Galileo: Citizen-led Experimentation Using a Social Computing System

Vineet Pandey
Harvard University, Cambridge, Massachusetts, United States

Tushar Koul
UC San Diego, La Jolla, California, United States

Chen Yang
UC San Diego, La Jolla, California, United States

Daniel Mcdonald
UC San Diego, La Jolla, California, United States

Mad Price Ball
Open Humans Foundation, Sanford, North Carolina, United States

Bastian Greshake Tzovaras
Université de Paris, INSERM U1284, Center for Research and Interdisciplinarity (CRI), Paris, France

Rob Knight
UC San Diego, La Jolla, California, United States

Scott Klemmer
UC San Diego, La Jolla, California, United States

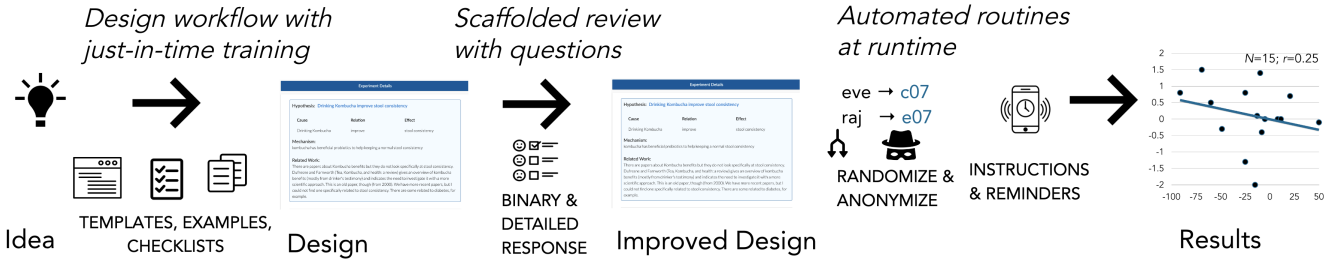

Figure 1: Galileo enables people to design and run experiments to test their intuitions. Experiment creators can invite others to review and participate in the experiment. Participants from around the world join experiments, follow instructions, and provide data in response to automated data collection reminders.

ABSTRACT

People have scientific questions and folk theories; yet most lack the expertise to investigate them. How might people transform their questions into experiments that inform both science and their lives? This paper demonstrates how online volunteers can collaboratively design and run experiments using a novel social computing system. The Galileo system provides procedural support using three techniques: 1) experimental design workflow that provides just-in-time training; 2) review workflow with scaffolded questions; and 3) automated routines for data collection. We present two empirical investigations: a study and a field deployment with online volunteers across 16 and 8 countries respectively. People generated structurally-sound experiments on personally meaningful topics; three communities ran a week-long experiment each. We identify two key challenges for citizen-led experimentation—supporting different expertise levels and providing recruitment guidance—and provide specific suggestions from the social computing literature. Our results highlight the promise and challenges of citizen-led knowledge work like experimentation.

CHI '21, May 08–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445668

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); Collaborative and social computing systems and tools.

KEYWORDS

Social computing systems, citizen science, crowdsourcing, experimentation

ACM Reference Format:
Vineet Pandey, Tushar Koul, Chen Yang, Daniel Mcdonald, Mad Price Ball, Bastian Greshake Tzovaras, Rob Knight, and Scott Klemmer. 2021. Galileo: Citizen-led Experimentation Using a Social Computing System. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445668

1 INTRODUCTION

Scientific experimentation features technical requirements and contextual choices that are inscrutable for a lay individual yet necessary

for success [45]. While professional scientists and commercial ventures run experiments every day, with notable exceptions [12, 42], empirical papers from non-professionals are vanishingly rare. People have questions about their health but lack the expertise and resources to scientifically investigate them. Broadening the pool of experimenters could help people investigate their curiosities, develop solutions to improve health and performance, and assist institutional researchers.

The main contribution of this paper is a demonstration that online volunteers can collaboratively design and run experiments. This paper achieves this goal with the Galileo social computing system that instantiates procedural support using three techniques: experimental design workflow that provides just-in-time training, review with scaffolded questions, and automated routines for data collection (Figure 1).

Two empirical investigations tested Galileo's approach. First, a deployment across 16 countries found that people generated structurally-sound experiments on personally meaningful topics. Second, in a field deployment, online users from three communities—kombucha, Open Humans, and beer—across 8 countries demonstrated that people designed, iterated on, and ran week-long experiments.

2 RELATED WORK

This paper draws on prior work in designing systems for novice-led inquiry. Citizen science and crowdsourcing are the domains closest to this work.

2.1 Citizen Scientists: From Collectors to Experimenters

Citizen science efforts span counting bird species, identifying galaxies, editing protein structures, and creating novel hypotheses [12, 51, 62]. One reason for citizen science's success is that different people provide different expertise that can vet claims and fix mistakes [30]. A humbling example of the power of fresh eyes: volunteer citizen scientists identified an entirely new class of galaxies ("green pea" galaxies) from Galaxy zoo images; experts had dismissed these images as apparatus error [5]. This volunteer-led discovery demonstrates the need for fostering independent perspectives while simultaneously cultivating sufficient knowledge for meaningful domain contributions.

Efforts to expand participation in scientific research are bearing fruit: Lab in the Wild recruits anyone with an internet connection for behavioral studies [54]; All of Us aims to recruit one million Americans from all strata of society (allofus.nih.gov). Distributed data contributions from people around the world—browsing online [13], using activity trackers, and joining scientific projects—have enabled valuable insights on topics including obesity [2], aesthetic preferences [53], sleep [20], and the human microbiome [47]. Our work draws on this general idea of people sharing data with institutional experts [27]; it adds ways to include people's complementary insights and cognitive surplus for citizen-led scientific work [4].

A number of health and behavioral research projects enlist citizens as helpers (e.g., HabitLab [36]). It remains rare for citizens to design experiments. CivilServant enables online communities' moderators to test policy ideas; moderators share these ideas with researchers who transform them to study designs [46]. Through the PatientsLikeMe website (patientslikeme.com), citizens and scientists created a study investigating whether consuming lithium alleviated ALS symptoms [58]. While an initial scientific study had provided positive benefits, both this citizen science study and a subsequent university study did not find benefits. Closest to our research, Tummy Trials asked participants to generate health questions, introducing a protocol for self-experimentation combining ideation and self-tracking [31].

This paper provides a general workflow for people to transform their intuition to an experimental design; our work focuses on controlled experiments as opposed to self-tracking or informal iteration. Our work is distinct from prior citizen science platforms in three interlocking ways. 1) By reporting local/personal facts (e.g. Audubon count, 23andme), citizens typically help answer experts' questions. In our work, people answer their own questions. 2) Current citizen science research does not systematically support causal theory generation & evaluation. Our work provides one way: a platform for randomized experiments. 3) Converting citizens' intuitions to study designs requires experts' involvement [46, 58]. Our work supports testing hypotheses without drawing on expert time. Given the surge of public interest in clinical trials, science, and knowledge creation, we believe these contributions will be broadly useful to the citizen science community.

2.2 Supporting Novice-Led Inquiry

Lived experience, a tight feedback loop, and strong personal motivation can yield different and sometimes better ideas than experts [28, 50]. Prior work has explored collaborative hypothesis generation and testing on pre-existing data sets [43, 61]. Galileo offers a complementary contribution: enabling citizens to generate data on topics of personal interest.

One way to make complex activities manageable is to divide them into distinct phases. Touchstone demonstrates the power of a semi-automated workflow integrating experiment design, testing, and analysis [44]. Crowdsourcing has similarly innovated by dividing larger activities into microtasks; algorithms specify the division, dependency, and agglomeration activities while workers perform small tasks supported by task-specific guidelines [38]. From these systems, our work draws the idea of dividing experimentation into multiple tasks—some self-sourced, others crowd-sourced; and provides just-in-time support. Furthermore, Galileo– the system presented in this paper–automatically manages four activities at runtime to reduce bias and experimenter workload.

