Galileo: Roles and Procedural Support in a Social Computing System for Citizen Experimentation

VINEET PANDEY, Design Lab, UC San Diego, La Jolla, CA
TUSHAR KOUL, Design Lab, UC San Diego, La Jolla, CA
CHEN YANG, Design Lab, UC San Diego, La Jolla, CA
DANIEL MCDONALD, Department of Pediatrics, Design Lab, UC San Diego, La Jolla, CA
MADELEINE PRICE, Open Humans Foundation, Sanford, NC
BASTIAN GRESHAKE TZOVARAS, Open Humans Foundation, Sanford, NC
ROB KNIGHT, Department of Pediatrics, Design Lab, UC San Diego, La Jolla, CA
SCOTT KLEMMER, Design Lab, UC San Diego, La Jolla, CA

This paper introduces a crowdsourcing architecture that integrates *role differentiation for experimentation* with *procedural support* using three techniques: 1) experimental design workflow that provides just-in-time training, 2) review with scaffolded questions, and 3) automated checkers that implement standardized behaviors. We instantiate this approach in a system for designing and running experiments, and present three empirical investigations of this architecture: 1) a deployment with three communities across 8 countries; 2) a study across 16 countries; and 3) a controlled between-subjects experiment. Communities successfully designed and ran experiments without prior training in experiment design. They generated structurally-sound experiments on personally-meaningful topics. Procedural training yielded higher quality experiments than watching lecture videos. Remarkably, participants using Galileo garnered higher condition-blind ratings than doctoral students trained in experiment design. These results highlight social computing's promise for scaffolding personally-meaningful knowledge work like experimentation.

KEYWORDS

Social computing systems; citizen science; crowdsourcing; online learning

INTRODUCTION: FROM EXPERIENCES TO EXPERIMENTATION

Scientific experimentation features personal inspiration, conceptual creativity, and technical requirements that are inscrutable for a lay individual but necessary for success. Broadening the pool of experimenters and participants could help people investigate their curiosities, develop solutions to improve performance, and assist institutional researchers. This paper introduces a crowdsourcing architecture that divides experimentation into roles supported via procedural training (Figure 1). We instantiate this approach in a system for designing and running experiments, and present empirical results demonstrating the success and limitations of this approach.

While professional scientists and commercial ventures run experiments every day, with notable exceptions [16,47], empirical papers from non-professionals are vanishingly rare. This

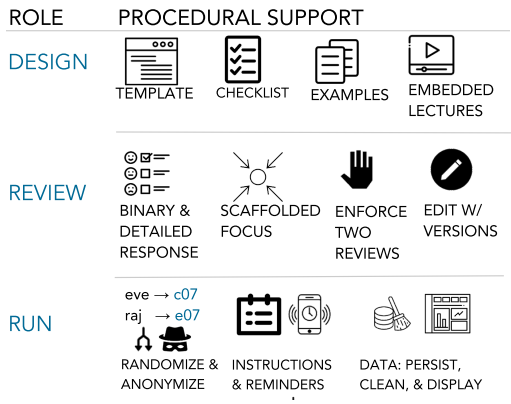


Figure 1: Galileo enables anyone to design and run experiments to test their intuitions. The key architectural insights are providing specific roles and supporting them with procedural training. Experiment creators can invite anyone to review and participate in the experiment. Participants from around the world join experiments, follow instructions, and provide data in response to automated data collection reminders.

STUDY	Design	Review	Run	Users	N	Main finding
1	X			Undergraduates	72	Structured procedural training yielded better experimental videos than video lectures with the same content
2	X	X		Online users	54	People designed structurally-sound experiments. Reviewers drew from personal intuitions and provided useful feedback on experiment structure and content
3	X	X	X	Communities	68	Communities designed, iterated upon, and ran week-long experiments without prior training in experiment design

Table 1: Two deployments and one between-subjects experiment examined the efficacy of procedural support for designing and running experiments. We found that creating roles and supporting them were helpful in enabling citizen experimentation.

biases the questions asked, studies run, and knowledge created [30]. Currently, both those asking scientific questions and those participating in studies are not representative of the global population. Behavioral science research's staple participants are university undergraduates [30], and medical research funding supports only a sliver of ideas and researchers compared to the wide range of possibilities and demographics [1].

Broadening the pool of experimenters and participants could help people investigate their curiosities, develop solutions to improve health and performance, and assist institutional researchers. Early efforts to diversify participation are bearing fruit: *Lab in the Wild* recruits anyone with an internet connection for behavioral studies [61]; *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 [18], using activity trackers, and joining scientific projects—have enabled valuable insights on topics including obesity [2], aesthetic preferences [62], sleep [24], and the human microbiome [52].

Building on these new *participation* channels, we suggest that democratizing *experimentation* may also expand the gamut of scientific knowledge. People have questions about their health, but lack the expertise and resources to scientifically investigate them. How might online systems support more complex activities that leverage the creativity and diversity of a global community? One reason for crowdsourcing's focus on brief tasks with a correct answer is such tasks require little training and produce verifiable responses [37]. Furthermore, drawing on personal experiences to make diverse contributions is actually a *benefit* for creative work, not a problem.

This paper presents three empirical investigations of role-based crowdsourcing with procedural support (Table 1). First, a controlled between-subjects experiment with 72 participants found that structured procedural training yielded significantly higher-quality experiment designs than with lecture videos. Remarkably, Galileo participants garnered higher condition-blind ratings than doctoral students trained in experiment design. Second, a deployment across 16 countries found that people generated structurally-sound experiments on personally-meaningful topics. Third, a field deployment with three communities—kombucha, Open Humans, and beer—across 8 countries demonstrated that communities designed, iterated on, and ran week-long experiments without prior training in experiment design. These results highlight social computing's promise for scaffolding personally-meaningful knowledge work like experimentation.

RELATED WORK

This work draws on and contributes to citizen science, crowdsourcing, and lead-user innovation. (Figure 2). From *citizen science*, we take the idea of people using their complementary insights

and cognitive surplus towards scientific work [4], and provide a system for scientific experimentation to focus citizens' effort towards personally-meaningful work. From *crowdsourcing*, we draw the idea of dividing a complex activity into multiple tasks—some self-sourced, others crowd-sourced; and introduce procedural support for roles to perform these tasks. From *lead-user innovation*, we learn that success requires possessing both the context (lived experience) and the tools (knowledge) to make changes and learn from them [32], and contribute ways to embed learning and collaboration to improve the odds of success.

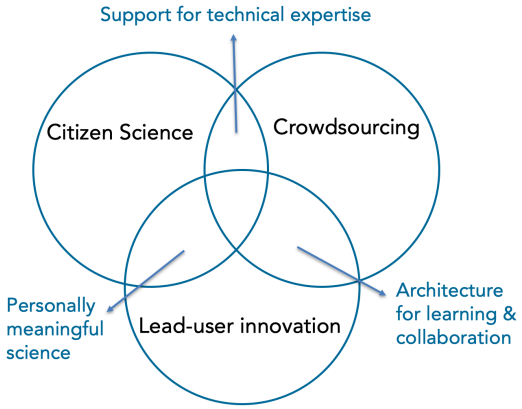


Figure 2: Galileo draws from—and contributes to—citizen science, crowdsourcing complex work, and lead-user innovation

The Future of Citizen Science: Running own experiments?

