying to assess something visually*". This cognitive overload may be eased if participants were given more training and were already comfortable with the system once the TA process began. In-person participants in our study were walked through the introduction by the experimenter and could ask questions about the system or data. We hence echo previous findings [70] that special care has to be taken in training crowdsourced participants, e.g., using unskippable videos, or techniques such as *feedforward* [91], which have been proven to be effective at encouraging interaction in crowdsourced studies [15]. It also may be important to specify what kind of feedback is desired from participants. In our studies, we asked for a wide range of feedback, including *'insights, questions or problems'*. This resulted in crowdsourced participants interpreting this differently on a case-by-case basis, with one participant beginning to give important usability feedback but stopping themselves as they did not think it was relevant, saying, "*[CS-TA] Do wish there was an option to "highlight all". Like, there's "clear selection", but I'm maybe talking outside the scope of the study at the moment. So I will be quiet while I highlight all of these points*"

7.4 Trade-off Between Instrumentation and Scalability

As discussed in subsection 5.4, there are potential time and expense savings to be realized by using a crowdsourced think-aloud study. However, instrumentation with CrowdAloud may not suit every

study, as there is a trade-off to using CrowdAloud: there is some additional effort required to instrument a stimulus. Provenance tracking provides ample benefits unrelated to TA studies, but it may be tedious to retrofit a larger system to properly track the state. Provenance tracking may also prove difficult to implement for less technical users, or for stimuli that are not created in code.

We note that provenance tracking is not required for any of the CrowdAloud features to work, and many studies that can benefit from TA (including our own Study 2) are conducted without any interaction with the stimulus.

Conceptually, CrowdAloud could also integrate video from screen recordings instead of action replays. The downside of video are (a) that screen capturing requires permissions, (b) that the video would substantially increase the data that has to be uploaded, and (c) that events couldn't be automatically captured. However, video also has advantages (in addition to the ease of implementation). Specifically, a screen recording would guarantee that the experimenter sees the interface exactly as the participant sees it, including issues specific to browsers and devices (e.g., screen resolution and window size). Because of this, we plan on integrating screen capture in future versions of CrowdAloud.

We argue that provenance-tracked CrowdAloud is most suited for evaluating complex interactive visualization methods, such as comparing interactive node-link diagrams and adjacency matrices [70]. For such studies, the overhead of instrumentation is likely small, and having both provenance data and think-aloud data enables experimenters to do a rich analysis of paths to solutions or insights [69] as well as participants' thoughts and concerns.

7.5 Trade-off Between Ecological Validity and Experimental Control

As discussed, crowdsourcing TA studies results in fewer logistical challenges and lower costs for conducting studies, while increasing the technical burden for preparing stimuli. However, Crowdsourced TA studies also result in more variability in the study, as researchers have less control over the conditions (hardware, browsers, operating systems) in which a study takes place. For example, many participants in Study 1 reported a lag when brushing, which might have been exaggerated by slower hardware the participants used. The resulting tradeoff between ecological validity and experimental control is common in crowdsourced studies [45]. Additionally, compared to traditional crowdsourced studies, TA may add another source of variability, as the quality of the microphone, background noise, and possible interruptions add additional differences between participants and may reduce the number of participants willing to take the study, as evidenced by the high return rates we observed for both studies. As discussed in Section 7.2 the absence of a researcher may also have a larger impact on results than for a crowdsourced study not involving TA.

However, most TA studies are focused on usability and insight formulation, studies that may benefit from the increased ecological validity that is associated with a wider range of environments. We note that audio quality was only a slight concern in our studies, with one participant who could not be automatically transcribed, and a few participants who did require heavy editing of the transcription due to background noise, typically a television. Given the low cost and overhead of recruiting participants, it is reasonable to compensate for such issues with higher participant numbers.

7.6 Sample Sizes

It may be tempting to conduct large-scale CrowdAloud TA studies, as the cost of recruiting additional participants is relatively low. However, TA data is still cumbersome to analyze. While our CrowdAloud analysis platform increases efficiency by enabling analysts to focus on just the relevant moments of a participant's session, they still have to listen to participants' speech and read transcripts, develop codebooks, and code the data.

However, the ability to generate large volumes of text feedback might make complementary AI-powered mixed initiative approaches a worthwhile analysis strategy [52].

