Galileo: Scaling Citizen-led Experimentation with a Procedural Training Platform

Vineet Pandey[1], Tushar Koul[1], Chen Yang[1], Daniel McDonald[2], Rob Knight[2], Scott Klemmer[1]
[1]Design Lab, [2]Department of Pediatrics
UC San Diego, La Jolla, CA
{vipandey, tkoul, chy099, danielmcdonald, robknight, srk}@ucsd.edu

ABSTRACT
Success with complex creative activities requires procedural knowledge (how to do things) in addition to conceptual knowledge (facts). While many resources offer facts, procedural learning is often ignored. This paper introduces procedural training for helping people with complex activities like designing and running experiments. We instantiate this approach in the Galileo socio-technical system for citizen-led experimentation. Galileo chunks tasks, integrates learning, and facilitates iteration. This paper reports on three empirical investigations. First, 40 participants from 7 countries and three communities — kombucha, Open Humans, and beer — used Galileo to design and run week-long experiments. Second, 42 participants from 13 countries generated 50 structurally-sound experiments in an online deployment. Third, a between-subjects experiment with 72 participants found that people created significantly better designs with procedural training than with video lectures. Remarkably, participants using Galileo garnered higher condition-blind ratings than doctoral students trained in experimental design.

Author Keywords
Social computing systems; citizen science; crowdsourcing; online learning.

ACM Classification Keywords
K.3.1. [Computer Uses in Education]: Distance learning, Collaborative learning

FROM EXPERIENCES TO EXPERIMENTATION
People around the world participate in experiments as data donors: browsing online [12], using activity trackers, and joining scientific projects. These distributed data contributions have enabled scientific discoveries and valuable insights on topics including obesity [2], aesthetic preferences [56], sleep [19], and the human microbiome [47].

While professional scientists and commercial ventures run experiments every day, with notable exceptions [10,41], empirical papers from non-professionals are vanishingly rare. This biases the questions asked, studies run, and knowledge created [25]. Currently, both those asking scientific questions and those participating in studies are not representative of the global population. Behavioral science research mainly recruits university undergraduates for participation [25], and medical research funding has supported research that disproportionately benefits some [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. For example, *Lab in the Wild* recruits anyone with an internet connection for behavioral studies [55]; and *All of Us* aims to recruit 1 million Americans from all strata of society (allofus.nih.gov). Building on this, we hypothesize that increasing *experimenter* diversity also expands the gamut of scientific knowledge.

How might online systems support more complex activities that leverage the creativity and diversity of a global community? Enabling diverse, online participants to perform complex activities (like experimentation) shares many challenges with crowdsourcing. One reason for crowdsourcing's focus on brief tasks with a correct answer is such tasks require little training and produce verifiable responses [32]. By contrast, creative activities usually require more expertise, and rarely have a single correct solution [44]. Furthermore, drawing on diverse experiences to make diverse contributions is a *benefit* for creative work, not a problem.

This paper offers three contributions. The first is *integrating procedural training into a social computing system* to enable diverse creative work. The second is an instantiation of this approach in *the Galileo system for designing and running experiments*. The third is *empirical results supporting this approach*.

Baking Expertise into Software
Creative endeavors often rely on a course, degree, or significant life experience. Citizen science systems like Foldit and EteRNA show how carefully-constructed interfaces provide novices with micro-expertise to solve problems that only experts previously could [10,37,38,63]. Our contribution beyond this prior work is a) an expansion to open-ended, creative tasks; and b) greater emphasis on learning materials as a way for participants to succeed on endeavors outside their current expertise.

The closest piece of crowdsourcing work to this paper is Crowd Research where people perform open-ended research guided by professors and graduate students [59]. Our work adopts its strategy of iteration and peer assessment

contributing an interface and system that explicitly integrates procedural training without any expert involvement.

Task Complexity Through Role Differentiation

Social computing systems have demonstrated role-differentiating approaches to creative work. Ensemble introduced a leader and follower approach for collaborative story writing [30]. Flash Organizations tackled collaborative projects like product design by introducing automated hiring for roles, a hierarchy with a central leader, and optional team leaders [60]. Foldit provides a 3D interface for protein folding; the tasks are specified by researchers and solved by crowds [10]. Started by Fields medal winners Timothy Gowers and Terrence Tao, Polymath Project engages mathematicians of varying levels — from university professors to high-school teachers — in proposing and collaboratively solving math problems [13]. Unsurprisingly, many participants already know sufficiently-advanced math to contribute. Experts also provide cachet that inspires novices to participate. For example, Crowd Research organized talks by tech luminaries [59].

Distributed Lead-user Innovation

The power of lead-user innovation is that lived experience, a tight feedback loop, and strong personal motivation can yield different products — and in some cases better ones — than experts [26]. For example, snowboarders improved their binding ergonomics, and diabetes patients have improved insulin delivery [41]. 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 [28].

Lead users have an advantage when the key ingredient is experience intensive; experts retain the advantage for solution-intensive innovations. Professionals have the advantages of training, conceptual knowledge, pre-existing organizational structure for collaboration and support, and direct access to resources like manufacturing. Our hope is that learning and collaboration architectures expand the ways that lead users contribute.