Modern science is increasingly collaborative [57]: citizens count bird species, identify galaxies, edit protein structures, and create novel hypotheses [16,59,73]. One reason is that different people provide different expertise that can vet claims and fix mistakes [33]. Collaboration benefits creativity when it brings different perspectives that build on each other; it impedes creativity (or worse, causes regression) when—through groupthink—it spreads biases rather than removing them [68]. A humbling example of the power of fresh eyes: volunteer citizen scientists identified a new class of galaxies ("green pea" galaxies) after researching green blots on Galaxy zoo images; experts had dismissed these images as apparatus error [6]. Such collaboration requires *strategic isolation:* providing just enough scaffolding to keep biases independent, while not stifling original ideas for bottom-up knowledge creation.

While public contributions have supported institutional science; it's rare for citizens to design their own experiments. Despite a predetermined goal and a formalized process, experimentation requires making situationally-appropriate decisions. A dependent variable may produce crisp numbers but feedback on the experiment design itself is more multifarious. Good experiment design is inherently user centered: how will participants interpret the instructions? Experiment designers need awareness of others' interpretation of their ideas and asks. Feedback and iteration might be key to creative success, especially for novices. Feedback can be provided by experts [20,64], peers [5,40], software [19,28], or even oneself [5,64]. While feedback from novices can potentially improve both structure and content, it can also emphasize superficial issues over the underlying structure [11].

Crowdsourcing with Procedural Support for Roles (Figure 3)

Canonical crowdsourcing breaks larger tasks into microtasks; algorithms specify the division, dependency, and agglomeration activities while workers perform small tasks supported by task-specific guidelines [37]. Crowdsourcing has worked around the challenge of limited procedural learning by providing *some* learning in the interface, or by leveraging experts.

Systems like Foldit and EteRNA powerfully show how carefully-constructed interfaces provide novices with the task-specific expertise to solve problems that only experts previously could [16,41,44,73]. Foldit introduces an interface and 3D game for specifying low-energy protein structures to a direct manipulation game. For tasks that don't have as a crisp visual analogue as protein folding, people need better conceptual support. For more abstract tasks, CrowdLayout and Cicero

provide guidelines and static rules that crowdworkers can use to reason about their choices and to improve layouts of biological networks [10,67].

Leveraging existing expertise is another approach for complex knowledge work. One strategy directly employs experts' just-in-time feedback to improve crowd work [20]. Workflows manage experts for open-ended work like developing interactive prototypes [63]. Flash Organizations uses automated hiring, a hierarchy with a central leader, and optional team leaders for collaborative projects like product design [70]. Another strategy creates roles that enable more experienced crowd members to orchestrate the work. Ensemble supports *leaders* in guiding and constraining crowd contributions [35]. Role-based approaches confer three benefits: 1) clean delineation of responsibilities improves chances of task completion, 2) clustering similar tasks reduces overhead and increases consistency; 3) people can decide their contribution levels. However, experts are expensive, in short supply, and sometimes prone to groupthink. How might groups of *novices* perform complex work like experimentation?

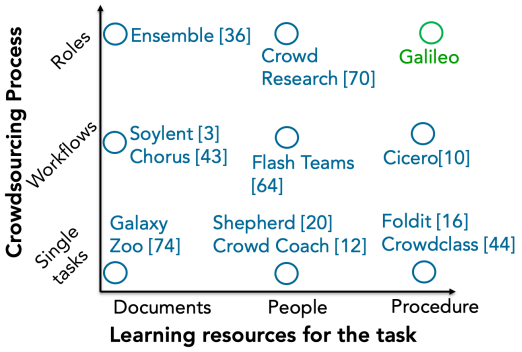


Figure 3: Crowdsourcing approaches vary in learning support for activities. Documents include guidelines or tutorials; people are peers/experts. Some, like Ensemble, provide interface support for different roles. Others, like Shepherd, provide timely feedback from experts. Galileo creates roles for experimentation and provides procedural support.

Lead-user Innovation: Successful when People Know What to Do and How to Do It

Lead-user innovation is both an inspiration and an application area for this work. Lead users are users of a product (or service) who experience advanced needs unmet by existing products. The power of lead-user innovation is that lived experience, a tight feedback loop, and strong personal motivation can yield different and sometimes better products than experts [32]. For example, diabetes patients have improved insulin delivery [47] and snowboarders have improved their binding ergonomics. Personal needs and challenges can be highly motivating, inspiring people to measure and seek correlations to improve their lives [55].

Lead users have an advantage when the key ingredient is experience intensive; experts retain the advantage for solution-intensive innovations [32]. Professionals have the advantages of training, conceptual knowledge, pre-existing organizational structure for collaboration and support, and direct access to resources like manufacturing. Closer to our research, Tummy Trials asked participants to generate health questions, introducing an experimental protocol combining ideation and self-tracking. We build on this prior work, introducing a structured approach for this distributed innovation, a platform embodying this approach, and focus on controlled experiments as opposed to self-tracking or informal iteration [34].

Self-tracking Offers Insights but Not Causality

People have questions about their health, but lack the expertise and resources to scientifically investigate them. These concerns are especially acute when multiple factors interact. Despite knowledge gained from lived experiences, people lack the procedural tools to gain the causal knowledge they seek. Many self-tracking efforts suffer from structural flaws that prohibit people from actually learning what they'd like to know [14,48]. A frequent error is mistaking correlation for causation [53]. People falsely believe that when one event follows another, the initial event is

the cause: *post-hoc ergo propter hoc*. At the same time, professional science suffers from structural biases. By creating controlled experiments (as opposed to tracking oneself), people can test their intuitions at a larger scale, potentially unearthing novel results. How can we train people in designing and running experiments to answer their personally-meaningful questions?

Contribution: Roles Support via Just-in-time Skill Acquisition

This paper introduces a crowdsourcing architecture that integrates *role differentiation for experimentation* with *procedural support* using three techniques: 1) experiment design workflow that provides just-in-time training, 2) review with scaffolded questions, and 3) automated checkers that implement standardized behaviors.

For every experiment, Galileo requires three roles: designer, reviewer, and participant. Roles bundle a set of smaller steps into an assignable unit. Galileo offers procedural support for every role, *e.g.* the experimenter need not have existing expertise; the system bears that load. Like the Shepherd system for writing product reviews [20], Galileo provides just-in-time support for experimentation. There are two key differences: 1) Galileo scaffolds the entire experiment creation process itself not just the post-draft feedback stage, and 2) Galileo does not draw on expert time – the knowledge is implemented in the software itself.

Possibly the closest design to our work are the workflows and contextual support provided in tax software like TurboTax (turbotax.intuit.com). With TurboTax, the task is pre-decided. Galileo provides more open-ended support: people decide their questions while the system provides learning resources about the paradigm and contextual task support for the requirements. Finally, successfully running an experiment requires managing multiple processes such as random assignment of participants, anonymizing participant details, and sending instructions and reminders for data collection. Galileo provides automated support for these processes.

Prior work has demonstrated the power of distinct phases for design, testing, and analysis [38]. Most related to this paper, Touchstone introduced a tabular interface for specifying experiment variables; the values specified in its design phase automatically flow into the run and analyze phases [49]. Galileo offers a complementary approach, introducing procedural training for those with little-to-no mental model of how experiments work. This paper focuses on whether this new approach enables specific outcomes; whether this creates better experiments than experts is left as future work.

