Doing Remote Controlled Studies with Humans: Tales from the COVID Trenches

Rupika Dikkala, Roli Khanna, Caleb Matthews, Jonathan Dodge, Sai Raja, Catherine Hu, Jed Irvine, Zeyad Shureih, Kin-Ho Lam,
Andrew Anderson, Minsuk Kahng, Alan Fern, and Margaret Burnett
Oregon State University
Corvallis, OR 97331, USA
{dikkalar, khannaro, mattheca, dodgej, rajasa, huca, irvine, shureihz, lamki, anderan2, minsuk.kahng, alan.fern, burnett}@oregonstate.edu

***Abstract*—How should empirical researchers conduct controlled, remote "lab" studies in the uncontrolled, noisy conditions of each participant's own home? Volatility in participant home environments, hardware, internet connection, and surrounding distractions takes the "controlled" out of controlled studies. This paper recounts our in-the-trenches mitigations for designing and conducting two complex controlled studies under COVID, in which participants, from home, interactively localized faults in an AI system. The studies with our COVID-era mitigations in 5 categories—Privacy/Security, Data Collection, Control, Technology Issues, Payment—ultimately produced crisp results beyond what we thought possible under such uncontrolled circumstances.**

Keywords—empirical studies in software engineering, COVID-era studies, qualitative empirical studies, quantitative empirical studies

I. Introduction

Doing controlled studies with human participants is often challenging. One challenge lies in the classic difficulty of isolating the independent variable(s)' effects (e.g., software tool A's effectiveness vs. software B's)—without extraneous variable(s) arising to change the results. Keeping such factors at bay is especially challenging if the controlled study cannot be held in a controlled environment. For example, humans in their own homes could be plagued with intermittent Wi-Fi, be interrupted by their children or roommates, be distracted by lawnmowers directly outside their windows, and so on.

However, in March 2020, our research group's ability to do controlled lab studies abruptly disappeared with the advent of COVID. At that time, we were in the midst of designing two in-person lab studies. Study 1 was a qualitative lab study to collect rich, detailed behavior data; Study 2 was a quantitative "control vs. treatment" lab study to compare people's problem-solving successes using certain features vs. not using those features.

The "treatment" was features for a new interactive fault localization process for domain-knowledgeable users of an AI-based system. This process is called the After-Action Review for AI (AAR/AI) process [6]. AAR/AI is a 7-step process. Most pertinent to this paper is the "inner loop" for each AI decision that the user decides to assess. In this loop, the user identifies what happened, describes why they think it happened, and formalizes learning from this decision, by identifying what should change. These AAR/AI steps were in a paper+online prototype for Study 1, and fully integrated into an online prototype for Study 2.

Study 1's central research question asked what *diverse behaviors* domain-knowledgeable users exhibit with AAR/AI in attempting to localize an AI's reasoning faults. Adding further to the diversity we encouraged, we needed participants to use their *own* standards to assess when the AI's reasoning was faulty. To allow these diverse behaviors and standards, we needed a very flexible prototype/environment, so that participants could go about fault localization however they pleased.

Our original Study 1 plan envisioned a conference room with a laptop and a large printout spread out on a table showing a detailed view of the explanation (Figure 1). We aimed to run about 15 participants, one at a time. The participant would move around the physical space to focus on parts of the diagram they wanted, marking it up when they found something problematic, interactively replaying/rewinding the game via laptop as desired, and answer the AAR/AI questions on the clipboard about faulty reasoning they had found. However, this set-up became impossible with the COVID lockdown.

Informed by Study 1's results, Study 2's main research question was quantitative, asking whether domain-knowledgeable users would be more successful localizing faults with the AAR/AI process than without it—again, as per their own standards.

As a statistical experiment, Study 2 needed about 60 participants to use our new system (half with AAR/AI, half without), without extraneous sources of variation interfering with statistical results. We planned to integrate the components of Figure 1 into a software implementation, with the idea that half the participants would use all three of Figure 1's components and half would use only the left and middle components. In a lab setting, we would have set participants up on identical computer systems, and continually monitored them to prevent interruptions, cell-phone calls, conversations with one another, etc.