Carefully-constructed interfaces can aid novices with task-specific expertise to solve problems that only experts previously could. Foldit introduced 3D game for specifying low-energy protein structures via direct manipulation [12]. Making a challenge visually salient is an effective way to on-board novices. Complementing this visual approach, crowdsourcing systems have successfully leveraged scaffolds and interactive guidance. For example, Cicero and CrowdLayout provide guidelines and rules to help workers reason about their choices and simplify complex activities like designing network layouts [8, 55]. Others, like CrowdSCIM and Crowdclass, scaffold pre-task interventions that provide procedural expertise for historical and scientific analysis [40, 57]. More generally, designing

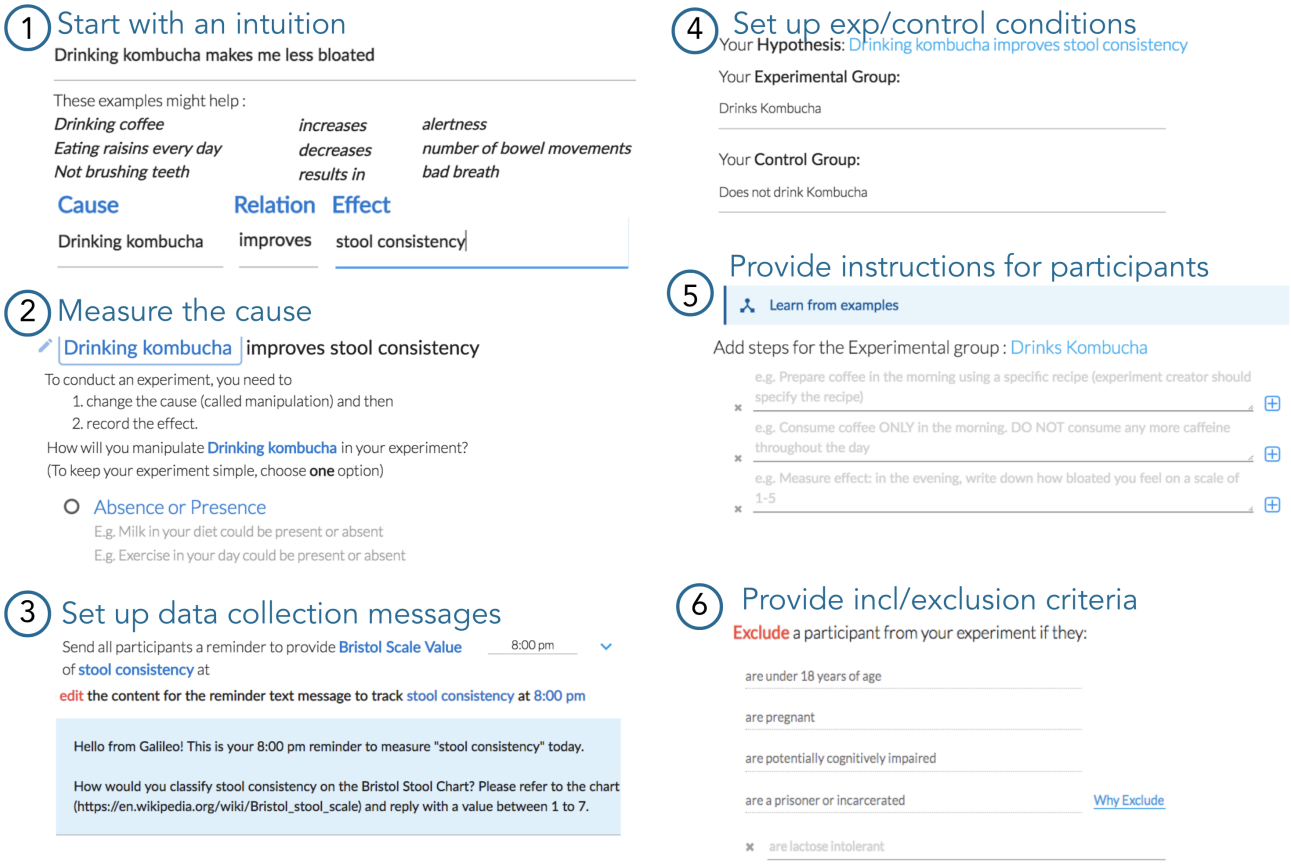

Figure 2: Galileo's design module helps people transform intuitions into experiment designs. It walks people through 1) converting an intuition to a hypothesis, 2,3) providing ways to manipulate/measure cause and effect, 4-5) specifying control and experimental conditions, and 6) providing inclusion/exclusion criteria.

complex tasks for crowdsourcing benefits from ideas in instructional design. For instance, providing step-by-step instruction and showing helpful supportive information help learners acquire complex cognitive skills [33]. Galileo introduces task-embedded support for people with little-to-no mental model of the knowledge domain. Like the Shepherd writing system [15], Galileo provides just-in-time support. There are two key differences: 1) Galileo begins earlier by scaffolding the entire creation process, not just the post-draft feedback stage, and 2) while Shepherd drew on expert time, Galileo does not– the knowledge is implemented in the software itself. Additionally, this work builds on personal informatics research that focuses on an iterative model of experiment design [14]. Our work introduces support for iterations without drawing on expert time: people design, ask for input, & then learn more via pilots.

3 THE GALILEO EXPERIMENTATION PLATFORM

Galileo introduces a system for end users to design experiments, get them reviewed, and run them with interested participants. It provides procedural support for these steps, an online collaboration platform, and automated data collection and reminders (Figure 1).

Despite a predetermined goal and a formalized process, experimentation requires making contextually-appropriate decisions [45]. Good experiment design is inherently user centered; designers need awareness of others' interpretation of their ideas and asks. Providing feedback on experiment designs requires knowing the success criteria and how to help improve. Finally, successfully running an experiment requires managing multiple processes such as random assignment, anonymizing participant details, and sending instructions and reminders for data collection.

3.1 Design-Review-Run: From Intuitions to Investigations

Galileo requires three roles for each experiment: designer, reviewer, and participant. Galileo offers procedural support for each: 1) a design workflow provides just-in-time training, 2) review with scaffolded questions, and 3) automated routines for runtime activities like data collection.

3.1.1 Design an Experiment from an Intuition. Galileo's design workflow helps people sharpen their hypotheses (Figure 2). Examples illustrate possible choices and how they relate; templates

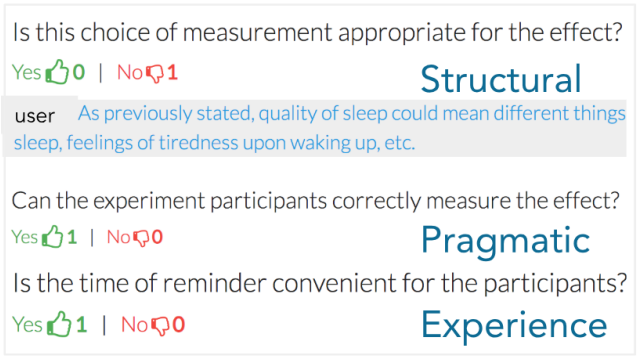

Figure 3: Reviewers walk through an experiment providing binary rubric assessments. A No response prompts reviewers to provide concerns and suggestions.

provide structure; and embedded videos explicate technical issues. Such procedural support can improve on-task performance [51]. A final self-review step provides an overview of the experiment. The design workflow does not mandate double-blindness or the use of placebo; designers can choose to specify these details.

3.1.2 Review the Design via Feedback from Others. Galileo requires at least two reviews before an experiment can be run. The designer invites reviewers: an online community member, a teacher, or anyone else who can provide useful feedback. Upon receiving reviews, the designer edits their experiment to address any issues. For research purposes, Galileo logs version changes. Reviewers provide both binary assessment and written responses to specific questions (Figure 3). These questions cover structure (*e.g.*, accounting for confounds), pragmatics (*e.g.*, measuring the real-world cause/effect), and participant experience (*e.g.*, data reminder time). Reviewers are ineligible to be participants in the same experiment. Similarly, creators may not review their own experiment.

3.1.3 Run an Experiment using Procedural Support. To launch an experiment, its designer shares a unique URL with potential participants. Galileo automatically manages four activities to reduce bias and workload:

1. Randomized placement of people into conditions [45];
2. Maintain a per-experiment participant map ([usernames] → [exp_id]) for anonymity;
3. Collect and clean data (sending data collection messages and reminders at time-zone appropriate times, parsing the responses, updating participant & experimenter views);
4. Prompt experimenters to perform tasks when conditions are met (e.g., setting the start date when enough participants have joined or reminding participants with missing data).