THE GALILEO EXPERIMENTATION PLATFORM

Galileo introduces an architecture for end users to design experiments, get them reviewed by a community, and run them with interested participants. It provides procedural training at different steps, an online collaboration platform, and automated data collection and reminders (Figure 1).

Design-Review-Run: From Intuitions to Investigations

Design an Experiment from an Intuition (Figure 4)

People have many, often poorly-framed, hypotheses. Galileo's design workflow helps people harvest and sharpen them. Examples illustrate possible choices and how they relate; templates provide structure; and embedded videos explicate technical issues. Such procedural training can improve on-task performance [59,65]. A final self-review step provides an overview of the experiment. To keep the platform safe, the primary author receives daily updates of platform activity. The design workflow does not mandate double-blindness or the use of placebo; designers can choose to specify these details.

Review the Design via Feedback from Others (Figure 5)

Experiments require at least two reviews before they can be run. The designer invites the reviewers, who might be online community members, a teacher, or anyone else who can provide useful feedback. Upon receiving reviews, the experiment designer edits the experiment to address the issues found. For our research team's benefit, Galileo logs version changes. Reviewers provide both binary assessment and written responses to specific questions (Figure 5). 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, experimenters may not review their own experiment.

Run an Experiment using Automated Procedural Support (Figure 6,7)

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 [50].

2. Maintaining a per-experiment participant map ([usernames]→ [exp_id]) for anonymity.

3. Collecting and cleaning data (sending data collection messages and reminders at timezone appropriate times, parsing the responses, updating participant and experimenter views).

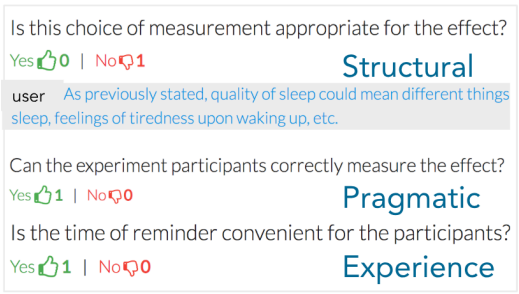


Figure 5: Reviewers walk through an experiment providing binary rubric assessments. A *No* response prompts reviewers to provide concerns and suggestions.

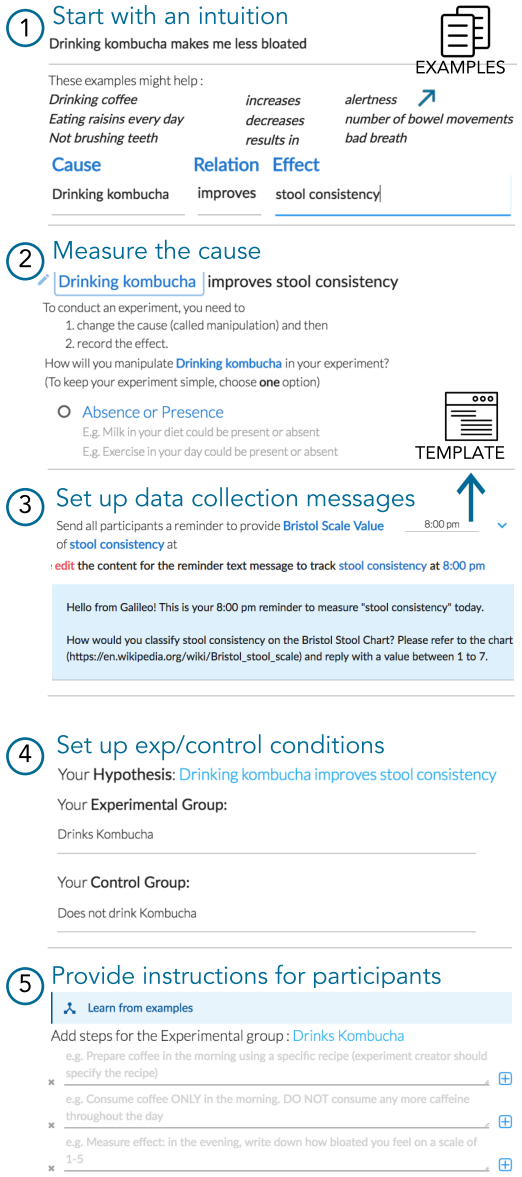


Figure 4: Galileo's design module helps people transform intuitions into experiment designs. It walks people through 1) converting an intuition to a hypothesis, 2,3) providing ways to manipulate/measure cause and effect, 4-5) specifying control and experimental conditions, and (not shown) providing inclusion/exclusion criteria.

4 Informing the experimenter to perform tasks when conditions are met (*e.g.*, setting the start date when enough participants have joined or reminding participants with missing data).
The experimenter's dashboard provides a task list to answer clarifying questions, remind/thank participants, or look at trends in data (Figure 6). Experiments have a minimum participation count but there's no limit to the number of participants who join an experiment cohort. People who sign up after a cohort begins are added to a waitlist.

The primary task for participants is responding to messages from the platform (Figure 7); the current implementation supports email, SMS, and WhatsApp. Self-reports provide the primary data collection mechanism. Participants can optionally answer follow-up questions that capture contextual insights. Galileo logs responses to a MongoDB database. Galileo presents participant data to experimenters using participant ID rather than their real name or username. When the experiment ends, participants receive a summary of the results. Participants can anonymously discuss the experiment at the end, so the experimenter can learn from their feedback.

The Galileo web application uses the Meteor (meteor.com) framework for synchronization, Jade for the front end (jade-lang.com), Materialize for styling (materializecss.com), and Twilio as the text message gateway (twilio.com). Galileo can be used at <URL>; its open source is at <URL>.

Designing the platform

Since its inception, over 80 people have designed and run experiments. 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 used a protocol to gather 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 others' experiments and receive feedback

We provide three examples of how pilot studies informed Galileo's design:
1. Embedded training over videos: Early versions included learning materials provided as lecture videos. However, users' struggles with staying engaged with the learning material and applying it to their context led to embedding learning in the interface.
2. Supporting actionable feedback: For the review interface, early versions only requested binary Yes/No responses similar to popular crowdsourcing platforms; 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.

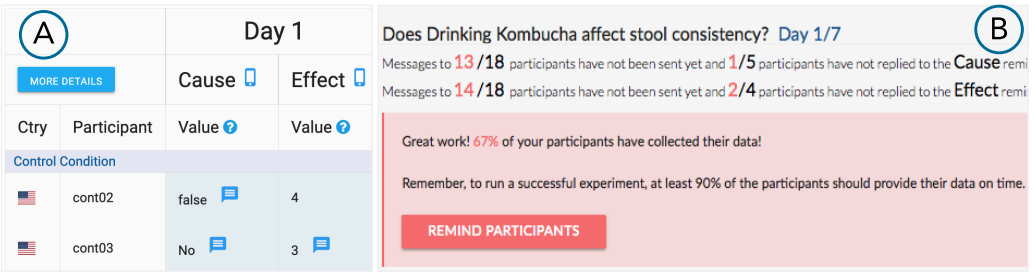


Figure 6: Galileo takes care of many experimenter responsibilities such as random placement of people, sending instructions and reminders, and cleaning and displaying data in both participant and experimenter dashboard. The dashboard enables experimenters to A) remind those with missing data; and B) see participants' data; and clarify questions raised by participants.