Self-tracking Offers Insights but Not Causality

Personal needs and challenges can be highly motivating [27], inspiring people to measure and seek correlations to improve their lives [50]. However, many self-tracking efforts suffer from structural flaws that prohibit people from actually learning what they'd like to know [8,42]. One frequent error is mistaking correlation for causation [48]. People falsely believe that when one event follows another, the initial event is the cause: *post-hoc ergo propter hoc*. Self-report bias and incorrect analyses can also undermine DIY studies [8,42]. 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. How can we train people in designing and running experiments to answer their personally-meaningful questions?

Creative, Complex Work: What to Do and How to Do It

Prior work has introduced methodologies for instructional designers to scaffold complex-task learning [31]. By contrast, our work presents heuristics and processes. Many genres of real-world knowledge work feature a common structure. For example, experiments feature a cause, an effect, and a way to investigate their relationship. Performing knowledge work in a genre requires both conceptual knowledge (what to do) and procedural knowledge (how to do it). We hypothesize that reifying genre conventions in software offers an on-ramp to creative activities otherwise outside novices' reach.

Feedback and iteration are key to creative success, especially for novices. Feedback can be provided by experts

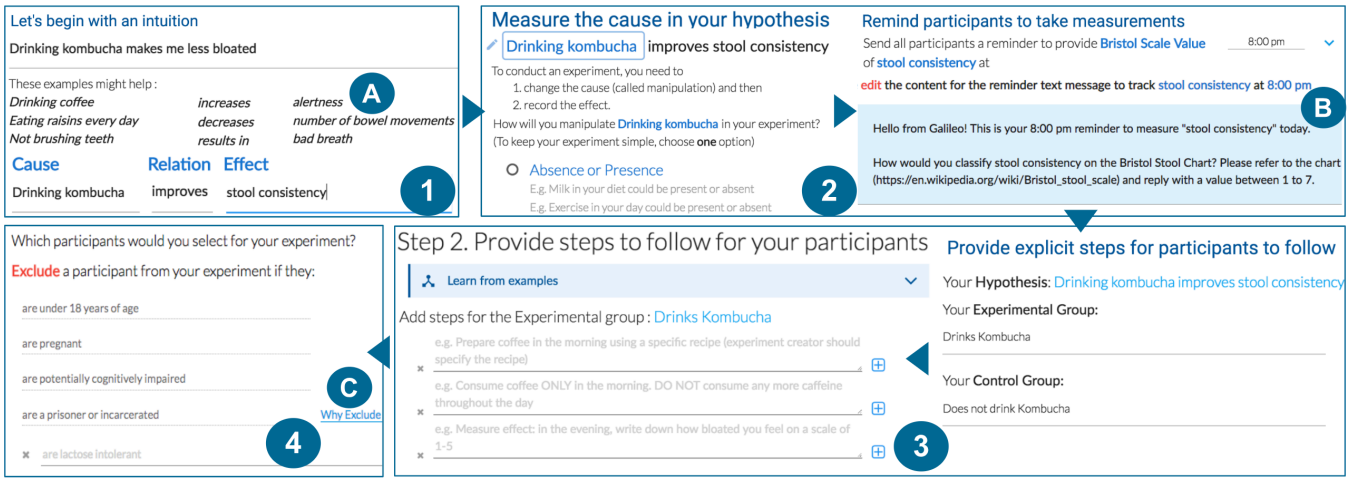

Figure 1: Galileo's design module helps people transform intuitions into experimental designs. It walks people through 1) converting an intuition to a hypothesis, 2) providing ways to manipulate/measure cause and effect, 3) specifying control and experimental conditions, and 4) providing inclusion/exclusion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

[15,57], peers [3,36], software [14,23], or even oneself [3,57]. Feedback can improve both structure and content, and when work is personally-driven, it is especially important to check that someone other than creator can understand it [29]. Consequently, this paper's approach includes an explicit review stage.

THE GALILEO EXPERIMENTATION PLATFORM

Galileo introduces an experimental design workflow 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). 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>.

The archetypical use of Galileo is to create between-subjects designs. Other design types are possible, especially with a bit of creativity. For example, a counterbalanced within-subjects design can be produced by changing the condition-specific instructions halfway through.

Design-Review-Run: From Intuitions to Investigations

Prior work has demonstrated the power of distinct phases for design, testing, and analysis [33]. Most related to this paper, Touchstone introduced a tabular interface for specifying experimental variables; the values specified in its design phase automatically flow into the run and analyze phases [43]. Galileo offers a complementary approach; Galileo introduces procedural training for those with little-to-no mental model of how experiments work.

Design an Experiment from an Intuition

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 (Figure 1A); templates provide structure (Figure 1B); and embedded videos explicate technical issues (Figure 1C). Such procedural training can improve on-task performance [53,58]. A final self-review step provides an overview of the experiment.