8 Future Work

8.1 Consider Speech Interventions

For our studies, we intentionally avoided any interventions that encouraged speech and were pleased to find that such interventions are not necessary for participants to speak at similar levels to in-person studies. However, as shown in Figure 7, the range of words spoken by crowdsourced participants is quite wide, and some participants spoke much less than we would have liked. Thus, employing interventions to increase crowdsourced speech may be useful. We did consider interventions that could be tested in future work and would like to discuss how our experiences have influenced our perception of such interventions.

First, we considered adding a time minimum on questions that had no given answer and only required participants to talk. In retrospect, we believe this unlikely to achieve positive results, as even on tasks where participants said nothing or very little, they almost always still took a reasonable amount of time on the task. We are also concerned about the potential for participants to be temporarily stuck on a task they have completed, and quit the study.

Second, we considered a gradually more salient reminder if we had not heard speech in a while, such as a flashing reminder or even an audio prompt (as done by Gamboa et al. [39]).

Although this may increase rates of speech, the feedback received has made us more hesitant to implement such a system. As discussed in subsection 5.4 and subsection 7.3, some participants believed they already were unable to focus on the actual task while speaking. Another participant in the post-study survey simply said, "there was more pressure" when asked about the use of TA. Reminders to speak would likely increase pressure on participants.

8.2 Expert Asynchronous TA Studies

Our validation focused on crowdsourced studies and did not investigate asynchronous TA studies on experts. The only asynchronous TA study we found [72] was run on experts, however, and this was the most common type of synchronous TA study. We generally believe that asynchronous TA studies with experts are of high value, as experts can participate in a study at their own convenience. They may also be more critical, as we found with crowdsourced participants. However, this effect might be mitigated by the expert knowing the authors of the study. The main downside of asynchronous studies is the inability to conduct semistructured interviews.

Instead, such studies would have to rely on structured feedback questions that can be answered either in text or by speech.

9 Conclusion

In this work, we introduce CrowdAloud, a platform for conducting and analyzing crowdsourced TA studies. We utilize CrowdAloud to conduct two studies, in order to test the feasibility of asynchronous TA studies. The first study, which explored both usability testing and elicitation of insights, compared crowdsourced TA and in-lab TA studies. We find that crowdsourced participants speak as much, if not more, than in-lab participants, without any necessary interventions to encourage speaking. For usability testing, we echo previous findings [39] that crowdsourced and in-lab usability tests produce similar results. We also find that crowdsourced participants elicit similar quality and quantity of insights in comparison to in-lab participants. Our second study, intended to compare crowdsourced TA responses to more traditional text responses, found similar quality of insights but a much higher quantity of insights using TA. Overall, we find that aside from increased instrumentation overhead, crowdsourced think-aloud studies are a viable alternative to lab studies.

Although future work should be done to determine if our results hold for a diverse set of studies and use cases, we hope that our findings encourage the visualization community to conduct more crowdsourced TA studies in the future.

Acknowledgments

We thank Max Lisnic for help with Study 2 and the reviewers for their constructive feedback throughout the revision process, which helped substantially strengthen the manuscript. This work is supported by the National Science Foundation (NSF CNS 2213756, 2213757, and 2313998).

References

[1] Moataz Abdelaal, Nathan D. Schiele, Katrin Angerbauer, Kuno Kurzhals, Michael Sedlmair, and Daniel Weiskopf. 2023. Comparative Evaluation of Bipartite, Node-Link, and Matrix-Based Network Representations. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 896–906. doi:10.1109/TVCG.2022.3209427

[2] Derya Akbaba, Lauren Klein, and Miriah Meyer. 2025. Entanglements for Visualization: Changing Research Outcomes through Feminist Theory. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (Jan. 2025), 1279–1289. doi:10.1109/TVCG.2024.3456171

[3] Derya Akbaba, Devin Lange, Michael Correll, Alexander Lex, and Miriah Meyer. 2023. Troubling Collaboration: Matters of Care for Visualization Design Study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–15. doi:10.1145/3544548.3581168

[4] Obead Alhadreti and Pam Mayhew. 2018. Rethinking Thinking Aloud: A Comparison of Three Think-Aloud Protocols. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173618

[5] Sriram Karthik Badam, Zhicheng Liu, and Niklas Elmqvist. 2019. Elastic Documents: Coupling Text and Tables through Contextual Visualizations for Enhanced Document Reading. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 661–671. doi:10.1109/TVCG.2018.2865119

[6] Andrea Batch, Andrew Cunningham, Maxime Cordeil, Niklas Elmqvist, Tim Dwyer, Bruce H. Thomas, and Kim Marriott. 2020. There Is No Spoon: Evaluating Performance, Space Use, and Presence with Expert Domain Users in Immersive Analytics. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 536–546. doi:10.1109/TVCG.2019.2934803