3. Ease of glancing at participants' data: Pilot users ran 6 trial experiments. The idea of a runtime dashboard (Figure 6A) came from observing experimenter's difficulty tracking participants' data and sending reminders to those who hadn't added their data. 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.

Integrating Procedural Training in the Design Workflow

Simple examples of procedural learning are activities like tying your shoes, roasting a chicken, or replacing a door handle. Recipes and instructions convey procedures in written form; demonstrations and hands-on learning make it more interactive. Creative tasks differ from rote procedures in that they require people to generate some artifact themselves.

Reduce Complexity by Streamlining, Chunking, & Carrying Context

Learning complex activities overwhelms working memory because of their many interrelated pieces [22]. Recalling work from previous steps and frequent context-switching are especially taxing [26]. Experts mitigate memory demands by integrating multiple elements into conceptual chunks [9]. When software has an explicit model of those relationships, it can help users by maintaining a dependency graph and requesting only needed information. We hypothesize that when novices work with interfaces that explicitly chunk elements for them, they attain more expert-like performance. At this step, we also follow the UI maxim to use familiar language [16,56] so that novices better understand what's being asked of them.

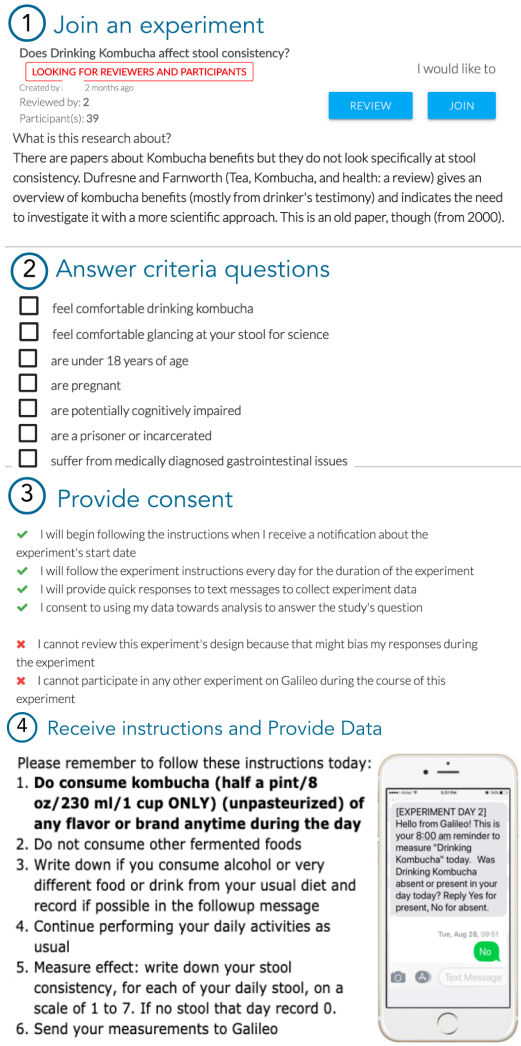


Figure 7: 1) Participants can view a list of experiments. When they elect to join one, they 2) answer inclusion/exclusion criteria, 3) consent to following the provided steps, and 4) receive instructions. 5) Participants receive daily, condition-specific requests, and respond with data and/or clarifying questions.

Embedding Just-in-Time Training

A well-chunked interface can still require knowledge that novices lack. Galileo provides missing knowledge by embedding learning materials in the interface. This *in-situ* embedding has three advantages: it is minimal [7], leverages teachable moments [27], and can be ability-specific [17]. Finally, as is good user interface practice, selecting good defaults for each step helps users see an example of appropriate choices.

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 learning.

Example: Training people to identify a cause

Controlled experiments seek to identify develop causal understanding by varying just the cause in experimental conditions. Many people do not understand the importance of having this minimal-pairs design, perhaps because they do not have the same issues in mind when thinking about the cause as when thinking about the conditions.

Galileo administers the following process to help designers select conditions that test a causal claim. It provides a simple description in common English with ~3 examples showing the data collection reminder text and times right after the designer decides on the cause and effect metrics. Galileo auto-populates text reminders with readable sentences [46] that people can edit. Finally, checklists help people review and improve their work. Such checklists refer to more context-specific challenges of making the experiment simple, safe, and comfortable for participants.

This paper presents three empirical investigations of role-based crowdsourcing with procedural support (Table 1). First, a controlled between-subjects experiment found that structured procedural training yielded significantly higher-quality experiment designs than with lecture videos. Remarkably, Galileo participants garnered higher condition-blind ratings than doctoral students trained in experiment design. Second, a deployment across 16 countries found that people generated structurally-sound experiments on personally-meaningful topics. Third, a field deployment with three communities—kombucha, Open Humans, and beer—across 8 countries demonstrated that communities designed, iterated on, and ran week-long experiments without prior training in experiment design.

STUDY 1: EXPERIMENT COMPARING PROCEDURAL TRAINING TO VIDEOS

Does procedural training help, or might people perform equivalently well without it? To investigate this, a between-subjects experiment tested the following hypothesis: *Structured procedural training yields higher quality experiment designs than learning from lecture videos.*

Procedural learning, when successful, helps people solve unique problems with similar structure. It is perhaps best studied in K-12 mathematics instruction [31]. We hypothesize that participants who carry out interactive procedural training create better experiment designs than those who watch videos on the topic.

Method

Participants were randomly assigned to one of two conditions: *Videos* and *Galileo* (Figure 8). The *Videos* condition provided a playlist of videos about experiment design from a Coursera MOOC that operationalized the specific concepts required for this task [72]. The Galileo condition provided participants access to Galileo. Both conditions provided the same content that contained all the attributes required to create a structurally-sound experiment. Moreover, participants were provided instructions that the resources (videos/Galileo) described the attributes that their designs should possess. Scripted study instructions ensured the same manipulation.

A lab session asked participants to design an experiment for a personal intuition. Participants could start from any intuition that came to mind. A researcher introduced the condition-appropriate material; *Videos* participants wrote their study designs in a Google doc. Participants were told that there was no lower or upper limit on time taken. Each session comprised the following steps: consent, experiment design task, survey, and interview. Participants could also use web

resources, such as Wikipedia and many did. The interview asked participants about confidence in their experiment design abilities and their experience using the system. The interview was tailored to participants' behavior and survey responses: for example, if a participant did not watch some videos, the interviewer enquired why. An independent rater (a professor who teaches experiment design) blind to the conditions rated each participant's experiment using the *Structure* rubric (Table 2).

Structure: 13 points
Hypothesis: 3 points Is the cause/relation/effect specific?
Measurement: 2 points Are the cause and effect manipulated/measured correctly?
Conditions: 3 points Are the control and experimental conditions appropriate? 2pts Do the conditions differ in manipulating the cause?
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

Table 2: Rubric for design-quality criteria for structure

Participants

Recruitment: 72 participants were recruited from a Western US Research University (Table 3). 11 had no prior experience with experiment design; 61 had taken a course or equivalent. Expertise was counterbalanced across conditions.