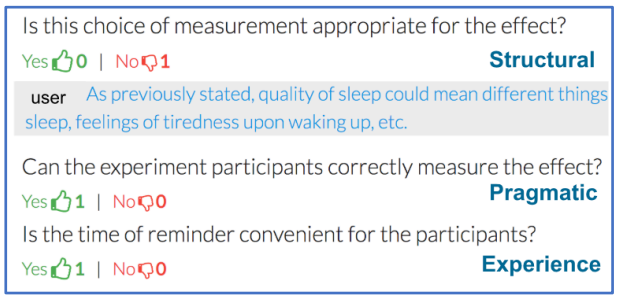

Figure 2: The review module walks reviewers through the experiment. Reviewers provide binary rubric assessments. A *No* response prompts reviewers to provide concerns and suggestions.

Review the Design via Feedback from Others

Galileo requires that experiments receive 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 2). 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

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 [44].
2 Maintaining a per-experiment participant map ([usernames]→ [exp_id]) to maintain anonymity.
3 Collecting and cleaning data (sending data collection messages and reminders at time-zone appropriate times, parsing the responses, updating participant and experi-

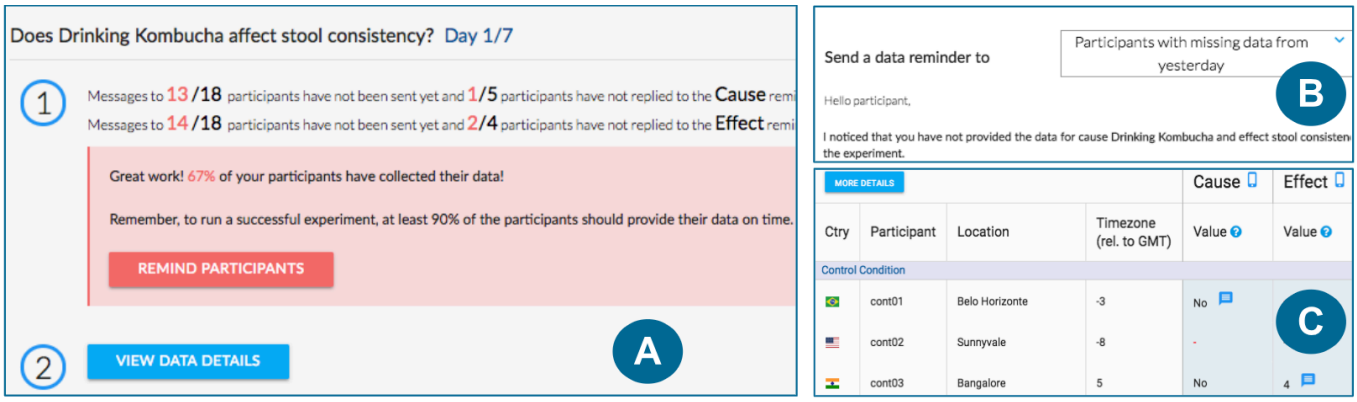

Figure 3: A) The dashboard enables experimenters to clarify questions raised by participants; B) remind those with missing data; and C) see participants' data.

menter views).

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 3). Experiments have both a minimum and maximum participation count. Once the maximum is reached, subsequent volunteers are added to a waitlist.

The primary task for experimental participants is responding to messages from the platform (Figure 4); the current implementation supports email, SMS, and WhatsApp. 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 an summary of the results. Participants can anonymously discuss the experiment at the end, so the experimenter can learn from their feedback.

INTEGRATING PROCEDURAL TRAINING

Simple examples of procedural learning are things 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 something themselves. While larger creative activities can overwhelm novices with their complexity, experts are less daunted because they see the system that undergirds the activity.

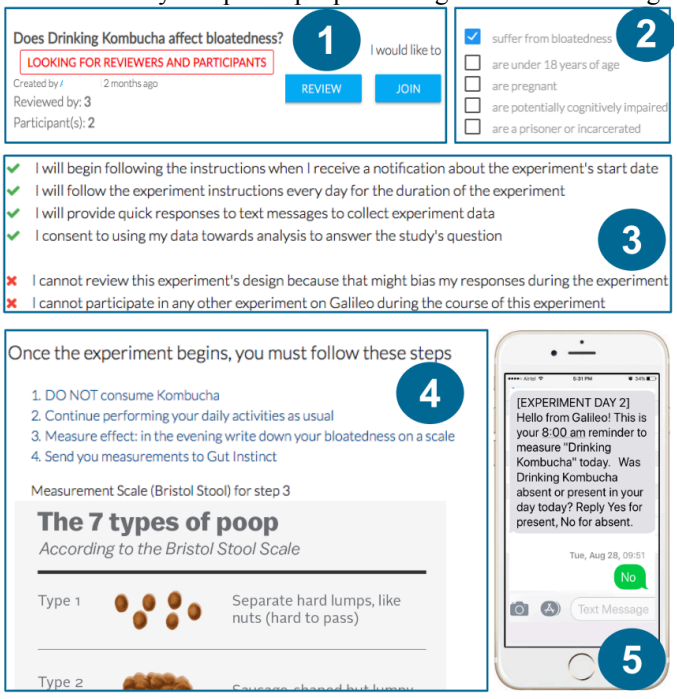

Figure 4: 1) Participants can view a list of experiments. When they elect to join one, they 2) answer inclusion criteria, 3) consent to following the provided steps, and 4) receive instructions. 5) Participants receive daily, condition-specific requests, and respond with data and/or clarifying questions.