Participation comprises following instructions (e.g., drink kombucha) and providing self-report responses to platform queries (Figure 4). Self-reports provide the primary data collection mechanism. Participants can optionally answer follow-up questions that capture contextual insights (e.g. changes in daily lifestyle due to travel). Galileo presents participant data to experimenters using

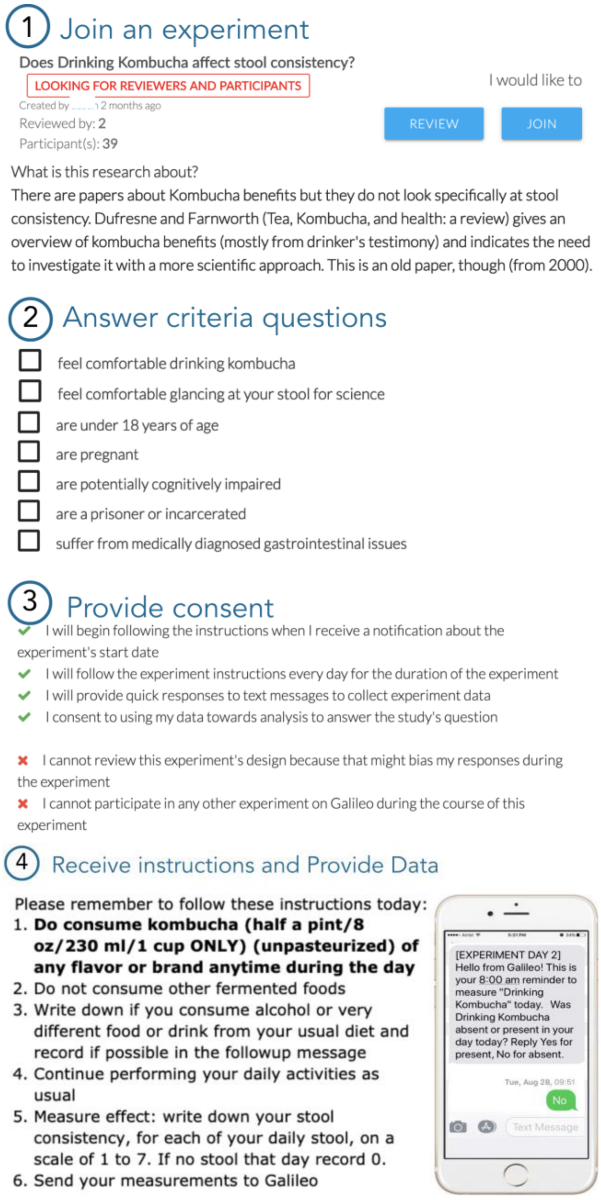

Figure 4: 1) Participants can view a list of experiments. When they elect to join one, they 2) answer inclusion/exclusion criteria, 3) consent to following the provided steps, and 4) receive instructions. Participants receive daily, condition-specific requests, and respond with data and/or clarifying questions.

participant ID rather than real name or username. When an experiment ends, Galileo sends a summary of results to participants. Participants can anonymously discuss experiments at the end, so the experimenter and other users on the platform can learn from their feedback. The experimenter's dashboard provides a summary of their experiment's progress and supports lightweight tasks to improve the quality of data collected. The dashboard lists tasks:

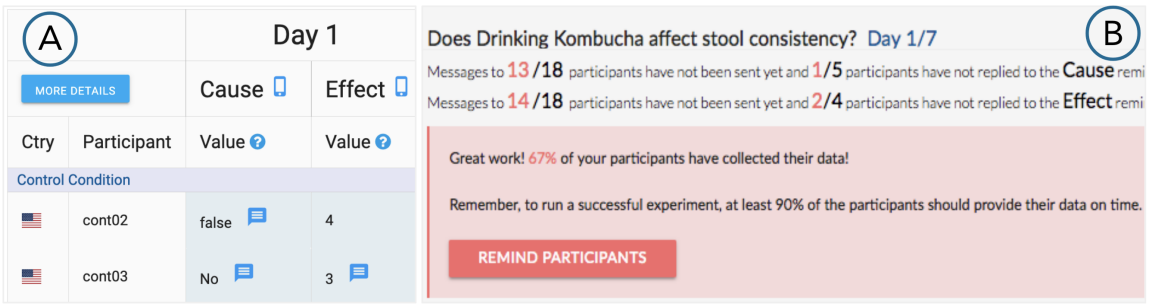

Figure 5: Galileo takes care of many experimenter responsibilities such as random placement of people, sending instructions and reminders, and cleaning and displaying data in both participant and experimenter dashboard. The dashboard enables experimenters to A) remind those with missing data; and B) see participants' data; and clarify questions raised by participants.

answer clarifying questions, remind/thank participants, or look at trends in data (Figure 5). Experiments have a minimum participation count; there's no upper limit to the number of participants. People who sign up after a cohort begins are waitlisted.

The Galileo web application uses the Meteor (meteor.com) framework for synchronization, Jade for the front end (jade-lang.com), and Materialize for styling (materializecss.com). The current Galileo implementation supports email, SMS with text message gateway Twilio (twilio.com), and WhatsApp. Galileo logs responses to a MongoDB database.

3.2 Designing the Platform Over Multiple Iterations

80 people designed, ran, or participated in experiments before formal evaluation of the platform. The system design evolved over a year of weekly in-person user-centered studies with lead users from different communities including kombucha and self-tracking enthusiasts. The pilot study gathered feedback on the usefulness of the interface items and resources. Students in an undergraduate Psychology class (Introduction to Research Methods) also used Galileo in a 90-minute classroom deployment to rapidly design and review each other's' experiments and receive feedback. We provide three examples of how pilot studies informed Galileo's design:

1. *Embedded written training over videos*: Early versions provided short, online lecture videos as the learning materials. Most users did not watch them end-to-end to extract the step-relevant insight(s). In response, each step's content now offers written examples, which are easier to skim and refer back to. Additionally, the users can peruse the lecture videos for additional information.
2. *Supporting successful reviews*: For the review interface, early versions only requested binary Yes/No responses similar to popular commercial and research crowdsourcing platforms [37]; both experiment designer and reviewers found this to be unsatisfactory. Galileo now provides a prompt for actionable feedback whenever the reviewer selects "No" to any question. Additionally, early Galileo users sometimes made poor choices, like listing effects that are difficult to measure. To help guide people, Galileo now presents a short checklist for verifying the choices made in each section. This self-review provides lightweight, just-in-time support.
3. *Introducing dashboards for experimenters and participants*: Pilot users ran six trial experiments. The idea of a run-time dashboard (Figure 5A) came from observing experimenter's difficulty tracking participants' data and sending reminders to those who hadn't added their data. Additionally, participants struggled with making suitable preparations for a week of experimentation (e.g. buying sufficient kombucha). The system now prompts experimenters to explicitly add preparation instructions that are sent to participants 2 days before the experiment begins; these instructions are also shown on the participant dashboard.

3.3 Designing the Experiment Design Workflow

We first asked people to follow experiment design steps in a more "standard" order (identify hypothesis & variables, create conditions, add participation steps, then add data collection logistics etc.) but realized two concerns: 1) the workflow was too long; 2) people needed to recall work from previous steps. Such problems are well-known in instructional design. Complex activities overwhelm working memory because of their many interrelated pieces [18]. Recalling work from previous steps & frequent context-switching are especially taxing. Experts mitigate memory demands by integrating multiple elements into conceptual chunks [7]. Using these ideas, the design interface clusters steps related to one variable; e.g. Steps 2 & 3 in Figure 2 streamline tasks related to the independent (and dependent) variable before moving to set up conditions.

4 STUDY 1: DESIGN & REVIEW EXPERIMENTS ONLINE

A deployment investigated the quality and nature of experiments: do people create experiment designs that a) are structurally-sound, and b) demonstrate insights from lived experiences? Further, do people provide useful feedback on experiment designs?

4.1 Method

Participants used Galileo to design their experiments and review others' designs. Galileo's landing page described the importance of

Table 1: Rubric for design-quality criteria for Structure (13 points), Content, and Novelty

Structure: 13 points
Hypothesis: 3 points
Is the cause/relation/effect specific? (1pt each)
Measurement 2 points
Are the cause/effect manipulated/measured correctly? (1pt each)
Conditions: 3 points
Are the control and experimental conditions appropriate? 2pts
Do the conditions differ in manipulating the cause? 1pt
Steps: 2 points
Are experimental steps clear for control/experimental conditions?
Criteria: 2 points
Are the exclusion criteria correct and complete?
Are the inclusion criteria correct?
Can the overall experiment be run as is? 1 point
Content
Personal? Did the hypothesis draw from lived experience?
Popular? Is the world already curious about this hypothesis (e.g. Are there online discussions about this hypothesis?)
Insightful? Does the hypothesis link to existing science?
Novelty
Is there a chance the world will learn something: absence of published research for this question?

experimentation to create scientific knowledge and how citizens can contribute towards making discoveries. Upon logging in, participants could design an experiment (see Figure 2), review existing experiments (see Figure 3), or join an experiment (see Figure 4).

