UNIVERSITY OF CALIFORNIA SAN DIEGO

Citizen-led Work using Social Computing and Procedural Guidance

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Vineet Pandey

Committee in charge:

Professor Scott R. Klemmer, Chair
Professor James D. Hollan
Professor Rob Knight
Professor Donald A. Norman
Professor Laurel D. Riek

2019

The dissertation of Vineet Pandey is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2019

DEDICATION

To the adventure of life

EPIGRAPH

Whenever you are in doubt, or when the self becomes too much with you, apply the following test. Recall the face of the poorest and the weakest person whom you may have seen, and ask yourself, if the step you contemplate is going to be of any use to them.

Mahatma Gandhi

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

This dissertation has been a long, exciting journey forming questions, finding answers, and just figuring things out. Along this terrific journey, I have been lucky to enjoy the support of many people and institutions. These words of acknowledgement are just some rushed thoughts; my gratitude goes way broader and deeper...

Ideas are cheap, implementation is expensive. I want to thank NSF-IGE grant, a Google Research Award, and a gift from SAP for supporting this dissertation research. Summers spent elsewhere were supported by Microsoft Research and a European Research Council Grant. Thank you UCSD Computer Science department for providing confirmed funding to students for the first year; this dissertation is a product of my broad explorations during that period. I express my sincere gratitude to everyone who keeps *things* running. Vanessa, Sara, Nina, Teenah, Ian, Olga, Clint, student assistants, and the broader Lab Ops team: your constant support made it easier to focus on research. Thank you Jennifer, Emily, Alan, and Vidula for your seamless support with reimbursements and more. Thank you, Michiko, for being so patient at one of the toughest tasks I know: managing Rob's schedule. Thank you Julie for the calm support you've provided me (and hundreds of CSE PhD students) over the years. I salute the UCSD Human Research Protection Program (HRPP) for their critical support of this and other research.

This dissertation brings together ideas from different universes. Scott Klemmer guided me through this process with both substance and style. When we first started working together, Scott made his expectations clear: do. important. research. Over the years, I've learned that success is a daily grind; *overnight* successes take years in the fermentation jar. If you've enjoyed using this dissertation's systems or reading the papers, you have also gained from Scott's wisdom about research, his feedback on my writing, and his interaction design sorcery. With Scott's encouragement, I've transferred many principles between research and life; here are three: 1) look outwards, 2) do

something, and 3) be empirical.

Years from now, I'll realize just how special my dissertation committee was. At every meeting, Rob inspired me with his dedication, knowledge (of microbiome and beyond), and argumentation. Before starting grad school, I loathed Don's book: it blamed the designer for users' mistakes. I stand corrected. Don's push to *start with people* has been life-changing for this engineer-turned-researcher. Jim's positive disposition, willingness to share his time, and a high tolerance for puns gifted me with a calm confidence. And if Jim could spend his mornings working in the lab, so could I! Professor Laurel Riek's excitement and probing questions about my research have both motivated and tested me. Thank you committee for the inspiration and the support.

Guiding students made me more responsible and taught me new techniques. Aliyah, Brian, Chen, Cody, Crystal, Dingmei, Hedy, Kaung, Liby, Orr, Rachel, Robert, Senyan, Tushar: Thank you and I hope I have been helpful. I'm around. Microbiome experts—Tomasz, Justine, Embriette, Daniel, and Amnon—helped me through the challenges of doing careful science. Community leaders and volunteers—Adriana, Ariel, Austin, Bastian, Mad—piloted and used my platform. Many deserve blame for introducing me to HCI research: Laura (my first and last HCI TA) suggested I might be decent at this stuff; Chinmay mentored my first project about learning and social computing; and Catherine introduced me to research design. Adam (Rule) is the oracle who transforms my stupid questions to coherent ideas. The jump from systems to HCI research was big, scary, and exciting; Tianyin's passion for sysconfig ideas eased the transition. Early on, Yasmine patiently showed me the ropes of MOOC research. My interest in research crystallized over multiple internships during undergrad. Sid Jaggi (Chinese University of Hong Kong), SS Rao (Seoul National University), and Bimal Roy (Indian Statistical Institute, Kolkata) supervised, inspired, and supported me even when I had little to show for my research efforts. Arvind Arasu and Krish Chatterjee were

always around to discuss ideas when they hosted me for a summer each. To all my collaborators, mentors, and the few thousands of people who tried this dissertation's research systems: *Thank you*!

The Design Lab has supported me in thinking unfettered by disciplinary boundaries. Thank you Eli, Eric, Lilly, Philip, Steven for feedback and alternate takes on my research. Thank you Michele for the banter and laughs. Thank you Ariel, Tricia, Ailie, crewtoms for providing feedback, sharing our grad school challenges, and for doing sushibooch(a). Ailie's flagrant generosity has benefitted me over many years now: my drafts would have the many more errors otherwise. You have been the light in our tiny, dank office. Colleen, Lars, and Deborah encouraged me to a fault. Research chats around town w/Derek, Seth, and the Shermanator are some of my favourite memories. I've received many thoughtful reviews from CHI, CSCW, and UIST reviewers. I've deeply appreciated your questions and insights; let's make great reviews the norm. UCSD has some incredibly creative minds; being in their vicinity is reason enough to try harder.