Reduce Complexity by Streamlining, Chunking, & Carrying Context

Learning complex activities overwhelms working memory because of their many interrelated pieces [17]. Recalling work from previous steps and frequent context-switching are especially taxing [20]. Experts mitigate memory demands by integrating multiple elements into conceptual chunks [6]. 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 [10,51] so that novices better understand what's being asked of them. When software has an explicit model of those relationships, it can help users by maintaining a dependency graph and requesting only needed information.

Embedding Just-in-Time Learning

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 [4], leverages teachable moments [22], and can be ability-specific [11]. Finally, as is good user interface practice, selecting good defaults for each step helps users see an example of appropriate choices.

In an early version of Galileo, designers sometimes made poor choices. For example, some people listed effects that are difficult to measure. To help guide people, Galileo now presents a short checklist for verifying the choices made in each section, *e.g.*, the reminder "times selected by you are appropriate" for participants. This self-review provides lightweight, just-in-time learning.

Example: Training people to identify a cause

Controlled experiments seek to identify a cause by varying experimental conditions in just one dimension. 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 participants select conditions that test a causal claim. It reduces complexity by 1) providing 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 [40] 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.

To understand the efficacy of Galileo's procedural training, we conducted three empirical investigations of this ap-

proach. First, 40 participants from three communities used Galileo to design and run 3 separate week-long experiments. Second, an online deployment with 42 participants from 13 countries generated 50 structurally-sound experiments. Third, a between-subjects experiment with 72 participants compared this approach to online video lectures.

THREE COMMUNITIES DESIGN & RUN EXPERIMENTS

Three communities — Kombucha, Open Humans, Beer — with diverse goals and structure designed and ran experiments with Galileo (Table 1). Kombucha is a fermented tea drink popular in many parts of the world. Open Humans enables people to contribute personal data (*e.g.*, genetic, social media, activity) for donation to research projects (openhumans.org). Finally, a group of graduate students sought to investigate the effect of alcohol consumption on sleep.

Does drinking Kombucha improve stool consistency? Fermented foods (miso, yogurt, ayran, kefir) have been a staple in many cultures for thousands of years [7]. While there is widespread belief that kombucha "benefits the gut", there is little published empirical evidence for these claims [18]. The experimenter hypothesized that kombucha supplies beneficial probiotics that help maintain normal stool consistency, and designed a within-subjects experiment.

Does using less social media increase optimism? The second experimenter investigated the relationship between social media and mood. Curious about the popular Facebook contagion study [12], 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 [54]. Published papers describe how alcohol helps people fall asleep but disrupts the REM cycle [16]. 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. Kombucha's 37 days included 17 days finding reviewers and 7 running a pilot. 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 experimental 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. The three experimental designs used appropriate measures, provided a minimal-pairs design, tracked confounds, and provided appropriate criteria for participation.

Reviewers provided a total of 104 boolean answers and 32 detailed comments. Reviewer 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 [61].

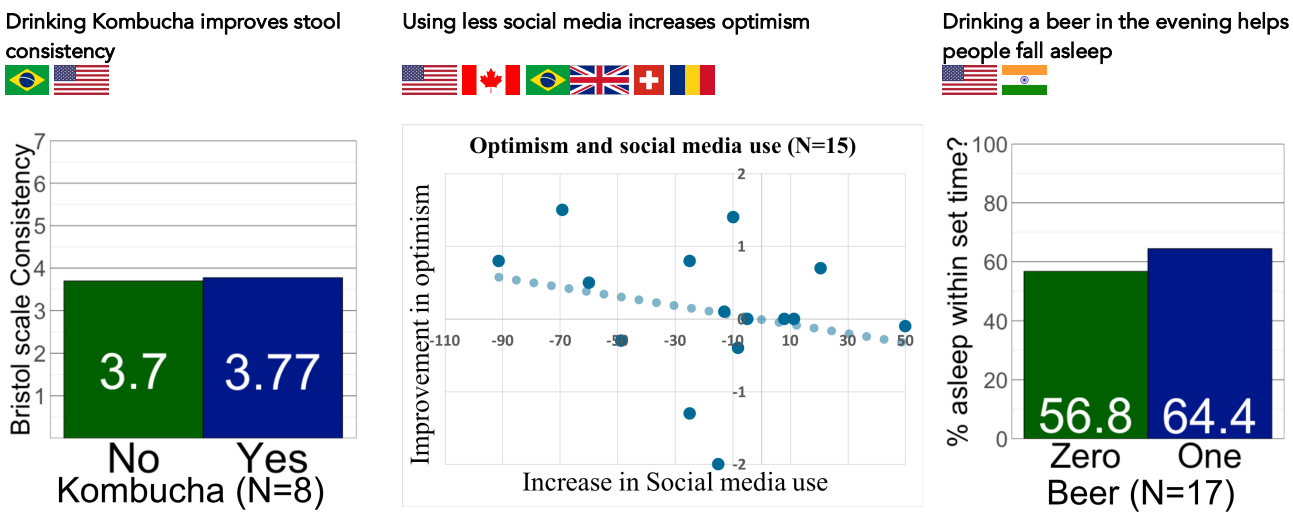

Table 1: Three communities—Kombucha, Open Humans, Beer— used Galileo to design and run experiments; each ran for a week. The flags represent participants' nationality.