Measures

The independent variable is access to Galileo/Videos. Dependent variables comprised design quality and time taken to design an experiment. Qualitative measures included how participants used the tool, where they faced challenges, and a post-experiment survey.

Results

A non-parametric Mann-Whitney test assessed the effect of access to Galileo (independent variable) on design quality. Galileo participants created higher-quality experiments (M=11.3) than *Videos* participants (M=5.6); *Mann–Whitney* U=108, $n1$=$n2$=36, p<0.005 (Figure 9A). There was no significant difference in the amount of time participants spent creating an experiment in the *Videos* condition (M=30.8 mins) vs Galileo (M=29 mins), Mann–Whitney U =734, $n1$=$n2$=36, p=0.33 two-tailed. Of the top 50% of experiment designs (36), 29 were from Galileo condition. Galileo participants performed better on five out of six sections (all except hypothesis) of experiment

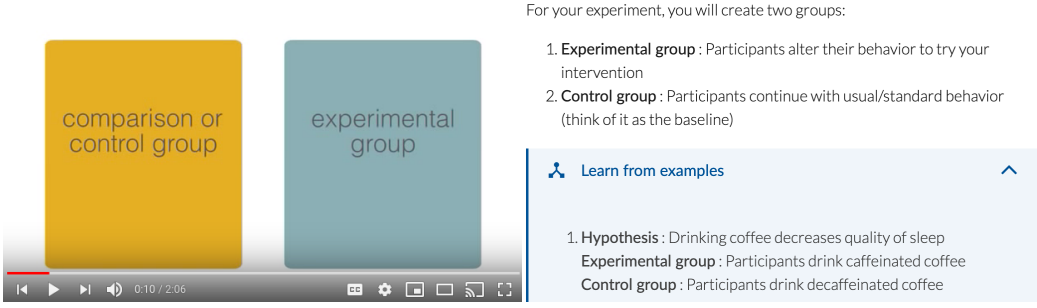


Figure 8: Two conditions for experiment. In the *Videos* condition where participants accessed videos about experiment design. In the *Galileo* condition where participants accessed Galileo tool. Both conditions provided the same content.

design, by a factor of 2. Preliminary analysis found no effects for experiment expertise, so these were excluded from further analyses.

Comparison to experts' design

How similar were the best designs to those that experts create? To assess this, four behavioral-science doctoral students designed experiments for five intuitions selected from the highest-scoring experiment designs in the Galileo condition. These students had prior training in designing and running experiments and were not provided access to Galileo. They designed 15 experiments that were rated by the same rater. Remarkably, Galileo participants created higher-quality designs (M=11.3) than experts without access to Galileo(M=8.9); *Mann–Whitney* U=104, $n1$=15, $n2$=36, p<0.005 (Figure 9B).

Nationality	USA = 37 No Answer = 6	China = 11 Others = 18
Gender	Female = 47	Male = 24
Native English	Yes = 38	No = 34
Age	18-20 = 40 21-25 = 31	26-30 = 1
Ethnicity	Asian/Pacific=36 White = 11	Hispanic/Latino=14 Others = 11
Major	Biology =12 Cognitive Sci=12	Psychology=20 Others = 20
Usedonline learning material	Never = 28 1 class=11	Occasional=16 2-5 classes=12

Table 3: Demography info for 72 participants (all undergraduate students). Some participants did not complete portions of the survey.

Discussion

Because Galileo's goal is to help people design experiments, the dependent variable is the quality of the design; the study does not measure learning gains. Publicly available online resources (like lectures) provide a relevant control condition closer to the real world. Videos participants followed one of two strategies: 1) watch all the videos at once and then begin writing the experiment; and 2) begin designing the experiment and use the videos to fill in the gap when stuck. Like cramming, all-at-once watching floods the mind, making retention difficult. By contrast, the search-when-needed approach interrupts people's flow, replacing the attention on design with a task of locating needed information. The Videos condition's lower score, in conjunction with these observations and the literature, suggest that videos out of context yield a worse learning experience than more contextually-integrated approaches like procedural training. Our observations of Galileo participants suggested that they maintained flow much better.

Participants reported that the videos were slow and the interface provided sufficient examples. Participants in the Galileo condition opened and closed the videos in quick succession. Participants in the *Videos* condition, however, felt that the videos provided a refresher of some concepts they vaguely knew about. Did too much information (e.g. the inclusion of other concepts) in the Coursera course dilute performance? It's possible; accessing the "right" moment in videos is a known research question [36].

Participants in both conditions seemed concerned about their choice of measures for cause and effect. Some participants spent over 15 minutes searching for good measures: one found a formal sleep-quality scale from Stanford researchers. Participants across both

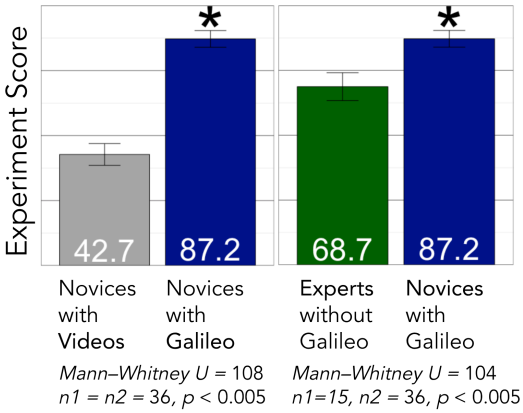


Figure 9A. *Access to Galileo improved the quality of experiment design

B. *Novices with Galileo created better experiment designs than experts without Galileo

conditions mentioned that they enjoyed reflecting on their lifestyle/health ideas and thinking through how to transform an intuition into an experiment. Participants wished that Galileo was integrated with their class, describing it as "hands on" and "DIY."

STUDY 2: PEOPLE DESIGN AND REVIEW EXPERIMENTS ONLINE

The second deployment investigated the quality and nature of experiments; specifically, whether people a) create experiment designs that are structurally-sound, demonstrate insights from lived experiences, and have novel ideas, and b) provide useful feedback on experiment designs.

Method

Participants used Galileo to design their experiments and review others'. Galileo's landing page described why experiments are important and the importance of citizens' contributions towards making discoveries. Upon logging in, participants could design an experiment (see Figure 4), review existing experiments (see Figure 5), or join an experiment (see Figure 7).

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 [51]. 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 [39]. Participation was voluntary and unpaid.

Structure: 13 points
Hypothesis: 3 points Is the cause/relation/effect specific?
Measurement: 2 points Are the cause and effect manipulated/measured correctly?
Conditions: 3 points Are the control and experimental conditions appropriate? 2pts Do the conditions differ in manipulating the cause?
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. discussions on online fora)?
Insightful? Does the hypothesis link to existing science?
Novel
Is there a chance the world will learn something: absence of published research for this question?

Table 4: Rubric for design-quality criteria for Structure, Content, and Novelty

Measures

Measures comprised structure, content, and novelty of experiment designs (Table 4) and usefulness of reviews on experiment design. Two independent raters with training in experiment design rated experiment designs using the following workflow: 1) calibrate: rate three experiments independently and discuss; 2) rate: independently rate all participant generated experiments; 3) combine: discuss ratings where different & develop a common score. High reliability was found between the two raters' measurements (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. Broad or vague hypotheses 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). Some participants edited their original experiment design after receiving reviews from others. 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. The anonymized data is at <URL>.