4.2 Recruitment

Participants were recruited via online publicity. One recruitment focus was people curious about the microbiome because it is a domain where lived experience may inspire intuitions, and the science is nascent [47]. Galileo was promoted on the American Gut's and their collaborators' Facebook and Twitter pages. Galileo was added as a project on Open Humans (openhumans.org), posted on multiple subreddits pertaining to health and lifestyle, and introduced as an optional activity in assignments on the Gut Check Coursera MOOC [34]. Participation was voluntary and unpaid.

4.3 Measures

Measures comprised experimental designs' structure, content, and novelty (Table 1) and reviews' usefulness. The rubric was developed iteratively by the lead author & an instructor (an expert in research methods instruction) during an early pilot in a class. The rubric checks whether people create correct specific elements of an experiment. The final rubric worked well for rating students' experiments; we reused it for this study. Two raters with experiment design training independently rated 5 experimental designs, then discussed them to form a shared view of assessment. Next, each independently rated all experiments. The final score is the mean of the independent ratings. Moderate reliability was found between the two raters' measurements [35]; m(ICC) = .62, 95% CI [.45, .75], ($F(64,64)= 4.33$, $p<.001$.

Structure measures whether the design is correct and includes appropriate components. *Content* measures the subject matter of the idea driving the experiment design; it was rated as personal focus, popularity, and insightfulness of the hypothesis. *Novelty* was assessed as the potential to create new knowledge and operationalized as the lack of research papers about the specific hypothesis. Raters were instructed to assign points for a component (say hypothesis) if the experiment provided appropriate details about it. For example, the hypothesis "*Text message reminder increases consumption of recovery snack*" was rated to have a specific cause, a specific effect, and a clear relation between the two, while "*Eating too much energy causes disturb [sic] sleep cycle*" did not have a clear cause or effect. "*Ingesting non-local food results in poor evacuation of fecal matter*" was rated as novel because no published research addresses this (as per first 100 Google Scholar search results). Broad or vague hypotheses or those related to well-studied topics were not deemed novel (e.g. "*Going to college increases grades*").

54 users from 16 countries created 66 complete experiment designs (*Mdn*=27 minutes). 37 users provided 205 descriptive review comments. Latest versions of complete experiment designs were scored as described above; incomplete experiments and older versions were removed from analysis.

4.4 Study 1 Results

4.4.1 People Designed Structurally-Sound Experiments and Drew from Personal Intuitions. The mean score for the experiment was 10.3/13. 75% of participants earned full scores on 8 of 13 measures. 38% of experiment designs came for people's lived experiences; e.g., "*eating yogurt makes a person have a more regular bowel movement*" (P52). Personal health and performance were big draws: 90% of experiments sought to improve a health outcome. 51% of the experiments were rated as popular; their hypotheses were discussed on other online fora; e.g., "*having dry mouth (or Sjogren's Syndrome) promotes the growth of less beneficial gut microbes*" (P24). Common

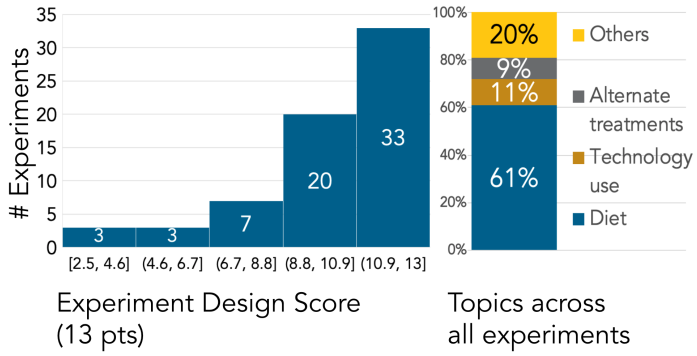

Figure 6: A) Most experiments were structurally-sound, scoring high on the structure rubric. B) Most experiments drew from personal experiences.

themes included diet (dietary styles, alcohol, fermented foods), technology use (social media, laptop, mood) and alternative treatments (homeopathy), and health (sleep, pain, gut issues) (Figure 6). Apart from being structurally-sound, the best experiment designs shared a personal experience and linked to known research. For example, one participant (P17) designed an experiment to test yogurt's effect on bowel movement and shared their motivation:

> "*For several months I have been producing Yogurt. This is fermented using commercial probiotics, Probiotic-10. My intuition was that since various microbe species were active in the making of the yogurt, this product can help relieve of the various digestive problems one persona can have. It happens that one of my sons was diagnosed with Ulcerative Colitis. among other things he was losing weight rapidly. After several weeks of consuming probiotics and/or the yogurt, he begun to recover.*"

17% of designs had novel insights that no published research addresses. For instance, "*Avoiding foods high in lectins cures long-term post-infectious diarrhea*" (P31) and "*Drinking kombucha regularly reduces joint inflammation/arthritis symptoms*" (P35) are both hypotheses of interest to citizens and microbiome researchers.

4.4.2 Reviewers Use Domain Knowledge to Improve Designs and Advocate for Participant Experience. 158 review comments (77%) were rated useful; i.e., that incorporating them would improve the experiment. Most were direct responses to a rubric question hinting that the review interface helped people focus on the salient parts of an experiment design (Figure 7A). Average comment length was 140 characters ranging from 3 characters ("yes") to 871 characters (Figure 7B,C).

Many comments (38%) requested specific details. For example, one reviewer questioned an experiment's choice of Likert scale for mood saying, "*A simplistic Likert scale seems like a bad idea. There has to be something better than this. At least a couple questions? Like, optimism, excitement, depression, anxiety?*" (P22). Reviewers provided the most comments (54%) about the hypothesis and cause & effect measures.

14% of comments demonstrated domain-specific knowledge. For example, one pointed out a conceptual mistake about a Type-1 diabetes experiment: "*A1C is measured monthly and won't change after 1g. You mean the BG value?*" (P10). A1C represents a 3-month average blood-glucose level: thus, by design it is less susceptible to short-term changes. BG here refers to the blood glucose value that depends on immediate glucose intake (among other factors). Surprisingly, reviewers did not draw from their personal experience when suggesting improvements (or at least, did not explicitly mention this was their personal experience). Some drew on counterfactual reasoning to consider about how participants might "hack" an experiment. For example, a comment on social media use and steps walked asked, "*. . .the timing of this [reporting steps taken] vs. social media use measure is off and that makes me worry about intervening use throwing things off (e.g. "phew! I've reported my Facebook for the day, now I can go use it"?)*" (P41).

People advocated for improving participant's experience (18%). Suggesting better data collection messages and times was a popular theme. We present two examples: 1) "People are not very good at remembering what they eat. Maybe an App like MyFitnessPal would be useful since it would allow participants to track all the food they eat without having to remember for too long." (P3), and 2) "How long do they [experiment participants] have to answer? What if they're eating dinner and can't get to it until 9pm?" (P6).

5 STUDY 2: PEOPLE DESIGN, REVIEW, & RUN EXPERIMENTS

The previous study found that people generated novel, structurally-sound experiments. Might they successfully run experiments with others? Participants from three communities — Kombucha, Open Humans, Beer —designed and ran experiments (Figure 8). These three experiments are a subset of the experiments from Study 1; experiment designers did not continue gathering participants for the remaining experiments.