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.

Pilots: Two issues emerged in the kombucha pilot. First, people did not want to be randomly assigned to a condition because some were curious to try kombucha for the first time. Institutional science would likely insist on random assignment. Because Galileo participants are volunteers, the experimenter changed the design from between- to within-subjects. Second, people forgot instructions delivered early in the morning, so the experimenter changed the reminder time to noon.

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 the experimenter abandoned the study.

Finding participants: The Kombucha experimenter publicized the experiment on Instagram, 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. 40 participants from 7 countries participated in the experiment.

During the Experiment

Retention: 17 people signed up for the kombucha experiment; 8 completed it (47%). Retention rates were higher for the other two: 63% for Open Humans, 90% for beer. 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: Each day, an average of 54% of participants in the beer experiment followed the condition requirement (drinking 1 or 0 beers by 8PM). 15 of 17 failed to comply on at least one day**.** Non-compliance included drinking wine rather than beer, drinking after 8PM**,** drinking more than one beer, or not drinking in the drink-one condition. Most Open Humans participants demonstrated high adherence, cutting social media use in half or more. Kombucha garnered 76% adherence: 86% for days of no kombucha, and 70% when asked to drink kombucha. 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.

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: Galileo requested that all participants adhere to the protocol as much as possible without harming their health. Participants were instructed to ask the experimenter if confused. Participants' clarifying questions focused on specificity of measurements (*e.g.*, measuring stool consistency once during the day or multiple times; quantifying optimism) and specific lifestyle choices (*e.g.*, consuming probiotics while drinking kombucha?).

Discussion: Fostering Better Experimentation

Three main challenges emerged in citizen-led experiments.

Finding & Retaining Participants in Citizen-led Experiments

All three experiments were underpowered to achieve statistical significance (Table 1). The Open Humans experiment was the most promising, with reduced social media correlating with increased optimism (r=0.25). This result of underpowered experiments shows Galileo's primary weakness.

While scientists build intuitions and test power to estimate the number of participants needed, novices lack both the technical skill and understanding of its importance. 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-led experiment platforms like this should more clearly convey the importance of getting enough participants; help experimenters estimate what "enough" is and help recruit such participants.

Why might people participate in citizen-led experiments? Common reasons why people join *expert-led* experiments include [52]: 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 [9].

We suggest five 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 facilitate posting to niche, interested communities; 3) leverage participation from communities with already strong ties and common goals; 4) plan for failure and hire twice the required participants by spending more time in participant recruit-

ment; 5) following the lead of data journalists [21], convey results through real-world effect sizes such as additional years you'll live.

Galileo does not provide experimenters or participants monetary compensation. Consequently, people's motivation is more intrinsic, which has benefits [49], but also seems to yield a high dropout rate. Compensation may help some citizen science experiments.

Providing Specific Instructions and Measures

Participants' clarifying questions requested more precise instructions. Participants interpreting the same steps differently or being unprepared can undermine the experiment's validity. Instructions that use multiple media, such as images and videos, can increase clarity [45]. Additionally, providing a more-structured means for pilots to give experimenters more detailed feedback to experimenters may yield greater improvement.

Improving Participant Adherence and Data Quality

Non-adherence causes problems in many experiments [44]. There are many potential levers for improving adherence. The first is motivation: telling people the importance of their work improves performance [5]. Experimenters might convey to participants the direct connection of their behavior to the study's ability to deliver, especially when participants seek to learn from the results. Second, because people are more likely to adhere to protocol when it does not conflict with other goals and obligations [28], protocols should be designed to minimize these conflicts.

Third, reducing the friction for participation will likely improve adherence. Different people prefer reminder times and channels. Allowing participants to customize these may improve adherence. For instance, one participant mentioned that the 5AM text with the day's instructions was lost in the plethora of messages she'd see upon waking up; another mentioned that receiving email reminders during workday would work better than text messages.

PEOPLE DESIGN EXPERIMENTS ONLINE

The second deployment investigated the quality and nature of experiments created by people drawing on their life experience. Specifically, we wanted to see whether people create structurally-sound experiments using Galileo that demonstrate personal insights and curiosity.

Method and Recruitment

Participants used Galileo to design their own experiments. Participants were recruited via online publicity. One recruitment focus was people curious about the microbiome because that is a domain where lived experience may inspire intuitions, and the science is nascent [46]. Galileo was promoted on the American Gut Project'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 (*e.g.*, reddit.com/r/soylent), and introduced as an optional activity in assignments on the *Gut Check* Coursera MOOC [34]. Participation was voluntary and unpaid.

Measures

Two raters numerically coded experimental designs for structure and content (Table 2). *Structure* measures whether the design is correct and includes appropriate components. *Content* measures the merit of the idea being tested. The average ICC measure for 2 raters was 0.7.