Results

People Designed Structurally-Sound Experiments, and Drew from Personal Intuitions

The mean score for the experiment was 10.3/13. On average, people scored higher than 75% 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". 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". Common 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 10). Apart from being structurally-sound, the best experiment designs shared two features: they shared a personal experience and linked to known research. For example, a user 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."

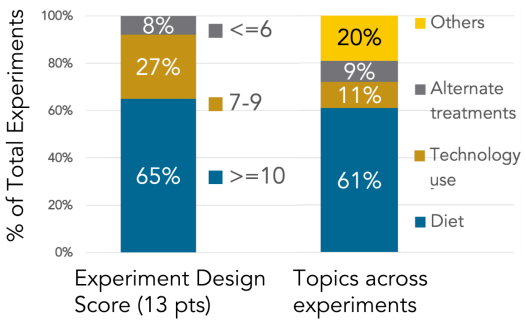


Figure 10: A) Most experiments were structurally-sound, scoring 10/13. B) Most experiments drew from personal experiences

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

Reviewers Use Domain Knowledge to Improve Designs and Advocate for Participant Experience

158 review comments (77%) were rated useful (by the raters); incorporating them would improve the experiment. Average comment length was 140 characters ranging from 3 characters ("yes") to 871 characters (Figure 11B,C). Most comments were direct responses to a

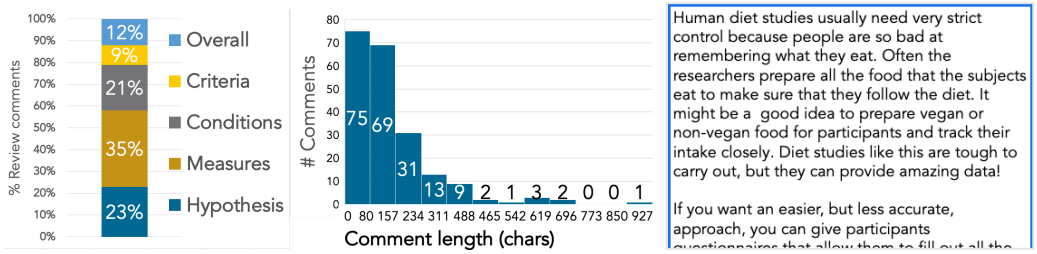


Figure 11: Summary of review A) Review comments were broadly distributed across all components of experimental design. B) Review comments ranged from 3 chars "yes" to 871 char long descriptions. C) The longest review comment described multiple problems with an experimental design while providing numerous actionable suggestions.

rubric question hinting that the review interface helped people focus on the salient parts of an experiment design.

The most common comments sought improving structural correctness (38%) by requesting 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?". Reviewers provided the most comments (54%) about the hypothesis and cause & effect measures.

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.", 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?".

14% of comments demonstrated domain-specific knowledge *E.g.*, one reviewer 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?". A1C refers to the average blood glucose value average levels over the past 3 months that 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 barely drew from their personal experience when suggesting improvements (or at least, did not explicitly mention this was their personal experience). Some comments drew on counterfactual reasoning while thinking about how participants might "hack" an experiment. A comment on an experiment about 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"?)"

STUDY 3: THREE COMMUNITIES DESIGN, REVIEW, & RUN EXPERIMENTS

Three communities—Kombucha, Open Humans, Beer—designed and ran 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 [13]. While there is widespread belief that kombucha "benefits the gut", there is little published empirical evidence for these claims [23]. The 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 [18], 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 [60]. Alcohol helps people fall asleep but disrupts the REM cycle [21]. 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.

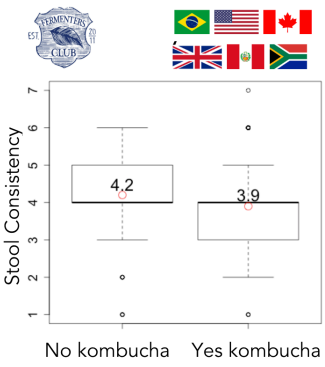
Results

Before the Experiment

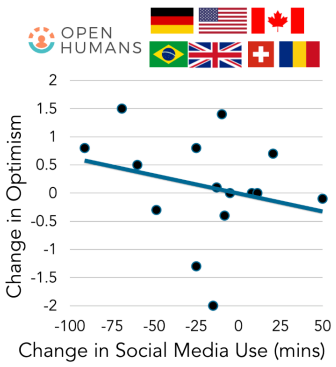
From initial design to launch—37 (kombucha), 13 (Open Humans), and 11 (beer) days elapsed. Each experiment ran for a week.

Design and Review: None of the experimenters had previously designed and run an experiment with people. 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.

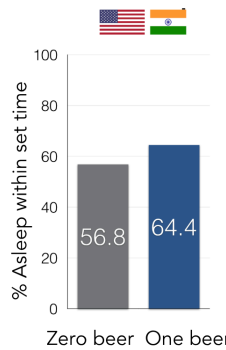
Reviewers provided a total of 104 boolean answers and 32 detailed comments. Comments focused on two themes. First, reviewers helped make the hypothesis and measures more specific; *e.g.*, an experimenter started with the question "Does drinking a beer in the evening help you get to bed on time?"; the 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



Drinking kombucha improves stool consistency N=36; *p<0.03*



Reducing social media use was positively correlated with improving optimism N=15; *r=0.25*



No trend was observed between drinking beer in the evening and time to fall asleep N=17; *p=0.56*

Figure 12: Three communities—Kombucha, Open Humans, Beer—designed and ran experiments; each ran for a week. The flags represent participants' nationality.

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 [71]. 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 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.

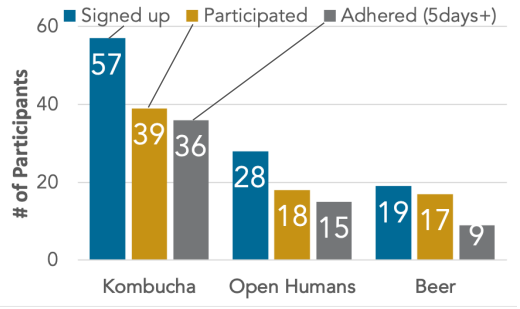


Figure 13: After signing up, fewer people participated in Kombucha (68%) and Open Humans (63%) experiments than Beer (90%). However, those who participated reported greater adherence in Kombucha (92%) and Open Humans (83%) compared to Open Humans (50%). Reasons for non-adherence included being busy, annual leave, and brewers needing to check on the taste of Kombucha.

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 enforcing 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 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.

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 13). 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 Open Humans participants reported high adherence, cutting social media use in half or more (Figure 13). 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 8PM, drinking more than one beer, or not drinking in the drink-one condition (when they were supposed to).

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.

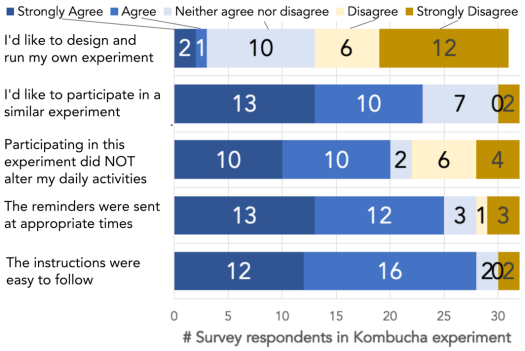


Figure 14: Participants in the kombucha experiment reported an overall positive experience expressing an interest to participant in another similar experiment (23/32). Most found the instructions easy to follow (28/32) and the reminders sent at appropriate times (25/32).