Does drinking Kombucha improve stool consistency? Kombucha is a fermented tea drink popular in many parts of the world. Fermented foods (miso, yogurt, ayran, kefir) have been a staple in many cultures for thousands of years [10]. While there is widespread belief that kombucha "benefits the gut"[1], there is little published empirical evidence for these claims [19]. The

[1]https://www.nytimes.com/2019/10/16/style/self-care/kombucha-benefits.html

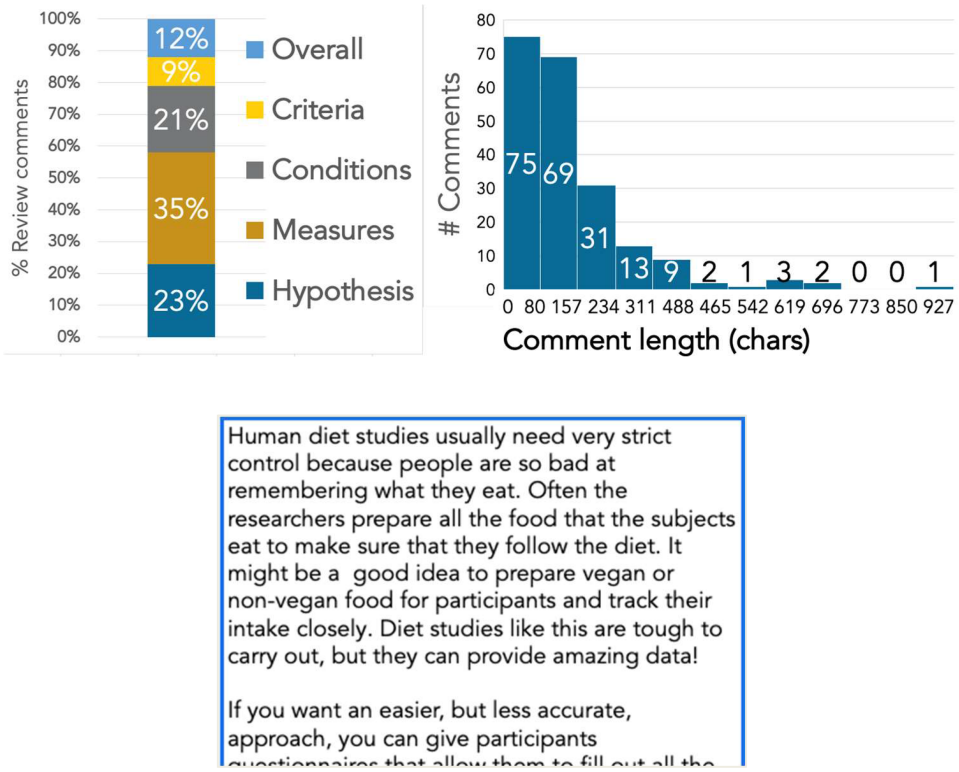

Figure 7: Summary of review (from top-left, clockwise): A) Review comments were broadly distributed across all components of experimental design. B) Review comments ranged from 3 chars "yes" to one 871 char long description. C) The longest review comment described multiple problems with an experimental design while providing numerous actionable suggestions (too long to share in its entirety).

experimenter hypothesized that kombucha supplies beneficial probiotics that help maintain normal stool consistency, and designed a between-subjects experiment.

Does reducing social media time increase optimism? Open Humans enables people to contribute personal data (e.g., genetic, social media, activity) for donation to research projects (openhumans.org). An experimenter investigated the relationship between social media and mood. Curious about the popular Facebook contagion study [13], an Open Humans member (openhumans.org) created a between-subjects experiment to investigate social media and optimism.

Does drinking a beer in the evening help people fall asleep? Some people believe that a pint of beer in the evening helps them sleep by relaxing them; others think alcohol disturbs their sleep [52]. Alcohol helps people fall asleep but disrupts the REM cycle [17]. Still, it can be more convincing to see the evidence oneself. The experimenter (a graduate student) tested the effect of beer on sleep time with a between-subjects experiment.

5.1 Results

5.1.1 Before the Experiment. From initial design to launch — 37 (kombucha), 13 (Open Humans), and 11 (beer) days elapsed. Each experiment ran for a week.

Design and Review: None of the experimenters had experience with human subjects research. All knew some concepts about experiment design; two have PhD degrees (in biology and ecology) and one is enrolled in a Computer Science PhD program. The experimenters are Brazilian, German, and US nationals. While the three experimenters had lived experience of their experiment's topic, they had never scientifically studied it.

Reviewers provided a total of 104 Boolean answers and 32 detailed comments. Comments focused on two themes. First, reviewers helped make the hypothesis and measures more specific; e.g., an experimenter started with the question "Does drinking a beer in the evening help you get to bed on time?"; the reviewers nudged the experimenter to creating the more specific hypothesis: "Drinking a 5% ABV (+-0.5%) beer between 6PM and 8PM local time helps people fall asleep no more than 30 minutes past their desired bed time." A reviewer criticized Kombucha experiment's 5-point Likert scale for bloatedness as overly vague. In response, the experimenter found and adopted the Bristol stool chart—a picture-based scale that is the industry standard [60]. Second, reviewers suggested improving data quality by instructing participants to skip confounding activities. For example, reviewers pointed out that caffeine and alcohol interact. The experimenter addressed this in instructions asking participants to abstain from coffee and alcohol. All issues that reviewers raised were tightly connected to Galileo's review

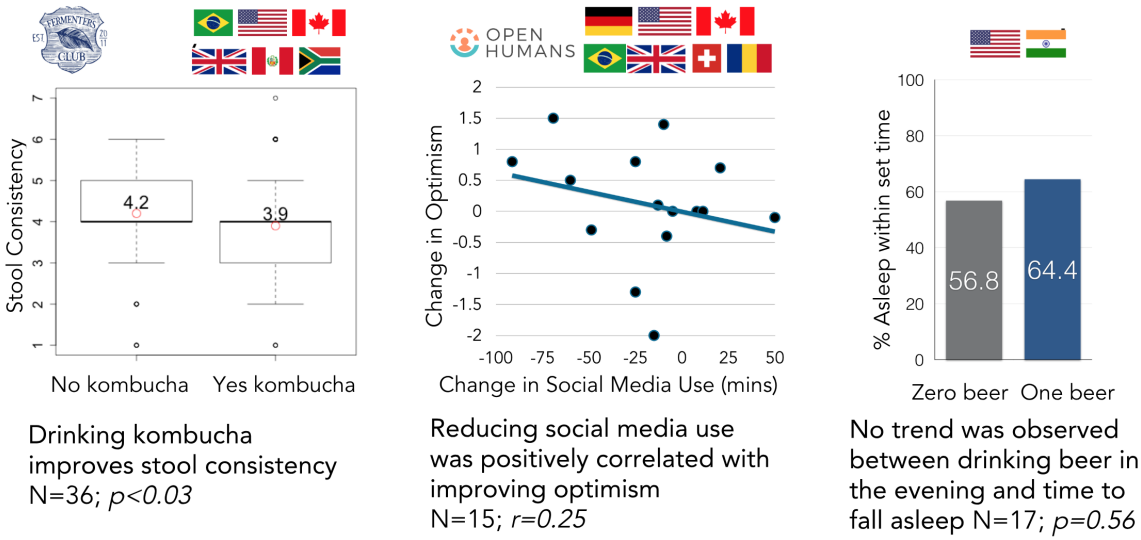

Figure 8: Three communities—Kombucha, Open Humans, Beer—designed and ran experiments; each ran for a week. The flags represent participants' nationality.

rubric. At the end of review, the three experiment designs used appropriate measures, provided a minimal-pairs design, tracked confounds, and provided appropriate criteria for participation.

Pilots: Three lessons emerged. First, some participants were loath to look at their stool. Since viewing one's stool is necessary, the experimenter added an inclusion criterion for this. Second, some participants reported eating other fermented foods in the process; the experimenter modified the instructions for participants to not consume these. Third, after failing to recruit sufficient participants, the experimenter collaborated with a kombucha fermenter in an American city who knew more kombucha enthusiasts. Before testing for the effect of social media, an Open Humans member piloted a study on the effect of 30 extra minutes of aerobic exercises on sleep. However, potential participants were loath to alter their lifestyle this dramatically, and so the experimenter abandoned the study.

Finding participants: The *Kombucha* experimenter publicized the experiment on Instagram, Twitter, and newsletter; they also created a poster, and reached out to enthusiasts in their city in Brazil and an American city. The Open Humans experimenter recruited on social media, a mailing list, and the Open Humans Slack channel. The beer experimenter reached out to peers interested in community experimentation and/or the effects of alcohol. At least one potential participant in each of the three experiments was excluded because of inclusion/exclusion criteria.

5.1.2 During the Experiment. Retention: 57 people signed up for the kombucha experiment; 36 completed it (68%). Retention rates were similar for the Open Humans experiment (63%) and higher for beer (90%) (Figure 9). 78% of dropouts occurred in the first 48 hours. The reasons participants reported for dropping out included lack of interest, holidays, and work travel.