[7] Leilani Battle and Alvitta Ottley. 2023. What Exactly is an Insight? A Literature Review. In *2023 IEEE Visualization and Visual Analytics (VIS)*. IEEE, Melbourne, Australia, 91–95. doi:10.1109/VIS54172.2023.00027

[8] Michael Behrisch, Tobias Schreck, and Hanspeter Pfister. 2020. GUIRO: User-Guided Matrix Reordering. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 184–194. doi:10.1109/TVCG.2019.2934300

[9] Donald Bertucci, Md Montaser Hamid, Yashwanthi Anand, Anita Ruangrotsakun, Delyar Tabatabai, Melissa Perez, and Minsuk Kahng. 2023. DendroMap: Visual Exploration of Large-Scale Image Datasets for Machine Learning with Treemaps. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 320–330. doi:10.1109/TVCG.2022.3209425

[10] Tanja Blascheck, Markus John, Kuno Kurzhals, Steffen Koch, and Thomas Ertl. 2016. VA2: A Visual Analytics Approach for Evaluating Visual Analytics Applications. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 61–70. doi:10.1109/TVCG.2015.2467871

[11] Tanja Blascheck, Lindsay MacDonald Vermeulen, Jo Vermeulen, Charles Perin, Wesley Willett, Thomas Ertl, and Sheelagh Carpendale. 2019. Exploration Strategies for Discovery of Interactivity in Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 2 (Feb. 2019), 1407–1420. doi:10.1109/TVCG.2018.2802520

[12] Jeremy E. Block, Shaghayegh Esmaeili, Eric D. Ragan, John R. Goodall, and G. David Richardson. 2023. The Influence of Visual Provenance Representations on Strategies in a Collaborative Hand-off Data Analysis Scenario. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 1113–1123. doi:10.1109/TVCG.2022.3209495

[13] T. Boren and J. Ramey. 2000. Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication* 43, 3 (Sept. 2000), 261–278. doi:10.1109/47.867942

[14] R. Borgo, L. Micallef, B. Bach, F. McGee, and B. Lee. 2018. Information Visualization Evaluation Using Crowdsourcing. *Computer Graphics Forum* 37, 3 (2018), 573–595. doi:10.1111/cgf.13444

[15] Jeremy Boy, Louis Eveillard, Françoise Detienne, and Jean-Daniel Fekete. 2016. Suggested Interactivity: Seeking Perceived Affordances for Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 639–648. doi:10.1109/TVCG.2015.2467201

[16] A.J. Bernheim Brush, Morgan Ames, and Janet Davis. 2004. A comparison of synchronous remote and local usability studies for an expert interface. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1179–1182. doi:10.1145/985921.986018

[17] Florian Brühlmann, Serge Petralito, Lena F. Aeschbach, and Klaus Opwis. 2020. The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology* 2 (Nov. 2020), 100022. doi:10.1016/j.metip.2020.100022

[18] Dylan Cashman, Shenyu Xu, Subhajit Das, Florian Heimerl, Cong Liu, Shah Rukh Humayoun, Michael Gleicher, Alex Endert, and Remco Chang. 2021. CAVA: A Visual Analytics System for Exploratory Columnar Data Augmentation Using Knowledge Graphs. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 1731–1741. doi:10.1109/TVCG.2020.3030443

[19] Krista Casler, Lydia Bickel, and Elizabeth Hackett. 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior* 29, 6 (Nov. 2013), 2156–2160. doi:10.1016/j.chb.2013.05.009

[20] Marco Cavallo and Çağatay Demiralp. 2019. Clustrophile 2: Guided Visual Clustering Analysis. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 267–276. doi:10.1109/TVCG.2018.2864477

[21] Davide Ceneda, Alessio Arleo, Theresia Gschwandtner, and Silvia Miksch. 2022. Show Me Your Face: Towards an Automated Method to Provide Timely Guidance in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (Dec. 2022), 4570–4581. doi:10.1109/TVCG.2021.3094870

[22] Qing Chen, Nan Chen, Wei Shuai, Guande Wu, Zhe Xu, Hanghang Tong, and Nan Cao. 2024. Calliope-Net: Automatic Generation of Graph Data Facts via Annotated Node-Link Diagrams. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 562–572. doi:10.1109/TVCG.2023.3326925

[23] Ran Chen, Di Weng, Yanwei Huang, Xinhuan Shu, Jiayi Zhou, Guodao Sun, and Yingcai Wu. 2023. Rigel: Transforming Tabular Data by Declarative Mapping. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 128–138. doi:10.1109/TVCG.2022.3209385