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 14).

DISCUSSION

Finding & Retaining Participants in Citizen Experiments

Two of the three completed experiments were underpowered. These experimenters learned what many scientists know: recruiting participants is time-consuming. Citizen experimenters aren't as ardent about sufficient participation numbers as professional scientists: *p*-values and similar hold much less sway. Citizen experimentation platforms like ours should more clearly convey the importance of getting enough participants; help experimenters estimate what "enough" is; and help recruit participants.

Why might people participate in citizen experiments? Common reasons why people join *expert-led* experiments include [58]: to help find an answer to a question that personally affects them, to gain access to potential treatments, and for credit or monetary compensation. Moreover, the trust placed in institutional researchers might not extend to citizen experimenters [15]. We suggest three remedies: 1) increase trust by sharing more information about the experiment's goals, approximate effort expected, and the experimenter's biography; 2) train experimenters to better publicize them, and leverage participation from communities with already strong ties and common goals; 3) plan for failure and hire twice the required participants by spending more time in participant recruitment. Galileo does not provide experimenters or participants monetary compensation. Consequently, people's motivation is more intrinsic, which has benefits [54] (*e.g.* telling people the importance of their work improves performance [8]), but also empirically shows a high dropout rate. Compensation may help some citizen science experiments. Finally, because

people are more likely to adhere to protocol when it does not conflict with other goals and obligations [34], protocols should be designed to minimize these conflicts.

Balancing the Fun with the Tedious

Both online users and undergraduate students designed structurally-sound experiments. The success of the design workflow shows the extent to which sophisticated tasks can be scaffolded by careful workflow creation. However, it also brings up the tension between the exciting and the tedious parts of knowledge work. *E.g.*, while 90% of online users created a concrete hypothesis, fewer created a minimal-pairs experiment (50%). Generating hypotheses is fun but fleshing it out to a complete experiment design requires motivation and time to both follow multiple steps and to incorporate feedback over multiple iterations. While such tedious details-oriented work may be less appealing than shorter, more creative tasks, it provides a necessary pre-condition for creating scientific knowledge.

Similar concerns show up for participating in experiments. The opportunity to contribute to science is exciting (e.g. kombucha experiment participants mentioned this as a motivation). While changing 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. Perhaps experts are distinct from novices not just in terms of their expertise but also in terms of being a professional and developing techniques to charge through the more mundane and/or difficult aspects of work.

One way to solve this is by distributing the more tedious parts (when possible) among a wider group (e.g. crowd workers can iterate on people's experiment design). Exploring trade-offs of effort, expense, and quality of work among volunteers and paid workers provides an interesting avenue for future work.

Do Citizen Experiments Benefit or Harm Society?

One challenge of modern life is the increasing layers of social and technical infrastructure that separate the creation of knowledge from its everyday use. This divorce makes it difficult to wisely assess and use knowledge. This paper has outlined the positive potential for citizen designed experiments, a greater diversity of perspectives, participation, and understanding. It's worth considering the risks. The primary concern we have is that a poorly designed experiment with a faulty conclusion influences people in fraught ways.

At its best, over time scientific experiments expand human knowledge and correct mistakes when they occur. However, sometimes the popular press reports a headline-grabbing result that is inaccurate, but not the subsequent correction and elaboration. Particularly with science, when ideas are newsworthy but low-quality, people can incorporate misguided ideas in a way that be difficult to dislodge. Perhaps the most notorious example 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. Furthermore, the 20th century offers many examples of pervasively-adopted chemicals (such as lead in paint and gasoline, and asbestos in buildings) that were later found to be toxic. Wakefield's publication linking MMR vaccine to autism (later retracted) was a serial case study [25], 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.) Currently, most readers cannot fully grasp the evidentiary difference between a small

case study and a rigorous controlled experiment. Our hope is that democratizing the doing of science may help the public interpret science news and reduce the risk of leaping to conclusions.

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. There is a community that believes colloidal silver (tiny particles suspended in liquid) to have beneficial properties [45]. 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 [29]. 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" indicates potential legal risks across many places around the world.

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 platform) is an idea with anecdotal benefits but the existence of venom implies non-trivial possibility of self-harm.

CONCLUSION

This paper introduced a crowdsourcing architecture that integrates *role differentiation for experimentation* with *procedural support* using three techniques: 1) experimental design workflow that provides just-in-time training, 2) review with scaffolded questions, and 3) automated checkers that implement standardized behaviors. We demonstrated this approach in the Galileo social computing system.

Three communities used Galileo to design and run structurally-sound experiments. All three drew on lived experiences to create personally-meaningful studies. Finding and retaining participants, providing specific instructions, and improving adherence emerged as key challenges. 54 participants from 16 countries generated 66 structurally-sound experiments; both designers and reviewers shared insights from lived experiences. A between-subjects experiment with 72 participants found procedural training to significantly improve novices' experiment designs over online video lectures. Remarkably, participants using Galileo garnered higher condition-blind ratings than doctoral students trained in experiment design.

In addition to the empirical work described here, Galileo has also been used in an undergraduate Psychology class introducing research methods. We observed two apparent benefits from Galileo's classroom usage. First, learners rapidly designed and reviewed experiment designs. Second, the instructor and Teaching Assistant were able to provide feedback on learners' experiment designs.

Citizen experimentation enables people to test intuitions from their lived experiences to create new knowledge. This distributed knowledge-generation has the potential to diversify the viewpoints and insights represented in scientific knowledge. Also, by exposing more people to the *doing* of science, citizens may gain a deeper and visceral understanding of the texture of scientific knowledge and how it accrues and changes.