When COVID prevented these studies from being run in controlled lab settings, we devised ways to conduct Study 1 and Study 2 remotely from participants' homes. This paper recounts the ways we found to mitigate our remote studies' risks and challenges, and touches upon one form of validation: whether these mitigations paid off.

II. Background

Doing controlled empirical work with human participants not co-located is not new. For example, one form of data collection often conducted with remote participants is one-on-

2574-1837/21/$31.00 ©2021 IEEE
DOI 10.1109/CHASE52884.2021.00022

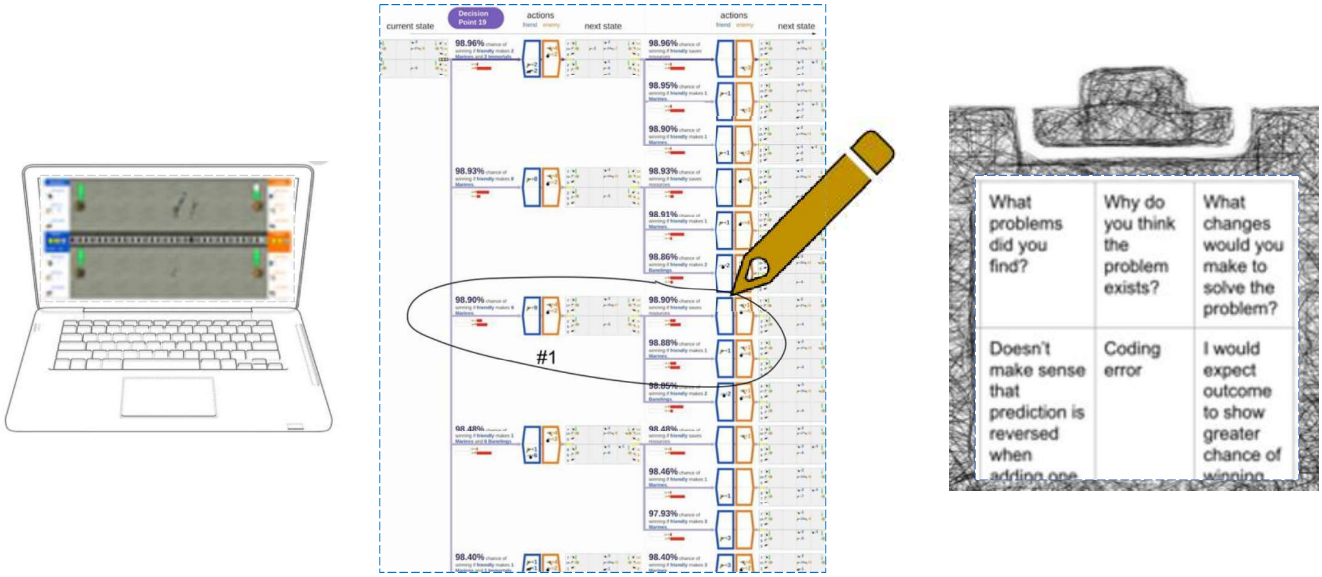

Figure 1: Study 1 plan before COVID lockdown. A conference room with (Left:) laptop for interacting with parts of the game; (Center:) a huge paper diagram explaining the AI's logic (only a portion is shown here), spread out on a table for participant to mark up; (Right:) clipboard with AAR/AI questions for participant to answer.

one on-camera interviews. On-line surveys are another way of going about data collection from remote participants. Both interviews and surveys are very different from controlled lab studies, but surveys have in common with our studies a need for quality control devices [5]. One such device is attention checks—questions inserted only to gauge whether the participant is paying attention, rather than just quickly marking things (e.g., [1, 2, 3]). However, neither surveys nor interviews can produce detailed behavior data on a complex system.

Conceptually closer to controlled lab studies are remote crowdsourced studies using online labor markets like Amazon's Mechanical Turk (mTurk). In such markets, empiricists can post web-based tasks, and participants can sign up to complete them for pay. Most mTurk participants are young, educated males [8], a potential problem given the diversity of participants we sought for Study 1. Also critical for our studies is mTurks' tendency to select tasks they are already familiar with. With this tendency, plus the fact that only 10% of the mTurks complete almost half the available tasks, mTurk respondents might have been overly familiar with our prior studies [8, 9]. mTurks are also risk-averse about their work being rejected, so they select tasks with well-defined goals, and frequently share information with each other about how to complete them successfully [7]—however, our tasks were (deliberately) poorly defined, to allow participants latitude in deciding what constituted as faulty reasoning. For all these reasons, we could not use online labor markets for our studies. However, the extensive use of attention checks in mTurk studies [4] influenced the way we inserted attention checks into our remote studies.