[24] Arjun Choudhry, Mandar Sharma, Pramod Chundury, Thomas Kapler, Derek W. S. Gray, Naren Ramakrishnan, and Niklas Elmqvist. 2021. Once Upon A Time In Visualization: Understanding the Use of Textual Narratives for Causality. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 1332–1342. doi:10.1109/TVCG.2020.3030358

[25] Tarik Crnovrsanin, Shilpika, Senthil Chandrasegaran, and Kwan-Liu Ma. 2021. Staged Animation Strategies for Online Dynamic Networks. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 539–549. doi:10.1109/TVCG.2020.3030385

[26] Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLOS ONE* 8, 3 (March 2013), e57410. doi:10.1371/journal.pone.0057410

[27] Yuan Cui, Lily W. Ge, Yiren Ding, Fumeng Yang, Lane Harrison, and Matthew Kay. 2024. Adaptive Assessment of Visualization Literacy. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 628–637. doi:10.1109/TVCG.2023.3327165

[28] Zach Cutler, Kiran Gadhave, and Alexander Lex. 2020. Trrack: A Library for Provenance-Tracking in Web-Based Visualizations. In *2020 IEEE Visualization Conference (VIS)*. IEEE, Salt Lake City, UT, USA, 116–120. doi:10.1109/VIS47514.2020.00030

[29] Aritra Dasgupta, Joon-Yong Lee, Ryan Wilson, Robert A. Lafrance, Nick Cramer, Kristin Cook, and Samuel Payne. 2017. Familiarity Vs Trust: A Comparative Study of Domain Scientists' Trust in Visual Analytics and Conventional Analysis Methods. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan. 2017), 271–280. doi:10.1109/TVCG.2016.2598544

[30] Zikun Deng, Shifu Chen, Tobias Schreck, Dazhen Deng, Tan Tang, Mingliang Xu, Di Weng, and Yingcai Wu. 2024. Visualizing Large-Scale Spatial Time Series with GeoChron. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 1194–1204. doi:10.1109/TVCG.2023.3327162

[31] Vaishali Dhanoa, Conny Walchshofer, Andreas Hinterreiter, Eduard Gröller, and Marc Streit. 2023. Fuzzy Spreadsheet: Understanding and Exploring Uncertainties in Tabular Calculations. *IEEE Transactions on Visualization and Computer Graphics* 29, 2 (Feb. 2023), 1463–1477. doi:10.1109/TVCG.2021.3119212

[32] Yiren Ding, Jack Wilburn, Hilson Shrestha, Akim Ndlovu, Kiran Gadhave, Carolina Nobre, Alexander Lex, and Lane Harrison. 2023. reVISit: Supporting Scalable Evaluation of Interactive Visualizations. In *2023 IEEE Visualization and Visual Analytics (VIS)*. IEEE, Melbourne, Australia, 31–35. doi:10.1109/VIS54172.2023.00015

[33] Will Epperson, Vaishnavi Gorantla, Dominik Moritz, and Adam Perer. 2024. Dead or Alive: Continuous Data Profiling for Interactive Data Science. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 197–207. doi:10.1109/TVCG.2023.3327367

[34] K. Anders Ericsson and Herbert A. Simon. 1980. Verbal reports as data. *Psychological Review* 87, 3 (May 1980), 215–251. doi:10.1037/0033-295X.87.3.215

[35] Mingming Fan, Ke Wu, Jian Zhao, Yue Li, Winter Wei, and Khai N. Truong. 2020. VisTA: Integrating Machine Intelligence with Visualization to Support the Investigation of Think-Aloud Sessions. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 343–352. doi:10.1109/TVCG.2019.2934797

[36] Takanori Fujiwara, Jian Zhao, Francine Chen, and Kwan-Liu Ma. 2020. A Visual Analytics Framework for Contrastive Network Analysis. In *2020 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Salt Lake City, 48–59. doi:10.1109/VAST50239.2020.00010

[37] Aimen Gaba, Vidya Setlur, Arjun Srinivasan, Jane Hoffswell, and Cindy Xiong. 2023. Comparison Conundrum and the Chamber of Visualizations: An Exploration of How Language Influences Visual Design. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 1211–1221. doi:10.1109/TVCG.2022.3209456

[38] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. 2017. Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, Daniel Archambault, Helen Purchase, and Tobias Hoßfeld (Eds.). Springer International Publishing, Cham, 6–26. doi:10.1007/978-3-319-66435-4_2