REFERENCES

1. Aaron E. Carroll. 2018. Why the Medical Research Grant System Could Be Costing Us Great Ideas. Retrieved from nytimes.com/2018/06/18/upshot/why-the-medical-research-grant-system-could-be-costing-us-great-ideas.html
2. Tim Althoff et al. 2017. Large-scale physical activity data reveal worldwide activity inequality. *Nature* 547, 7663: 336–339.
3. Michael S. Bernstein et al. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*, 313–322.
4. Rick Bonney et al. 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience* 59, 11: 977–984. https://doi.org/10.1525/bio.2009.59.11.9
5. David Boud. 1995. *Enhancing learning through self-assessment*. Kogan Page, London.
6. Carolin Cardamone et al. 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.
7. John M Carroll et al. 1987. The minimal manual. *Human-Computer Interaction* 3, 2: 123–153.
8. Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior and Organization* 90: 123–133.
9. William G Chase and Herbert A Simon. 1973. Perception in chess. *Cognitive psychology* 4, 1: 55–81.
10. Quanze Chen et al. 2018. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. *CoRR* abs/1810.1. Retrieved from http://arxiv.org/abs/1810.10733
11. Michelene T H Chi et al. 1981. *Expertise in problem solving*.
12. Chun-Wei Chiang et al. 2018. Crowd Coach: Peer Coaching for Crowd Workers' Skill Growth. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: 37.
13. Stephanie N Chilton et al. 2015. Inclusion of fermented foods in food guides around the world. *Nutrients* 7, 1: 390–404.
14. Eun Kyoung Choe et al. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*: 1143–1152.
15. Caren B. Cooper et al. 2014. The invisible prevalence of citizen science in global research: Migratory birds and climate change. *PLoS ONE* 9, 9.
16. Seth Cooper et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307: 756–760.
17. Albert T Corbett et al. 1997. Intelligent tutoring systems. *Handbook of human-computer interaction* 5: 849–874.
18. Lorenzo Coviello et al. 2014. Detecting emotional contagion in massive social networks. *PLoS ONE*.
19. Loris D'antoni et al. 2015. How Can Automatic Feedback Help Students Construct Automata? *ACM Trans. Comput.-Hum. Interact.* 22, 2: 9:1–9:24.
20. Steven P. Dow et al. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*, 1013–1022.
21. Irshaad O Ebrahim et al. 2013. Alcohol and sleep I: effects on normal sleep. *Alcoholism: Clinical and Experimental Research* 37, 4: 539–549.
22. Randall W Engle. 2002. Working memory capacity as executive attention. *Current directions in psychological science* 11, 1: 19–23.
23. E Ernst. 2003. Kombucha: a systematic review of the clinical evidence. *Complementary Medicine Research* 10, 2: 85–87.
24. f.lux. 2018. f.lux: sleep research. Retrieved from justgetflux.com/research.html
25. Fiona Godlee et al. 2011. Wakefield's article linking MMR vaccine and autism was fraudulent.
26. Victor M González and Gloria Mark. 2004. Constant, constant, multi-tasking craziness: managing multiple working spheres. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 113–120.
27. Robert J Havighurst. 1953. Human development and education.
28. Andrew Head et al. 2017. Writing Reusable Code Feedback at Scale with Mixed-Initiative Program Synthesis. In *Learning@Scale 2017*.
29. National Institute of Health. 2018. Colloidal Silver | NCCIH. Retrieved from nccih.nih.gov/health/silver
30. Joseph Henrich et al. 2010. Most People are not WEIRD. *Nature* 466, July 2010.
31. James Hiebert and Patricia Lefevre. 1986. Conceptual and procedural knowledge in mathematics: An introductory analysis. *Conceptual and procedural knowledge: The case of mathematics* 2: 1–27.

32. Eric von Hippel. 2005. *Democratizing innovation: The evolving phenomenon of user innovation*. MIT.
33. 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.
34. Ravi Karkar et al. 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.
35. Joy Kim et al. 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 745–755.
36. Juho Kim et al. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*: 4017–4026. https://doi.org/10.1145/2556288.2556986
37. Aniket Kittur et al. 2013. The future of crowd work. In *ACM Conference on Computer Supported Coooperative Work (CSCW 2013)*.
38. Scott R Klemmer et al. 2000. Suede: a Wizard of Oz prototyping tool for speech user interfaces. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*, 1–10.
39. Rob Knight et al. 2016. Gut Check: Exploring Your Microbiome. Coursera. Retrieved from https://www.coursera.org/learn/microbiome
40. Chinmay Kulkarni et al. 2015. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Learning at Scale*.
41. Walter Lasecki et al. 2012. Real-time captioning by groups of non-experts. *Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12*: 23. https://doi.org/10.1145/2380116.2380122
42. Walter S Lasecki et al. 2013. Chorus: A Crowd-powered Conversational Assistant. *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*: 151–162. https://doi.org/10.1145/2501988.2502057
43. Doris Lee et al. 2016. Crowdclass: Designing classification-based citizen science learning modules. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP '16)*.
44. Jeehyung Lee et al. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* 111, 6: 2122–2127.
45. Jayne Leonard. 2016. 15 Reasons You Need A Bottle Of Colloidal Silver In Your Home. Retrieved from naturallivingideas.com/colloidal-silver-benefits-and-uses/
46. Laura Levy et al. 2013. Health Mashups : Presenting Statistical Patterns between Wellbeing Data and Context in Natural Language to Promote Behavior Change. 20, 5: 1–27.
47. Dana Lewis and Scott Leibrand. 2016. Real-World Use of Open Source Artificial Pancreas Systems. *Journal of Diabetes Science and Technology* 10, 6.
48. Ian Li et al. 2010. A stage-based model of personal informatics systems. *Proceedings of the 28th international conference on Human factors in computing systems*: 557.
49. Wendy E Mackay et al. 2007. Touchstone: exploratory design of experiments. *CHI '07 Proceedings of the SIGCHI Conference on Human Factors in Computing System*: 1425–1434.
50. D. W. Martin. 2007. *Doing psychology experiments*. Cengage Learning.
51. Daniel McDonald et al. 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* 3, 3: e00031-18.
52. Daniel McDonald et al. 2018. American Gut: an Open Platform for Citizen-Science Microbiome Research. *bioRxiv*. Retrieved from biorxiv.org/content/early/2018/03/07/277970.abstract
53. Randall Munroe. 2009. Correlation. *XKCD*. Retrieved from xkcd.com/552/
54. UK National Council for Voluntary Organisations. 2018. Why Volunteer? Retrieved from ncvo.org.uk/ncvo-volunteering/why-volunteer
55. D. Neff, G., & Nafus. 2016. *Self-Tracking*. MIT Press.
56. Jakob Nielsen. 1999. *Designing web usability: The practice of simplicity*. New Riders Publishing.
57. Michael Nielsen. 2012. *Reinventing discovery: the new era of networked science*. Princeton University.
58. NIH. 2015. NIH Clinical Trials Research and You. Retrieved from nih.gov/health-information/nih-clinical-research-trials-you/basics
59. Vineet Pandey et al. 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.
60. 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
61. Katharina Reinecke et al. 2015. LabintheWild : Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*.
62. 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.
63. Daniela Retelny et al. 2014. Expert Crowdsourcing with Flash Teams. *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14*: 75–85. https://doi.org/10.1145/2642918.2647409
64. Donald A Schön. 1984. *The reflective practitioner: How professionals think in action*. Basic books.
65. Daniel L Schwartz and John D Bransford. 1998. A time for telling. *Cognition and Instruction* 16, 4: 475–522.
66. Robert Simpson et al. 2014. Zooniverse: observing the world's largest citizen science platform. *... of the*

Companion Publication of the …: 1049–1054. https://doi.org/10.1145/2567948.2579215

67. Divit P Singh et al. 2018. CrowdLayout: Crowdsourced Design and Evaluation of Biological Network Visualizations. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 232:1--232:14. https://doi.org/10.1145/3173574.3173806
68. Kate Starbird et al. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 Proceedings*.
69. Rajan Vaish et al. 2017. Crowd research: Open and scalable university laboratories. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, 829–843.
70. Melissa A Valentine et al. 2017. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3523–3537.
71. Wikipedia. 2018. Bristol stool scale. Retrieved from en.wikipedia.org/wiki/Bristol_stool_scale
72. Jacob Wobbrock and Scott Klemmer. 2018. Designing, Running, and Analyzing Experiments. Retrieved from coursera.org/learn/designexperiments
73. Zooniverse. 2007. Galaxy Zoo. Retrieved December 31, 2016 from galaxyzoo.org