Adherence: *Kombucha* garnered 76% adherence: 86% for days of no kombucha, and 70% when asked to drink kombucha. Most

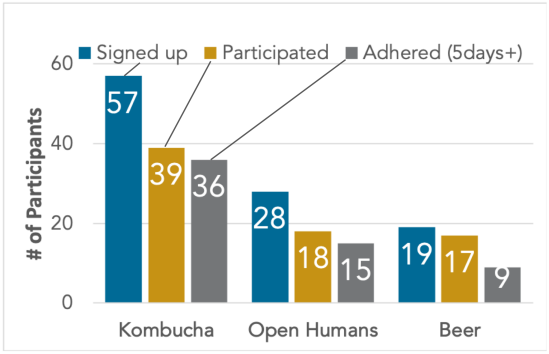

Figure 9: After signing up, a smaller fraction of people participated in Kombucha (68%) and Open Humans (63%) experiments than Beer (90%). However, those who participated reported greater adherence in *Kombucha* (92%) and Open Humans (83%) compared to Open Humans (50%). Reasons for non-adherence included being busy, annual leave, and brewers needing to check on the taste of kombucha.

Open Humans participants reported high adherence, cutting social media use in half or more (Figure 9). Each day, an average of 54% of participants in the beer experiment reported following the condition requirement (drinking 1 or 0 beers by 8PM). 15 of 17 failed to comply on at least one day.

Some participants disclosed confounds and reasons for non-adherence. For example, drinking alcohol was a reported confound, because it might affect kombucha's impact on the body. Similarly, participants' non-adherence reports included scheduled disruptions like travel and holidays and work responsibilities like brewers needing to check on the taste of kombucha. Non-adherence for the beer experiment included drinking wine rather than beer, drinking after

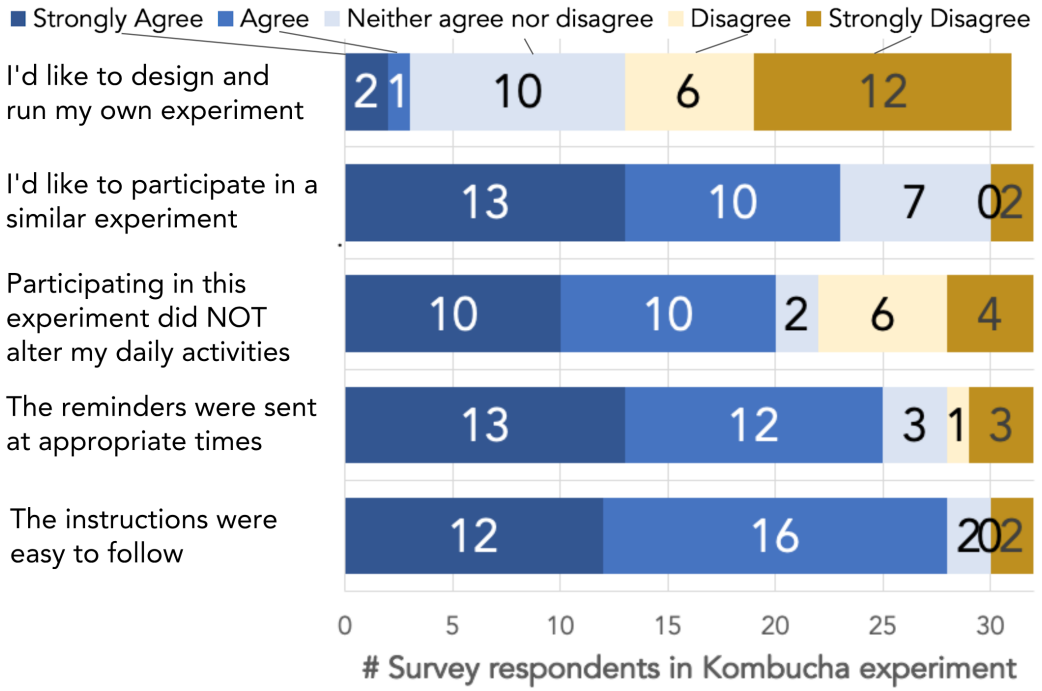

Figure 10: Kombucha participants reported an overall positive experience; nearly all expressed an interest in participating in similar experiments (23/32). Most reported that its instructions were easy to follow (28/32) and that reminder times were appropriate (25/32).

8PM, drinking more than one beer, or not drinking in the drink-one condition.

Data Collection: Most American participants selected text solicitations (86%); participants elsewhere received email solicitations due to varying regulations around automated text messages (e.g., replying to an automated text message in Brazil or India is infeasible since the source number is masked). 56% of participant responses came within 30 minutes of the solicitation; 21% of responses took more than 90 mins. Participants sparingly responded to follow-up questions. Experimenters used the remind participant button 2 (kombucha) and 3 (Open Humans) times to remind participants with missing data.

Clarifying questions: The experiment requested that all participants adhere to the protocol as much as possible without harming their health. Participants could ask the experimenter (via the platform) if confused. Participants' clarifying questions focused on measurements (e.g., measuring stool consistency once during the day or multiple times) and specific lifestyle choices (e.g., consuming probiotics while drinking kombucha?). Participants in kombucha experiment reported an overall positive experience (Figure 10).

6 DISCUSSION

This paper's results surface three challenges in democratizing complex tasks like experimentation: 1) all three experimenters had advanced degrees; 2) two of the three completed experiments were underpowered; and 3) participants demonstrated varying levels of adherence. In this section, we reflect on our findings and provide specific suggestions from the social computing literature.

6.1 Supporting Experimenters Without Advanced Degrees

Our results don't demonstrate that "anyone" can design & run experiments. A long line of research documents self-selection of users with advanced degrees on online platforms. Most citizen science participants (regardless of participation level) have advanced degrees [1]. Our results reflect data contributions from such typical participants while also supporting some participants in creating experiments. This is a net positive outcome. Consider the shift in scientific knowledge creation: even people with relevant knowhow avoid running experiments [56] while a group of kombucha fans (many brewers) used Galileo to successfully test a common community intuition. We think such interest-based involvement in science is an exciting possibility even with the current limitations of who the experiment designers are.

Contributions to web platforms vary across educational levels. MOOCs are disproportionately completed by learners from more-affluent and better-educated neighborhoods [25], and 73% of citizen scientists and Wikipedia contributors have advanced degrees [1, 59]. While all 36 Kombucha participants wanted to participate in future experiments, only two participants wanted to run their own; both have advanced degrees. An advanced degree is not a prerequisite to use Galileo or many other internet platforms but having one confers multiple advantages that may lead to self-selection. The complex knowledge needs of experimentation can potentially amplify such participation inequality. Long-term platform use with interventions

for 'non-power-users' to design experiments would be valuable future work

Simply asking people to contribute data might work for citizen science projects but running experiments might be a bigger leap. We suggest two improvements. First, reduce effort by providing ready to run experiments; common health and lifestyle topics such as coffee consumption and sleep might be good candidates. Running a sample experiment enables people to pilot the platform before testing their ideas while also potentially making them comfortable with the idea of experimentation itself. Second, support a growth mindset [16] that emphasizes that anyone can learn how to run an experiment. While our platform's interface provided plenty of messaging around testing one's ideas with experiments, it did not suggest strongly enough that using the platform required no prior knowledge and people could learn new skills—-like minimal pairs design—by using the platform.

Another reason why those with advanced degrees might have run experiments: they were aware of potential participants. All three experimenters had access to people who were interested in similar topics; e.g., the Open Humans experimenter received both participants and feedback for their idea from the group's slack community. Such affinity spaces are known to provide potential participants as well as social support [22]. To tackle this, the design workflow can nudge the creator to start their experiment design by thinking of topics relevant to their social connections.

While our suggestions could likely broaden the pool of experimenters, they do not help with fundamental socio-economic challenges. For example, our suggestions assume that running an experiment is a good use of volunteers' time; this might not be true for people already overburdened with life and work-related activities. We clarify that our experience and reflection with the Galileo research prototype suggests one way ahead for greater citizen-led inquiry; we do not claim that it overcomes existing social, educational, and other disparities.

6.2 Guidance Techniques to Enable Citizens to Recruit Others

Two of the three completed experiments were underpowered. The kombucha experiment gained critical mass after the original designer collaborated with another fermenter to reach out to more people across countries. Citizen experimenters learned what many scientists know: recruiting participants is difficult and time-consuming. This suggests that a good experimental design is not enough and recruiting is the next challenge for citizen scientists on their way to develop meaningful knowledge. Galileo did not provide any explicit knowledge support for recruiting participants beyond providing a sharable link to join the experiment. While the absence of shared knowledge with experts can sometimes give novices' work a boost (e.g. identifying green pea galaxies on Galaxy Zoo [5]), it is less useful when the lack of knowledge is a hindrance.