[39] Edwin Gamboa, Rahul Galda, Cindy Mayas, and Matthias Hirth. 2021. The Crowd Thinks Aloud: Crowdsourcing Usability Testing with the Thinking Aloud Method. In *HCI International 2021 - Late Breaking Papers: Design and User Experience: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 24–39. doi:10.1007/978-3-030-90238-4_3

[40] Lin Gao, Zekai Shao, Ziqin Luo, Haibo Hu, Cagatay Turkay, and Siming Chen. 2024. TransforLearn: Interactive Visual Tutorial for the Transformer Model. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 891–901. doi:10.1109/TVCG.2023.3327353

[41] Bhavya Ghai and Klaus Mueller. 2023. D-BIAS: A Causality-Based Human-in-the-Loop System for Tackling Algorithmic Bias. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 473–482. doi:10.1109/TVCG.2022.3209484

[42] Hua Guo, Steven R. Gomez, Caroline Ziemkiewicz, and David H. Laidlaw. 2016. A Case Study Using Visualization Interaction Logs and Insight Metrics to Understand How Analysts Arrive at Insights. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 51–60. doi:10.1109/TVCG.2015.2467613

[43] Hua Guo and David H. Laidlaw. 2020. Topic-Based Exploration and Embedded Visualizations for Research Idea Generation. *IEEE Transactions on Visualization and Computer Graphics* 26, 3 (March 2020), 1592–1607. doi:10.1109/TVCG.2018.2873011

[44] Qi Han, Dennis Thom, Markus John, Steffen Koch, Florian Heimerl, and Thomas Ertl. 2020. Visual Quality Guidance for Document Exploration with Focus+Context Techniques. *IEEE Transactions on Visualization and Computer Graphics* 26, 8 (Aug. 2020), 2715–2731. doi:10.1109/TVCG.2019.2895073

[45] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Atlanta Georgia USA, 203–212. doi:10.1145/1753326.1753357

[46] Md Naimul Hoque and Klaus Mueller. 2022. Outcome-Explorer: A Causality Guided Interactive Visual Interface for Interpretable Algorithmic Decision Making. *IEEE Transactions on Visualization and Computer Graphics* 28, 12 (Dec. 2022), 4728–4740. doi:10.1109/TVCG.2021.3102051

[47] Tom Horak, Norine Coenen, Niklas Metzger, Christopher Hahn, Tamara Flemisch, Julián Méndez, Dennis Dimov, Bernd Finkbeiner, and Raimund Dachselt. 2022. Visual Analysis of Hyperproperties for Understanding Model Checking Results. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 357–367. doi:10.1109/TVCG.2021.3114866

[48] Jinbin Huang, Aditi Mishra, Bum Chul Kwon, and Chris Bryan. 2023. ConceptExplainer: Interactive Explanation for Deep Neural Networks from a Concept Perspective. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 831–841. doi:10.1109/TVCG.2022.3209384

[49] Jessica Hullman, Robert Kosara, and Heidi Lam. 2017. Finding a Clear Path: Structuring Strategies for Visualization Sequences. *Computer Graphics Forum* 36, 3 (June 2017), 365–375. doi:10.1111/cgf.13194

[50] Sujin Jang, Niklas Elmqvist, and Karthik Ramani. 2016. MotionFlow: Visual Abstraction and Aggregation of Sequential Patterns in Human Motion Tracking Data. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 21–30. doi:10.1109/TVCG.2015.2468292

[51] Alex Kale, Ziyang Guo, Xiao Li Qiao, Jeffrey Heer, and Jessica Hullman. 2024. EVM: Incorporating Model Checking into Exploratory Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 208–218. doi:10.1109/TVCG.2023.3326516

[52] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–28. doi:10.1145/3613904.3642830

[53] Devin Lange. 2024. Vispubs.com: A Visualization Publications Repository. doi:10.31219/osf.io/dg3p2

[54] Ricardo Langner, Tom Horak, and Raimund Dachselt. 2018. VisTiles: Coordinating and Combining Co-located Mobile Devices for Visual Data Exploration. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 626–636. doi:10.1109/TVCG.2017.2744019

[55] Sukwon Lee, Sung-Hee Kim, Ya-Hsin Hung, Heidi Lam, Youn-Ah Kang, and Ji Soo Yi. 2016. How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice's Information Visualization Sensemaking. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 499–508. doi:10.1109/TVCG.2015.2467195