Results: People Designed Structurally-sound Experiments and Drew from Personal Intuitions

42 participants created 50 complete experiments with 68 versions (*Mdn*=27 minutes). Participants edited their original version after receiving reviews from others. Additionally, 48 experimental designs were started but not completed. The anonymized data is at <URL>.

Structure: The mean structure score for the experiment was 10/13. People scored more than 90% on 10 of 13 measures. Fewer participants created a minimal-pairs experiment (50%), providing clear steps (75%), and providing criteria (75%).

Experimental Content and Novelty

Common experimentation themes included diet (dietary styles, alcohol, fermented foods), medicines and alternative treatments (homeopathy), and health (sleep, pain, gut issues). Unsurprising, personal health and performance were big draws: 90% of experiments sought to improve a person-

Criteria	Operationalized as
Structure	
Hypothesis	**Is the hypothesis concrete? 3 points** Is the cause specific? Is the relation clear? Is the effect specific?
Measurement	**Are the cause and effect properly manipulated/measured? 2 points** Is the cause manipulated correctly? Is the effect measured correctly?
Conditions	**Are the conditions designed correctly? 3 points** Is the control condition appropriate? Is the experimental condition appropriate? Do the conditions differ in manipulating the cause?
Steps	**Are experimental steps clear? 2 points** For Control condition? For Experimental condition?
Criteria	**Are the participation criteria appropriate? 2 points** Are the exclusion criteria correct and complete? Are the inclusion criteria correct?
Run	**Can the overall experiment be run as is? 1 point**
Content	
Novelty	Is there a chance the world will learn something?
Popularity	Is the world already curious about this hypothesis?
Lived experience?	Did the hypothesis come from personal experience?

Table 2: The binary design-quality criteria for structure and criteria.

al health outcome.

86% of experiments came for people's lived experiences; *e.g.*, "eating yogurt makes a person have a more regular bowel movement". 82% of the experiments were rated popular; their hypotheses were discussed on other online for a; *e.g.*, "having dry mouth (or Sjogren's Syndrome) promotes the growth of less beneficial gut microbes".

Raters identified 28% (14/50) experiments as having 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.

Incomplete and Removed Experiments

48 of the 50 complete designs were posted online for others to view; 2 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 [39]. However, the NIH has investigated these claims, finding no benefits. 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 [24]. Galileo offers keyword triggers for alerting the research team of possibly dangerous experiments. For example, an experiment containing "cancer" triggers an email to the research team who then assess the experiment's risk.

In addition to the 50 completed experiments, participants began designing 48 others but did not complete them, presumably due to lack of interest. Half were left at the first step without a hypothesis. Eight experiments were dropped off at setting the data collection notifications and the experimental steps, both of which require using existing templates and examples to create your own details.

Results: Useful Feedback from Personal Experience

Reviewers provided the most comments (54%) about the hypothesis and cause & effect measures. These items come early in the review process, and the assessment is more straightforward than attributes like a minimal-pairs design and confounds.

We saw three main comment types. The most common sought improving structural correctness. For example, "*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?*" The second provided domain-specific knowledge, *e.g.*, "*A1C is measured monthly and won't change after 1g. You mean the BG value?*" for an experiment about Type-1 diabetes patients. The third advocated for participant's experience. Some comments contained multiple insights.

Discussion: Improving Design and Review

Creators need to understand what an experiment can actually measure. Looking at incomplete experiments and talking to potential creators, we realized that people did not always know what makes a testable hypothesis. Specifically, people did not understand that experiments should manipulate the cause. For example, people proposed studying the CRISPR gene-editing tool. This is not something citizens could likely tackle because few have access to these advanced tools. Others required an impractically-long time horizon; *e.g.*, better understanding Alzheimer's Disease. Others chose niche hypotheses. For example, a kombucha fermenter (independent of the one running the Kombucha deployment) wanted to run experiments to improve the fermentation process itself. One area for future work is supporting hypothesis exploration and generation.

Variety of Data Source and Roles

Galileo uses Open Humans to collect data from Fitbit, microbiome, and other data. Some participants sought connections to other resources, *e.g.* psych toolbox. We believe integrating other data sources will be useful.

Dedicated reviewers paid attention to the entire experiment. For instance, one mentioned: "I missed questions regarding the viability of the study (less from a scientist perspective and more from a user perspective): 'How often in the last two weeks would you have been able to perform the required behavior and measurements?'" Enabling such reviewers to directly edit experiments might be useful. By contrast, others provided only a few comments on some experiments. It may be wise to assign time-limited reviewers different chunks and/or concerns.

EXPERIMENT: PROCEDURAL TRAINING VS VIDEOS

The previous section demonstrated that people create structurally-sound experiments using Galileo. Are the experiments strong because of its procedural training, or might people have performed equivalently well without it. To investigate this, a between-subjects experiment tested the following hypothesis:

Structured procedural training yields higher quality experimental designs than learning from lecture videos.