Tools for training and collaboration can help by clearly conveying the importance of getting enough participants; enabling experimenters estimate what "enough" is; and providing sources and strategies to recruit participants. Citizen experimenters aren't as ardent about sufficient participation numbers as professional scientists. One important piece of technical knowledge is performing power analysis before running the experiment. Additionally, following the lead of data journalists [24], conveying future results through real-world effect sizes—such as additional years you'll live—to both experimenters and potential participants might be useful. Moreover, the experimenter need not find all the participants by themselves. Akin to a Clinical Research Coordinator, a separate recruitment role can help the experimenter rope in others to help out. Participants signed up for an experiment can also assist by suggesting others (snowball sampling).

6.3 Supporting Participant Retention And Adherence

The opportunity to contribute to science is exciting; *Kombucha* participants mentioned this as a motivation. While altering one's lifestyle for a day might not be very difficult for many people, doing the same for a week (or more) might be tedious enough to entirely avoid participating, drop out after signing up, or not adhere to the instructions. It's also likely that the trust placed in institutional researchers might not extend to citizen experimenters [11].

Why might participants join citizen-led experiments and adhere to the instructions? Common reasons why people join *expert-led* experiments include [49]: to help find an answer to a question that personally affects them, to gain access to potential treatments, and for credit or monetary compensation. Drawing on findings from social computing and crowdfunding [29, 32], we suggest four remedies to improve both participation and adherence numbers: 1) increase participant trust by sharing more information about the experiment's goals, approximate effort expected, and the experimenter's biography; 2) implement activation thresholds to make social reciprocity explicit for group activities and to reduce potentially wasted efforts [9]; 3) leverage participation from communities with already strong ties and common goals; 4) allow people to pre-register for topics of interest so they might join relevant experiments created at a later date [3].

Our study did not provide experimenters or participants monetary compensation. Consequently, people's motivation is more intrinsic, which has benefits [48] (e.g. telling people the importance of their work improves performance [6]), but also empirically shows a high dropout rate. Compensation may help some citizen science experiments.

6.4 Do Citizen Experiments Benefit or Harm Society?

This paper has outlined the positive potential for citizen designed experiments. It's worth considering the risks. We see two major concerns: 1) a poorly designed experiment with a faulty conclusion can influence people in dangerous ways; and 2) people might hurt others/themselves by running/joining experiments that include potentially harmful steps.

6.4.1 Challenges of Misinformation And Harmful Participation. At its best, over time scientific experiments expand human knowledge and correct mistakes when they occur. However, sometimes the popular press (and/or netizens with no training in scientific journalism) report a headline-grabbing result that is inaccurate, but not the subsequent correction and elaboration [21]. Particularly with science, when ideas are newsworthy but low-quality, people can incorporate misguided ideas in a way that be difficult to dislodge. One of

the most notorious examples is the (debunked) claim that vaccines, especially MMR vaccine, cause autism by disrupting the body's microbial composition and/or introducing harmful chemicals. At a time of rising autism diagnoses, this claim terrified parents and continues to impede childhood vaccination more than two decades later. Wakefield's publication linking MMR vaccine to autism (later retracted) was a serial case study [23], not an experiment. While sharing case studies can help identify valuable leads for further study, the small size and biased selection create enormous risk of confounds and spurious relationships. (In this case, unidentified correlated timing in the measures and undisclosed financial ties by the author further clouded the picture.) Furthermore, novices overly rely on surface details of scientific explanations instead of understanding the underlying logic [39]. Our hope is that democratizing the doing of science may help the public interpret science news and reduce the risk of leaping to conclusions.

Furthermore, not all experiments are appropriate for people to run and some gatekeeping of citizen experiments might be necessary. 62 of the 66 complete designs were posted online on Galileo for others to view; 4 were taken down because the research team identified them as risky. For example, one removed design sought to investigate the effect of colloidal silver on cognitive performance. Some online communities believe that colloidal silver (tiny silver particles suspended in liquid) to have beneficial properties [41]. While the designer may be well-intentioned, consuming colloidal silver can cause irreversible damage such as skin discoloration, and the NIH has sued manufacturers for misleading claims [26]. Galileo offers keyword triggers for alerting both the designer and the research team of possibly dangerous experiments. For example, an experiment containing "cancer" or "CBD" triggers an email to the research team; use of the word "cancer" indicates potential health risks for participants (who might be cancer patients) while "CBD" (Cannabidiol) indicates potential legal risks across many places around the world.

6.4.2 How to Proceed? We clearly note that sending research prototypes like Galileo out to the real world requires a huge amount of work that is not the focus of this paper. For instance, gathering inputs from experts—subject matter experts, ethicists, psychologists, misinformation scholars—provides one starting point to understand the effects of citizens conducting experiments. Using experts' insights requires careful, slow, and small-scale collaborative design and testing. This work does not tackle these important challenges; this paper provides a proof-of-concept for how citizens might run experiments.

Sifting through ideas expressed by people for experimentation, we believe citizen experiments seem well suited for ideas that meet three criteria; they must 1) be scientifically tenable, 2) combine high excitement with low efforts, and 3) provide zero to no risk. Scientifically tenable means that the experiment answers a gap in research literature, minimizes placebo effects, and yields results in a week with a high likelihood. To be low-effort, all the experimental steps (including reporting data) should be easy to understand and perform. Finally, the experiment should not provide any cause of harm to participants and it should be legally and ethically permissible across countries and cultures. As a crude beginning, this can be operationalized as the existence of numerous anecdotes about potential upsides with none or well understood downsides. For instance, bee venom reduces Lyme disease symptoms (an idea proposed on the Galileo platform) is an idea with anecdotal benefits but the existence of venom implies non-trivial possibility of self-harm; therefore, such an experiment is an unlikely candidate for citizen-led experimentation in our view.

7 CONCLUSION

This paper investigated citizen-led experimentation with the novel Galileo social computing system. Two empirical investigations tested this approach. For us, the most striking result is that online volunteers collaboratively designed and ran experiments by drawing on their lived experience. Our work also illustrates the challenge of helping novices successfully execute a complex knowledge task like experimentation. Specifically, finding and retaining participants and making the platform accessible and useful to a broader audience emerged as key challenges. With systems that enable citizen-led experimentation, people can potentially match scientists' knowledge with their lived experiences to create insights both for themselves and for the scientific community. A future of science that includes deeper contributions from more people is a future worth striving for.

ACKNOWLEDGMENTS

We thank NSF award #1735234 for funding. We thank Dingmei Gu, Liby Lee, Kaung Yang, Orr Toledano, and Aliyah Clayton for help developing the website and running pilot studies. We thank Adriana Daudt Grativol and Austin Durant (Fermenter's Club, San Diego) for their inputs on the experiment review and participant gathering phases. Steven Dow and Ailie Fraser provided different useful approaches to frame this work. We thank anonymous reviewers for their thoughtful critiques over multiple cycles; their inputs deepened the lead author's understanding of the sociotechnical nature of this research. Finally, we are grateful to volunteers who participated in the experiments on Galileo.

REFERENCES

[1] National Academies of Sciences , Engineering, and Medicine. 2018. Learning through citizen science: enhancing opportunities by design. National Academies Press.

[2] Tim Althoff, Rok Sosič, Jennifer L. Hicks, Abby C. King, Scott L. Delp, and Jure Leskovec. 2017. Large-scale physical activity data reveal worldwide activity inequality. *Nature* 547, 7663: 336–339.

[3] Michael S Bernstein, Joel Brandt, Robert C Miller, and David R Karger. 2011. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 33–42.

[4] Rick Bonney, Caren B. Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V. Rosenberg, and Jennifer Shirk. 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience* 59, 11: 977–984.

[5] Carolin Cardamone, Kevin Schawinski, Marc Sarzi, Steven P Bamford, Nicola Bennert, C Megan Urry, Chris Lintott, William C Keel, John Parejko, Robert C Nichol, and others. 2009. Galaxy Zoo Green Peas: discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society* 399, 3: 1191–1205.

[6] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior and Organization* 90: 123–133.

[7] William G Chase and Herbert A Simon. 1973. Perception in chess. *Cognitive psychology* 4, 1: 55–81.