[56] Fritz Lekschas, Benjamin Bach, Peter Kerpedjiev, Nils Gehlenborg, and Hanspeter Pfister. 2018. HiPiler: Visual Exploration of Large Genome Interaction Matrices with Interactive Small Multiples. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 522–531. doi:10.1109/TVCG.2017.2745978

[57] Yanna Lin, Haotian Li, Leni Yang, Aoyu Wu, and Huamin Qu. 2024. InkSight: Leveraging Sketch Interaction for Documenting Chart Findings in Computational Notebooks. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 944–954. doi:10.1109/TVCG.2023.3327170

[58] Maxim Lisnic, Zach Cutler, Marina Kogan, and Alexander Lex. 2025. Visualization Guardrails: Designing Interventions Against Cherry-Picking in Interactive Data Explorers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Tokyo Japan, 1–19. doi:10.1145/3706598.3713385

[59] Qiangqiang Liu, Yukun Ren, Zhihua Zhu, Dai Li, Xiaojuan Ma, and Quan Li. 2023. RankAxis: Towards a Systematic Combination of Projection and Ranking in Multi-Attribute Data Exploration. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 701–711. doi:10.1109/TVCG.2022.3209463

[60] Zhicheng Liu and Jeffrey Heer. 2014. The Effects of Interactive Latency on Exploratory Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 2122–2131. doi:10.1109/TVCG.2014.2346452

[61] Yafeng Lu, Robert Krüger, Dennis Thom, Feng Wang, Steffen Koch, Thomas Ertl, and Ross Maciejewski. 2014. Integrating predictive analytics and social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Paris, France, 193–202. doi:10.1109/VAST.2014.7042495

[62] Joyce Ma, Kwan-Liu Ma, and Jennifer Frazier. 2019. Decoding a Complex Visualization in a Science Museum - An Empirical Study. *IEEE Transactions on Visualization and Computer Graphics* 26 (2019), 1–1. Issue 1. doi:10.1109/TVCG.2019.2934401

[63] Maxqda.com. 2025. MAXQDA. https://www.maxqda.com/

[64] Sharon McDonald, Helen M. Edwards, and Tingting Zhao. 2012. Exploring Think-Alouds in Usability Testing: An International Survey. *IEEE Transactions on Professional Communication* 55, 1 (March 2012), 2–19. doi:10.1109/TPC.2011.2182569

[65] Roberta Mota, Nivan Ferreira, Julio Daniel Silva, Marius Horga, Marcos Lage, Luis Ceferino, Usman Alim, Ehud Sharlin, and Fabio Miranda. 2023. A Comparison of Spatiotemporal Visualizations for 3D Urban Analytics. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 1277–1287. doi:10.1109/TVCG.2022.3209474

[66] Arpit Narechania, Adam Coscia, Emily Wall, and Alex Endert. 2022. Lumos: Increasing Awareness of Analytic Behavior during Visual Data Analysis. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 1009–1018. doi:10.1109/TVCG.2021.3114827

[67] Phong H. Nguyen, Kai Xu, Ashley Wheat, B.L. William Wong, Simon Attfield, and Bob Fields. 2016. SensePath: Understanding the Sensemaking Process Through Analytic Provenance. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (Jan. 2016), 41–50. doi:10.1109/TVCG.2015.2467611

[68] Jakob Nielsen. 2012. Thinking Aloud: The #1 Usability Tool. https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/

[69] Carolina Nobre, Dylan Wootton, Zach Cutler, Lane Harrison, Hanspeter Pfister, and Alexander Lex. 2021. reVISit: Looking Under the Hood of Interactive Visualization Studies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. doi:10.1145/3411764.3445382

[70] Carolina Nobre, Dylan Wootton, Lane Harrison, and Alexander Lex. 2020. Evaluating Multivariate Network Visualization Techniques Using a Validated Design and Crowdsourcing Approach. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–12. doi:10.1145/3313831.3376381

[71] Stan Nowak and Lyn Bartram. 2024. Designing for Ambiguity in Visual Analytics: Lessons from Risk Assessment and Prediction. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 924–933. doi:10.1109/TVCG.2023.3326571

[72] Stan Nowak, Lyn Bartram, and Pascal Haegeli. 2020. Designing for Ambiguity: Visual Analytics in Avalanche Forecasting. In *2020 IEEE Visualization Conference (VIS)*. IEEE, Salt Lake City, 81–85. doi:10.1109/VIS47514.2020.00023

[73] Liam O'Brien and Stephanie Wilson. 2023. Talking about Thinking Aloud: Perspectives from Interactive Think-Aloud Practitioners. *J. User Exper.* 18, 3 (June 2023), 113–132.