Learning research, especially online, has focused on improving learners' understanding of concepts by providing conceptual material or feedback on their artifacts [35]. Procedural learning, when successful, helps people solve unique problems with similar structure. Procedural learning is perhaps best studied in K-12 mathematics instruction. We hypothesize that participants who carry out an interactive procedural training workflow create better experimental designs than those who watch videos describing the process. One question this raises is whether learners benefit from procedural training when their conceptual knowledge is still nascent.

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
Undergraduate Major	Biology = 12 Cognitive Sci = 12	Psychology = 20 Others = 20
Used online learning material	Never = 28 2-5 classes = 12	Occasional = 16 One class = 11

Table 3: Demography info for 72 participants. Some participants did not complete portions of the survey.

Method

Participants were randomly assigned to one of two conditions: *Videos* and *Galileo* (Figure 5). The *Videos* condition provided a playlist of videos about experimental design from a Coursera MOOC that operationalized the specific concepts required for this task [62]. The Galileo condition provided participants access to Galileo. Both condiitons 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.

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, experimental 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 any videos, the interviewer enquired why. An independent rater (a professor who teaches experimental design) blind to the conditions rated each participant's experiment using the *Structure* rubric in Table 2.

Participants

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

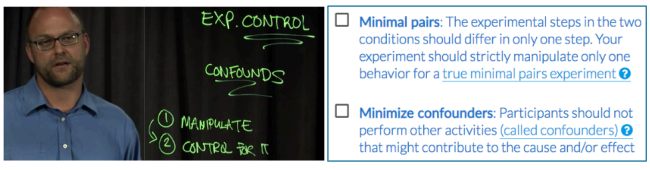

Figure 5: Two conditions for experiment. A) *Videos* condition where participants accessed videos about experimental design from a Coursera class [62] **B) *Galileo* condition where participants accessed Galileo tool**

Measures

Dependent variables comprised design quality (described in Table 2: Structure) and time taken to design the experiment. Qualitative measures included how participants used the tool, where they faced challenges, and a post-experiment survey.

Results

Non-parametric Mann-Whitney tested 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 6A). 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 experimental designs (36), 29 were from Galileo condition. Galileo participants performed better on five out of six sections (all except hypothesis) of experimental design, by a factor of 2. Preliminary analysis found no effects for experimental 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 experimental 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 6B).

Discussion

Both conditions covered the same content.

Videos participants followed two strategies: 1) watch all the videos at once and then begin writing the experiment; or 2)

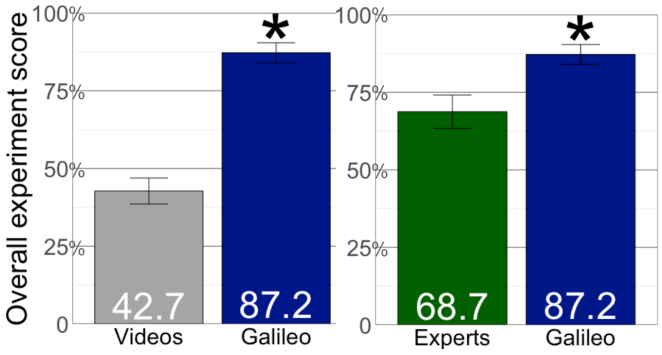

Figure 6A. *Access to Galileo improved the overall quality of experiment design

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

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 is that they maintained flow much better.

Participants across both 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."

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 felt that the videos were slow and the interface provided sufficient examples, *e.g.*, Figure 1A. Galileo participants 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.

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 experimental designs. Second, the instructor and Teaching Assistant were able to provide feedback on learners' experimental designs. The instructor staff suggested supporting study designs beyond between-subjects experiment; *e.g.*, observational studies.

CONCLUSION

This paper introduced procedural training for complex activities like designing and running experiments. We demonstrate this approach in the Galileo socio-technical system for citizen-led experimentation. Galileo chunks tasks, integrates learning, and facilitates iteration.

Three communities used Galileo to design and run structurally-sound experiments. One explored a novel hypothesis; one repeated a published comparison; the third was a personal curiosity. All three experimenters drew on their lived experiences to create personally meaningful studies. 40 participants from 7 countries completed participating in these experiments. Finding and retaining participants, providing specific instructions, and improving adherence emerged as key challenges.

42 participants from 13 countries generated 50 structurally-sound experiments where both designers and reviewers shared insights from lived experiences. Providing diverse data collection support and creating different reviewer roles emerged as avenues of future work.

A between-subjects experiment with 72 participants found procedural training to significantly improve novices' experimental designs over online video lectures. Remarkably, participants using Galileo garnered higher condition-blind ratings than doctoral students trained in experimental design.

Supporting citizen-led 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 more 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. David Boud. 1995. *Enhancing learning through self-assessment*. Kogan Page, London.
4. John M Carroll et al. 1987. The minimal manual. *Human-Computer Interaction* 3, 2: 123–153.
5. Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior and Organization* 90: 123–133.
6. William G Chase and Herbert A Simon. 1973. Perception in chess. *Cognitive psychology* 4, 1: 55–81.
7. Stephanie N Chilton et al. 2015. Inclusion of fermented foods in food guides around the world. *Nutrients* 7, 1: 390–404.
8. 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.
9. Caren B. Cooper et al. 2014. The invisible prevalence of citizen science in global research: Migratory birds and climate change. *PLoS ONE* 9, 9.
10. Seth Cooper et al. 2010. Predicting protein

structures with a multiplayer online game. *Nature* 466, 7307: 756–760.