[8] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. 2019. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*,

531.
[9] Justin Cheng and Michael Bernstein. 2014. Catalyst: triggering collective action with thresholds. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 1211–1221.
[10] Stephanie N Chilton, Jeremy P Burton, and Gregor Reid. 2015. Inclusion of fermented foods in food guides around the world. *Nutrients* 7, 1: 390–404.
[11] Caren B. Cooper, Jennifer Shirk, and Benjamin Zuckerberg. 2014. The invisible prevalence of citizen science in global research: Migratory birds and climate change. *PLoS ONE* 9, 9.
[12] Seth Cooper, Firas Khatib, Adrien Treuille, and Et Al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307: 756–760.
[13] Lorenzo Coviello, Yunkyu Sohn, Adam D.I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler. 2014. Detecting emotional contagion in massive social networks. *PLoS ONE*.
[14] Nediyana Daskalova, Karthik Desingh, Alexandra Papoutsaki, Diane Schulze, Han Sha, and Jeff Huang. 2017. Lessons learned from two cohorts of personal informatics self-experiments. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3: 1–22.
[15] Steven P. Dow, Anand Kulkarni, Scott R. Klemmer, and Bjorn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*, 1013–1022.
[16] Carol Dweck. 2016. What having a "growth mindset" actually means. *Harvard Business Review* 13: 213–226.
[17] Irshaad O Ebrahim, Colin M Shapiro, Adrian J Williams, and Peter B Fenwick. 2013. Alcohol and sleep I: effects on normal sleep. *Alcoholism: Clinical and Experimental Research* 37, 4: 539–549.
[18] Randall W Engle. 2002. Working memory capacity as executive attention. *Current directions in psychological science* 11, 1: 19–23.
[19] E Ernst. 2003. Kombucha: a systematic review of the clinical evidence. *Complementary Medicine Research* 10, 2: 85–87.
[20] f.lux. 2019. f.lux: sleep research. Retrieved from justgetflux.com/research.html
[21] Björn Fjæstad. 2007. Why journalists report science as they do. *Journalism, science and society*: 123.
[22] James Paul Gee. 2005. Semiotic social spaces and affinity spaces. *Beyond communities of practice language power and social context* 214232.
[23] Fiona Godlee, Jane Smith, and Harvey Marcovitch. 2011. Wakefield's article linking MMR vaccine and autism was fraudulent. *BMJ* 342.
[24] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. 2012. *The data journalism handbook: How journalists can use data to improve the news*. O'Reilly Media, Inc.
[25] John D Hansen and Justin Reich. 2015. Democratizing education? Examining access and usage patterns in massive open online courses. *Science* 350, 6265: 1245–1248.
[26] National Institute of Health. 2018. Colloidal Silver | NCCIH. Retrieved from http://nccih.nih.gov/health/silver
[27] Susanne Hecker, Muki Haklay, Anne Bowser, Zen Makuch, and Johannes Vogel. 2018. *Citizen science: innovation in open science, society and policy*. UCL Press.
[28] Eric von Hippel. 2005. *Democratizing innovation: The evolving phenomenon of user innovation*. MIT.
[29] Julie S Hui, Elizabeth M Gerber, and Darren Gergle. 2014. Understanding and leveraging social networks for crowdfunding: opportunities and challenges. In *Proceedings of the 2014 conference on Designing interactive systems*, 677–680.
[30] Gerald C Kane. 2009. It's a Network, Not an Encyclopedia: A Social Network Perspective on Wikipedia Collaboration. In *Academy of management proceedings*, 1–6.
[31] Ravi Karkar, Jessica Schroeder, Daniel A Epstein, Laura R Pina, Jeffrey Scofield, James Fogarty, Julie A Kientz, Sean A Munson, Roger Vilardaga, and Jasmine Zia. 2017. Tummytrials: a feasibility study of using self-experimentation to detect individualized food triggers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6850–6863.
[32] Jennifer G Kim, Ha Kyung Kong, Karrie Karahalios, Wai-Tat Fu, and Hwajung Hong. 2016. The power of collective endorsements: credibility factors in medical crowdfunding campaigns. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4538–4549.
[33] Paul A Kirschner and Jeroen Van Merriënboer. 2008. Ten steps to complex learning a new approach to instruction and instructional design.
[34] Rob Knight, J. Metcalf, and K. Amato. 2016. Gut Check: Exploring Your Microbiome. Coursera. Retrieved from https://www.coursera.org/learn/microbiome
[35] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15, 2: 155–163.
[36] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. 2018. Rotating Online Behavior Change Interventions Increases Effectiveness But Also Increases Attrition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: 95.
[37] Chinmay E Kulkarni, Michael S Bernstein, and Scott R Klemmer. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 75–84.
[38] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, 23–34.
[39] Samuel Lau, Tricia J Ngoon, Vineet Pandey, and Scott Klemmer. 2019. Experiment Reconstruction Reduces Fixation on Surface Details of Explanations. In *Proceedings of the 2019 on Creativity and Cognition*. 578–582.
[40] Doris Lee, Joanne Lo, Moonhyok Kim, and Eric Paulos. 2016. Crowdclass: Designing classification-based citizen science learning modules. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP '16)*.
[41] Jayne Leonard. 2016. 15 Reasons You Need A Bottle Of Colloidal Silver In Your Home. Retrieved from naturallivingideas.com/colloidal-silver-benefits-and-uses/
[42] Dana Lewis and Scott Leibrand. 2016. Real-World Use of Open Source Artificial Pancreas Systems. *Journal of Diabetes Science and Technology* 10, 6.
[43] Kurt Luther, Scott Counts, Kristin B Stecher, Aaron Hoff, and Paul Johns. 2009. Pathfinder: an online collaboration environment for citizen scientists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 239–248.
[44] Wendy E Mackay, Caroline Appert, Michel Beaudouin-Lafon, Olivier Chapuis, Yangzhou Du, Jean-Daniel Fekete, and Yves Guiard. 2007. Touchstone: exploratory design of experiments. *CHI '07 Proceedings of the SIGCHI Conference on Human Factors in Computing System*: 1425–1434.
[45] D. W. Martin. 2007. *Doing psychology experiments*. Cengage Learning.
[46] J Nathan Matias and Merry Mou. 2018. CivilServant: Community-led experiments in platform governance. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, 9.
[47] Daniel McDonald, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, and Yingfeng Chen. 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* 3, 3: e00031-18.
[48] UK National Council for Voluntary Organisations. 2018. Why Volunteer? Retrieved from ncvo.org.uk/ncvo-volunteering/why-volunteer
[49] NIH. 2015. NIH Clinical Trials Research and You. Retrieved from nih.gov/health-information/nih-clinical-research-trials-you/basics
[50] Vineet Pandey, Amnon Amir, Justine Debelius, Embriette R Hyde, Rob Knight, and Scott Klemmer. 2017. Gut Instinct: Creating Scientific Theories with Online Learners. In *2017 CHI Conference on Human Factors in Computing Systems (pp. 6825-6836). ACM.*
[51] Vineet Pandey, Justine Debelius, Embriette R Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2018. Docent: transforming personal intuitions to scientific hypotheses through content learning and process training. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 9.
[52] Michael J Breus Ph.D. Alcohol and Sleep: What You Need to Know. *Psychology Today*. Retrieved from psychologytoday.com/us/blog/sleep-newzzz/201801/alcohol-and-sleep-what-you-need-know
[53] Katharina Reinecke and Krzysztof Z Gajos. 2014. Quantifying visual preferences around the world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 11–20.
[54] Katharina Reinecke and Krzysztof Z Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*.
[55] Divit P Singh, Lee Lisle, T M Murali, and Kurt Luther. 2018. CrowdLayout. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (CHI '18), 1–14.
[56] Stefan Thomke. 2020. Building a culture of experimentation. *Harvard Business Review* 98, 2: 40–47.
[57] Nai-Ching Wang, David Hicks, and Kurt Luther. 2018. Exploring Trade-Offs Between Learning and Productivity in Crowdsourced History. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: 178.
[58] Paul Wicks, Timothy E Vaughan, Michael P Massagli, and James Heywood. 2011. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology* 29, 5: 411–414.
[59] Wikipedia. Community Insights/2018 Report. *2018*. Retrieved from https://meta.wikimedia.org/wiki/Community_Insights/2018_Report#What_progress_has_been_made_in_the_diversity_of_Wikimedia_communities
[60] Wikipedia. 2018. Bristol stool scale. Retrieved from en.wikipedia.org/wiki/Bristol_stool_scale
[61] Wesley Willett, Jeffrey Heer, Joseph Hellerstein, and Maneesh Agrawala. 2011. CommentSpace: structured support for collaborative visual analysis. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 3131–3140.
[62] Zooniverse. 2007. Galaxy Zoo. Retrieved December 31, 2016 from galaxyzoo.org