[74] Erica L. Olmsted-Hawala, Elizabeth D. Murphy, Sam Hawala, and Kathleen T. Ashenfelter. 2010. Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Atlanta Georgia USA, 2381–2390. doi:10.1145/1753326.1753685

[75] Jorge Piazentin Ono, Sonia Castelo, Roque Lopez, Enrico Bertini, Juliana Freire, and Claudio Silva. 2021. PipelineProfiler: A Visual Analytics Tool for the Exploration of AutoML Pipelines. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 390–400. doi:10.1109/TVCG.2020.3030361

[76] Heungseok Park, Yoonsoo Nam, Ji-Hoon Kim, and Jaegul Choo. 2021. HyperTendril: Visual Analytics for User-Driven Hyperparameter Optimization of Deep Neural Networks. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 1407–1416. doi:10.1109/TVCG.2020.3030380

[77] Annie Preston and Kwan-Liu Ma. 2023. Communicating Uncertainty and Risk in Air Quality Maps. *IEEE Transactions on Visualization and Computer Graphics* 29, 9 (Sept. 2023), 3746–3757. doi:10.1109/TVCG.2022.3171443

[78] Donghao Ren, Tobias Höllerer, and Xiaoru Yuan. 2014. iVisDesigner: Expressive Interactive Design of Information Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 2092–2101. doi:10.1109/TVCG.2014.2346291

[79] Donghao Ren, Bongshin Lee, and Matthew Brehmer. 2019. Charticulator: Interactive Construction of Bespoke Chart Layouts. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 789–799. doi:10.1109/TVCG.2018.2865158

[80] Daniela Retelny, Sébastien Robaszkiewicz, Alexandra To, Walter S. Lasecki, Jay Patel, Negar Rahmati, Tulsee Doshi, Melissa Valentine, and Michael S. Bernstein. 2014. Expert crowdsourcing with flash teams. In *Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST '14)*. Association for Computing Machinery, New York, NY, USA, 75–85. doi:10.1145/2642918.2647409

[81] P. Saraiya, C. North, and K. Duca. 2005. An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 11, 4 (July 2005), 443–456. doi:10.1109/TVCG.2005.53

[82] Yang Shi, Pei Liu, Siji Chen, Mengdi Sun, and Nan Cao. 2023. Supporting Expressive and Faithful Pictorial Visualization Design with Visual Style Transfer. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 236–246. doi:10.1109/TVCG.2022.3209486

[83] Ehsan Jahangirzadeh Soure, Emily Kuang, Mingming Fan, and Jian Zhao. 2022. CoUX: Collaborative Visual Analysis of Think-Aloud Usability Test Videos for Digital Interfaces. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 643–653. doi:10.1109/TVCG.2021.3114822

[84] Fabian Sperrle, Rita Sevastjanova, Rebecca Kehlbeck, and Mennatallah El-Assady. 2019. VIANA: Visual Interactive Annotation of Argumentation. In *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Vancouver, Canada, 11–22. doi:10.1109/VAST47406.2019.8986917

[85] Arjun Srinivasan, Steven M. Drucker, Alex Endert, and John Stasko. 2019. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 672–681. doi:10.1109/TVCG.2018.2865145

[86] Mengdi Sun, Ligan Cai, Weiwei Cui, Yanqiu Wu, Yang Shi, and Nan Cao. 2023. Erato: Cooperative Data Story Editing via Fact Interpolation. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 983–993. doi:10.1109/TVCG.2022.3209428

[87] Katherine E. Thompson, Evelyn P. Rozanski, and Anne R. Haake. 2004. Here, there, anywhere: remote usability testing that works. In *Proceedings of the 5th conference on Information technology education (CITC5 '04)*. Association for Computing Machinery, New York, NY, USA, 132–137. doi:10.1145/1029533.1029567

[88] Wai Tong, Chen Zhu-Tian, Meng Xia, Leo Yu-Ho Lo, Linping Yuan, Benjamin Bach, and Huamin Qu. 2023. Exploring Interactions with Printed Data Visualizations in Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 418–428. doi:10.1109/TVCG.2022.3209386

[89] Jakob Troidl, Simon Warchol, Jinhan Choi, Jordan Matelsky, Nagaraju Dhanyasi, Xueying Wang, Brock Wester, Donglai Wei, Jeff W. Lichtman, Hanspeter Pfister, and Johanna Beyer. 2024. ViMO - Visual Analysis of Neuronal Connectivity Motifs. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 748–758. doi:10.1109/TVCG.2023.3327388