III. Tales from the COVID-19 Trenches

We ran Study 1 as a one-on-one think-aloud and Study 2 in sessions, with one to seven participants per session (average was two). We initially thought the main risks were going to lie in Controls, but five categories of risks arose: Privacy/Security, Technology Issues, Controls, Data Collection, and Payment.

Privacy/security issues faced both researchers and participants. Some of these arose because we did not have time to start over with a new IRB application—but these issues would unlikely pass most IRBs anyway. For example, our IRB protocol did not cover (and therefore would not allow) capturing video of participants' home environments. Even screen-sharing could be problematic, as the participant might inadvertently share information not included in our IRB protocol (e.g., their email) or material offensive to our researchers. Further, installing our technology on their home computers (1) was not covered by our IRB protocol; and (2) could be a participation barrier for people uncomfortable installing third-party software, potentially limiting our goal of attracting diverse viewpoints. Also, providing participants remote access to our team's personal computers posed a security risk to our team.

We mitigated these challenges in Study 1 by running our programs on *researchers'* computers, i.e., sharing *our* screens with participants; by Study 2, we had implemented an entirely web-based prototype. During screen share, we safeguarded researchers' privacy by sharing only the relevant window, not the entire desktop, and by hiding the bookmark bar in their internet browser. Finally, we collected only audio (not video) for both data anonymization and personal privacy reasons, with both researchers and participants switching off their cameras before recording began.

These mitigations worked well, but after we finished data collection, we realized we had missed an important privacy risk in Study 1. Our IRB approvals require data to be stored anonymously, e.g., by a participant ID for that study, not by name. However, we used a Google-based software platform (discussed in the Technology subsection), and later discovered

that its stored version histories included participant names (if the participant happened to be logged into their personal Gmail account during the study). This personally identifiable information became part of our data collection, visible to all researchers during analysis. In future COVID studies, if we use a Google-based platform, we plan to mitigate this by creating temporary Gmail accounts "named" by participant ID (e.g. P234@gmail.com), to ensure that no personal accounts are inadvertently recorded.

Technology issues arose for both software and hardware. The first software issue was the difficulty of designing a software set-up that would provide the kind of flexibility we had envisioned for Study 1 (Figure 1). Screens are much smaller than conference tables, and the classic uses of progressive disclosure (showing just an overview, then allowing people to expand details) would have forced users into a prescribed subset of problem-solving behaviors—thereby subverting Study 1's research question. We ultimately decided to provide them with *too much* information, then allowing them to freely pan and zoom, adding mark-ups as they desired.

However, finding a software platform with this functionality was problematic. We spent days experimenting with several products that ultimately did not suffice, with problems like automatically down-sampling the very high-resolution images our explanations needed, or delaying feedback on annotations enough to discourage its use. Finally we settled on Google-Draw with preloaded explanations (Figure 1 (Center), without the pencil) and a table of AAR/AI answer space (Figure 1 (Right)), which supported the functionality we needed for Study 1.

The one-on-one think-aloud set-up made Study 1 easier to run than Study 2. For example, although we could not see the participants, we could pseudo-observe them in Study 1 by watching their live changes to the shared Google-Draw document. Also in Study 1, the researcher could replay parts of the game on demand if a thinking-aloud participant requested it. Since we were recording the researcher's screen (which showed the video and the Google-Draw activity live) along with participant audio, the full context was recorded and available for later analysis. However, Study 2 had multiple participants at once and, because we did not want them to see/hear each others' work, Study 2 needed a different approach.