11. Albert T Corbett et al. 1997. Intelligent tutoring systems. *Handbook of human-computer interaction* 5: 849–874.

12. Lorenzo Coviello et al. 2014. Detecting emotional contagion in massive social networks. *PLoS ONE*.

13. Justin Cranshaw and Aniket Kittur. 2011. The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1865–1874.

14. 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.

15. 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.

16. Irshaad O Ebrahim et al. 2013. Alcohol and sleep I: effects on normal sleep. *Alcoholism: Clinical and Experimental Research* 37, 4: 539–549.

17. Randall W Engle. 2002. Working memory capacity as executive attention. *Current directions in psychological science* 11, 1: 19–23.

18. E Ernst. 2003. Kombucha: a systematic review of the clinical evidence. *Complementary Medicine Research* 10, 2: 85–87.

19. f.lux. 2018. f.lux: sleep research. Retrieved from justgetflux.com/research.html

20. 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.

21. Jonathan Gray et al. 2012. *The data journalism handbook: How journalists can use data to improve the news*. " O'Reilly Media, Inc."

22. Robert J Havighurst. 1953. Human development and education.

23. Andrew Head et al. 2017. Writing Reusable Code Feedback at Scale with Mixed-Initiative Program Synthesis. In *Learning@Scale 2017*.

24. National Institute of Health. 2018. Colloidal Silver | NCCIH. Retrieved from nccih.nih.gov/health/silver

25. Joseph Henrich et al. 2010. Most People are not WEIRD. *Nature* 466, July 2010.

26. Eric von Hippel. 2005. *Democratizing innovation: The evolving phenomenon of user innovation*. MIT.

27. Charlene Jennett et al. 2016. Motivations, learning and creativity in online citizen science. *Journal of Science Communication* 15, 3.

28. 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.

29. Tom Kelley. 2001. *The art of innovation: Lessons in creativity from IDEO, America's leading design firm*. Broadway Business.

30. 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.

31. Paul A Kirschner and Jeroen Van Merriënboer. 2008. Ten steps to complex learning a new approach to instruction and instructional design.

32. Aniket Kittur et al. 2013. The future of crowd work. In *ACM Conference on Computer Supported Coooperative Work (CSCW 2013)*.

33. 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.

34. Rob Knight et al. 2016. Gut Check: Exploring Your Microbiome. Coursera. Retrieved from https://www.coursera.org/learn/microbiome

35. Sean Kross and Philip J Guo. 2018. Students, systems, and interactions: synthesizing the first four years of learning@ scale and charting the future. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 2.

36. Chinmay E Kulkarni et al. 2015. PeerStudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 75–84.

37. 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

38. Jeehyung Lee et al. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* 111, 6: 2122–2127.

39. Jayne Leonard. 2016. 15 Reasons You Need A Bottle Of Colloidal Silver In Your Home. Retrieved from naturallivingideas.com/colloidal-silver-

benefits-and-uses/

40. 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.

41. Dana Lewis and Scott Leibrand. 2016. Real-World Use of Open Source Artificial Pancreas Systems. *Journal of Diabetes Science and Technology* 10, 6.

42. 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.

43. 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.

44. D. W. Martin. 2007. *Doing psychology experiments*. Cengage Learning.

45. Richard E Mayer. 2002. Multimedia learning. In *Psychology of learning and motivation*. Elsevier, 85–139.

46. Daniel McDonald et al. 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* 3, 3: e00031-18.

47. 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

48. Randall Munroe. 2009. Correlation. *XKCD*. Retrieved from xkcd.com/552/

49. UK National Council for Voluntary Organisations. 2018. Why Volunteer? Retrieved from ncvo.org.uk/ncvo-volunteering/why-volunteer

50. D. Neff, G., & Nafus. 2016. *Self-Tracking*. MIT Press.

51. Jakob Nielsen. 1999. *Designing web usability: The practice of simplicity*. New Riders Publishing.

52. NIH. 2015. NIH Clinical Trials Research and You. Retrieved from nih.gov/health-information/nih-clinical-research-trials-you/basics

53. 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.

54. 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

55. 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*.

56. 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.

57. Donald A Schön. 1984. *The reflective practitioner: How professionals think in action*. Basic books.

58. Daniel L Schwartz and John D Bransford. 1998. A time for telling. *Cognition and Instruction* 16, 4: 475–522.

59. 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.

60. 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.

61. Wikipedia. 2018. Bristol stool scale. Retrieved from en.wikipedia.org/wiki/Bristol_stool_scale

62. Jacob Wobbrock and Scott Klemmer. 2018. Designing, Running, and Analyzing Experiments. Retrieved from coursera.org/learn/designexperiments

63. Zooniverse. 2007. Galaxy Zoo. Retrieved December 31, 2016 from galaxyzoo.org