[90] usertesting.com. 2025. UserTesting. https://www.usertesting.com

[91] Jo Vermeulen, Kris Luyten, Elise Van Den Hoven, and Karin Coninx. 2013. Crossing the bridge over Norman's Gulf of Execution: revealing feedforward's true identity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 1931–1940. doi:10.1145/2470654.2466255

[92] Chenglong Wang, John Thompson, and Bongshin Lee. 2024. Data Formulator: AI-Powered Concept-Driven Visualization Authoring. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 1128–1138. doi:10.1109/TVCG.2023.3326585

[93] Xiaoyi Wang, Alexander Eiselmayer, Wendy E. Mackay, Kasper Hornbaek, and Chat Wacharamanotham. 2021. Argus: Interactive a priori Power Analysis. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 432–442. doi:10.1109/TVCG.2020.3028894

[94] Xingbo Wang, Renfei Huang, Zhihua Jin, Tianqing Fang, and Huamin Qu. 2024. CommonsenseVIS: Visualizing and Understanding Commonsense Reasoning Capabilities of Natural Language Models. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 273–283. doi:10.1109/TVCG.2023.3327153

[95] Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. 2020. DataShot: Automatic Generation of Fact Sheets from Tabular Data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 895–905. doi:10.1109/TVCG.2019.2934398

[96] Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Polo Chau. 2021. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 1396–1406. doi:10.1109/TVCG.2020.3030418

[97] Zhen Wen, Wei Zeng, Luoxuan Weng, Yihan Liu, Mingliang Xu, and Wei Chen. 2023. Effects of View Layout on Situated Analytics for Multiple-View Representations in Immersive Visualization. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 440–450. doi:10.1109/TVCG.2022.3209475

[98] Peter Xenopoulos, João Rulff, Luis Gustavo Nonato, Brian Barr, and Claudio Silva. 2023. Calibrate: Interactive Analysis of Probabilistic Model Output. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (Jan. 2023), 853–863. doi:10.1109/TVCG.2022.3209489

[99] Cindy Xiong, Vidya Setlur, Benjamin Bach, Eunyee Koh, Kylie Lin, and Steven Franconeri. 2022. Visual Arrangements of Bar Charts Influence Comparisons in Viewer Takeaways. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 955–965. doi:10.1109/TVCG.2021.3114823

[100] Mehmet Adil Yalçın, Niklas Elmqvist, and Benjamin B. Bederson. 2018. Keshif: Rapid and Expressive Tabular Data Exploration for Novices. *IEEE Transactions on Visualization and Computer Graphics* 24, 8 (Aug. 2018), 2339–2352. doi:10.1109/TVCG.2017.2723393

[101] Fumeng Yang, Mandi Cai, Chloe Mortenson, Hoda Fakhari, Ayse D. Lokmanoglu, Jessica Hullman, Steven Franconeri, Nicholas Diakopoulos, Erik C. Nisbet, and Matthew Kay. 2024. Swaying the Public? Impacts of Election Forecast Visualizations on Emotion, Trust, and Intention in the 2022 U.S. Midterms. *IEEE Transactions on Visualization and Computer Graphics* 30, 1 (Jan. 2024), 23–33. doi:10.1109/TVCG.2023.3327356

[102] Emanuel Zgraggen, Alex Galakatos, Andrew Crotty, Jean-Daniel Fekete, and Tim Kraska. 2017. How Progressive Visualizations Affect Exploratory Analysis. *IEEE Transactions on Visualization and Computer Graphics* 23, 8 (Aug. 2017), 1977–1987. doi:10.1109/TVCG.2016.2607714

[103] Emanuel Zgraggen, Robert Zeleznik, and Steven M. Drucker. 2014. PanoramicData: Data Analysis through Pen & Touch. *IEEE Transactions on Visualization*

and Computer Graphics 20, 12 (Dec. 2014), 2112–2121. doi:10.1109/TVCG.2014.2346293

[104] Xun Zhao, Yanhong Wu, Dik Lun Lee, and Weiwei Cui. 2019. iForest: Interpreting Random Forests via Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 407–416. doi:10.1109/TVCG.2018.2864475

[105] Chen Zhu-Tian, Shuainan Ye, Xiangtong Chu, Haijun Xia, Hui Zhang, Huamin Qu, and Yingcai Wu. 2022. Augmenting Sports Videos with VisCommentator. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 824–834. doi:10.1109/TVCG.2021.3114806