Further, even if privacy concerns were not an issue, we could not have asked Study 2 participants to share their screens—Zoom does not support multiple synchronous screen-sharing. Thus, we created a web-based platform that allowed participants to (1) pinpoint faulty reasoning in the explanation, (2) replay certain portions of the game (dependent on how far they were in the task), and (3) answer the AAR/AI questions. It also (4) logged everything participants were doing and displayed a live dashboard to the researcher.

Hardware issues were also problematic, which we cover more in the "Controls" subsection. Suffice to say here that, because "anything that can go wrong will" (Murphy's Law), we set up contingency plans for choppy/lost internet connections, inaudible directions, computers running out of power, and broken chat links—and almost all of these actually occurred when we ran the studies.

Controls are strong with in-person lab studies, since everyone is using an identical system in the same environment, under constant researcher oversight. Without this uniformity, we risked uncontrolled factors affecting our results. For example, we had no control over participants' *technical* environment, such as their monitor size, internet quality/dependability, operating system, or computer memory. We also could not control factors that might *distract* the participant, such as clamors for attention from children, phone notifications, or multitasking. Any of these factors risked influencing the quality of the data.

We mitigated Control risks using three strategies. First, Study 1 sessions were one participant at a time and Study 2 sessions averaged two participants at a time, so interacting conversationally with participants was natural. We used this fact in a few ways to insert attention checks throughout the study, both to check attention and to confirm absence of technical difficulties: For example, the tutorial was interactive—the researcher checked in with participants along the way, visually led them to a feature of interest with a large mouse cursor, and the tutorial included "pop quiz" questions to ensure they were attentive. Since we could not discern where they were looking, our "quizzes" asked questions like: "how many <objects> do you see on the left side of the screen?" (See Supplemental Document: https://doi.org/10.6084/m9.figshare.13696132.)

We also used this conversational strategy to ask them to interrupt us if internet connection became problematic, or if they could not see the items we pointed at. Then in the main task, we monitored their level of attention throughout the session: in Study 1 reminding them to think aloud if they fell silent, and in Study 2 watching the dashboard for signs of participant inactivity. However, this Study 2 device tasked the researcher with the intense effort of keeping up with all the participants on the dashboard. This was in addition to communicating special instructions to each participant via private Zoom messages about individual log-ins, payment codes, or responses to individual questions.

Second, we reduced distractions by minimizing context switches (e.g., moving from Zoom screenshare to the web browser), which could have been particularly disruptive on small screens. (We had to do a surprising amount of planning to achieve this goal.)

Third, we elected not to provide a "live" game system, for two reasons. The first was to avoid the game behaving differently on unforeseen/untested computer environments. The web platforms described in the Technology subsection solved this by allowing participants to use the browser they were comfortable with. The second was to prevent the participant from interacting in ways that would take them "off the rails" of our study sequence. Instead, we relied on game replays, and pre-prepared browser-based explanations of the portions of the games they saw, for them to mark up and comment upon while we watched.

Data collection mechanisms were challenged by the no-videos-no-participant-screen-capture mitigations of the Privacy/Security risks.

One issue was with the audio data, particularly key in Study 1. Qualitative data from think-alouds without facial expressions and without continuous screen capture of participants' mouse movements introduced both loss of information and extra effort. We had hoped for a silver lining via Zoom's automatic transcription, which we expected to eliminate much of the manual work associated with audio recording. However, our optimism was misplaced. Some transcripts had processing issues, requiring waiting for tech support follow-up. Furthermore, machine translation was often low quality, and we had to entirely redo them. As a result, no time savings materialized.

Another data collection problem in both studies was synchronizing data from multiple platforms. For example, Study 1 used one platform with clear, precise data storage (Google-Draw) for participant markups, and another with low resolution (Zoom recordings) for interaction with the researcher and the shared display. We also used transcribed audio for ease of qualitative coding. These three separate, unsynchronized data repositories required frequent window switching during analysis to align participants' actions with their verbalizations. In Study 2, the custom-built web-based system alongside a separate click-log replaced Google-Draw, but made analysis much less intuitive. As in Study 1, synchronizing Study 2's distinct repositories of participants' work, without the benefit of a high-quality unifying video, was a painstaking process.

Payment: Even paying participants was challenging. During in-person studies, we usually pay cash and provide a duplicated signed receipt, which produces an audit trail for all parties. In contrast, moving money was hard under COVID, and doing so with an audit trail was harder. We chose not to do a cash e-transfer participants because (1) it would require banking with particular banks/digital payment services, potentially introducing a sampling bias; and (2) some lacked audit trails.

Ultimately, we elected to transfer money from researchers to participants using Amazon eGift Cards. The researcher stayed on the Zoom call until the participant acknowledged the gift card receipt, both verbally and with an email reply. Amazon also notified the researchers when we purchased these gift cards. We needed these receipts from Amazon to document our expenditures for university accounting, but this added the burden of manual bookkeeping.

IV. Epilog and Concluding Remarks

None of the above mitigations came cheaply—we iteratively piloted and adjusted for weeks before conducting the studies with real participants.

Study 1 (qualitative) was first, with 17 participants. As we had hoped, it did yield a diversity of problem-solving approaches, due mostly to the fact that over 1/3 of the participants were women. These women added a diverse set of backgrounds and approaches to the fairly homogeneous batch the men had brought to the study.

Despite Study 1's success, we did not have high hopes for Study 2, because the setup for Study 2 was more complex than for Study 1, and the need for controls with statistical studies is very high. Still, we ran it, with 65 participants. To our surprise, Study 2 produced results with a strong signal-to-experimental-noise ratio. Due to its surprisingly clean data, Study 2 produced a collection of statistically significant results with moderate to strong effect sizes, such as the result snippet depicted in Figure 2.

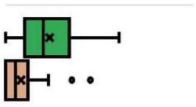

Figure 2: One snippet of Study 2's statistically significant AAR/AI vs. non-AAR/AI differences. (Green (top) is AAR/AI, brown (bottom) is non-AAR/AI.)

As these tales from the COVID-19 trenches show, running a controlled study with remote human participants—one that actually *is* reasonably well-controlled—is not easy, and is a lot of work. However, it can be achieved with careful planning. We hope that, by sharing the challenges we ran into and our remedies, other researchers can gain from our experiences and be able to run their own controlled studies with human participants, even before the pandemic loosens its grip.

Acknowledgments

This work was supported in part by DARPA N66001-17-2-4030, NSF 1901031, and NSF 2042324.

References

[1] James D. Abbey and Margaret G. Meloy. "Attention by design: Using attention checks to detect inattentive respondents and improve data quality." Journal of Operations Management, vol. 53, 2017, pp. 63-70.

[2] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. "Separating the shirkers from the workers? Making sure respondents pay attention on self - administered surveys." American Journal of Political Science, vol. 58, no. 3, 2014, pp. 739-753.

[3] Tobias Gummer, Joss Roßmann, and Henning Silber. "Using instructed response items as attention checks in web surveys: Properties and implementation." Sociological Methods & Research vol. 50, no. 1, 2021, pp. 238-264.

[4] David J. Hauser and Norbert Schwarz. "Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants." Behavior Research Methods vol. 48, no. 1, 2016, pp. 400-407.

[5] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. "Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality". ACM CHI Conference on Human Factors in Computing Systems, May 2019, (12 pages.) https://doi.org/10.1145/3290605.3300316

[6] Theresa Mai, Roli Khanna, Jonathan Dodge, Jed Irvine, Kin-Ho Lam, Zhengxian Lin, Nicholas Kiddle, Evan Newman, Sai Raja, Caleb Matthews, Christopher Perdriau, Margaret Burnett, & Alan Fern, "Keeping it 'organized and logical': After-Action Review for AI (AAR/AI)", ACM Int. Conf. Intelligent User Interfaces, March 2020, pp. 465-476.

[7] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. "Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk." ACM CHI Conference on Human Factors in Computing Systems, pp. 2271-2282. 2016.

[8] Gabriele Paolacci and Jesse Chandler. "Inside the Turk: Understanding Mechanical Turk as a participant pool." Current Directions in Psychological Science, vol. 23, no. 3, 2014, pp. 184-188.

[9] Sang Eun Woo, Melissa Keith, and Meghan A. Thornton. "Amazon Mechanical Turk for industrial and organizational psychology: Advantages, challenges, and practical recommendations." Industrial and Organizational Psychology, vol. 8, no. 2, 2015, pp. 171-179.