What's life without friends? Karteek and Soumyadeep listened to my complaints and shared their grad school wisdom. Vicky and Eduardo dealt with messy kitchen sinks, undergrad friends—WING, Gaonwaale, PCr, more— made fun of me, and grad school friends—Varsha, Radhe, Skanda, Saman, Moein, Pushp, Pragya, Rangmaster, Anupriya, Arjun, John, Manish, Mario, Dimo, Marcela, Akshay, Val, Tateda, Niki—reminded me of life's potential with interesting conversations, movies, music, food, and more. Our differences pale in comparison to our similarities. San Diego is a special place; its residents Martha and Lori deserve special applause for dealing with this tired grad student and introducing him to special places like Barrio Logan and OB. Through counseling opportunities offered by the university and supported by insurance, I've learnt plenty about myself. Feeling joyful should not be a privilege.

Culture creates the first design, the rest are just edits. My deepest regards are

reserved for my family. Thank you for living in ways many find worthy of emulating. You've blessed me with the freedom to follow some elusive light that I do not understand and cannot explain. As they say in California: *Namaste*.

CHAPTER 3, in part, includes portions of material as it appears in *Gut Instinct: Creating Scientific Theories with Online Learners* by Vineet Pandey, Amnon Amir, Justine W. Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, and Scott R. Klemmer in the proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2017). The dissertation author was the primary investigator and author of this paper.

CHAPTER 4, in part, includes portions of material as it appears in *Docent: Transforming Personal Intuitions to Scientific Hypotheses through Content Learning and Process Training* by Vineet Pandey, Justine W. Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, and Scott R. Klemmer in the proceedings of the ACM Conference on Learning at Scale (L@S 2018). The dissertation author was the primary investigator and author of this paper.

CHAPTER 5, in part, includes portions of material as it appears in the paper under preparation *Galileo: Procedural Support for Citizen Experimentation* by Vineet Pandey, Tushar Koul, Chen Yang, Daniel McDonald, Mad Price Ball, Bastian Greshake Tzovaras, Rob Knight, and Scott R. Klemmer. The dissertation author was the primary investigator and author of this paper.

VITA

2011	B. Engineering in Computer Science, Birla Institute of Technology & Science, Pilani, India
2016	M.S. in Computer Science, University of California San Diego
2019	Ph. D. in Computer Science, University of California San Diego

PUBLICATIONS

Vineet Pandey, Amnon Amir, Justine W. Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, Scott R. Klemmer. "Gut Instinct: Creating Scientific Theories with Online Learners", *ACM Conference on Human Factors in Computing Systems (CHI)*, 6825-6836, 2017.

Vineet Pandey, Justine W. Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, Scott R. Klemmer. "Docent: Transforming Personal Intuitions to Scientific Hypotheses through Content Learning and Process Training", *ACM Learning at Scale*, 9:1-9:10, 2018.

Daniel Mcdonald, Rob Knight, Vineet Pandey, Scott R. Klemmer, American Gut Consortium. "American gut: an open platform for citizen science microbiome research", *American Society for Microbiology* mSystems, 2018.

Vineet Pandey, Tushar Koul, Chen Yang, Daniel McDonald, Mad Price Ball, Bastian Greshake Tzovaras, Rob Knight, Scott R. Klemmer. "Galileo: Procedural Support for Citizen Experimentation", *In Preparation*, 2019.

Gerth S. Brodal, Mark Greve, Vineet Pandey, S. Srinivasa Rao. "Integer Representations towards Efficient Counting in the Bit Probe Model". *Journal of Discrete Algorithms*, 2014.

Tianyin Xu, Vineet Pandey, Scott Klemmer. An HCI View of Configuration Problems. *arXIv.org*, 2015.

Catherine M Hicks, Vineet Pandey, C Ailie Fraser, Scott R. Klemmer. "Framing Feedback: Choosing Review Environment Features that Support High Quality Peer Assessment". *ACM Conference on Human Factors in Computing Systems (CHI)*, 2016.

Arvind Arasu, Ken Eguro, Raghav K., Donald Kossmann, Pingfan Meng, Vineet Pandey, Ravi R. "Concerto: A High Concurrency Key-Value Store with Integrity". *ACM SIGMOD*, 2017.

ABSTRACT OF THE DISSERTATION

Citizen-led Work using Social Computing and Procedural Guidance

by

Vineet Pandey

Doctor of Philosophy in Computer Science

University of California San Diego, 2019

Professor Scott R. Klemmer, Chair

Online platforms enable people to interact with friends, family, and the world at large. How might people go beyond sharing stories and ideas to building and testing theories in the real world? While many are motivated to dig deeper into their lived experience, limited expertise and lack of platform support make complex activities like experimentation dauntingly hard. Novices benefit greatly from expert guidance: this thesis advocates baking the guidance into the interface itself.

This dissertation introduces *procedural guidance* to build just-in-time expertise for difficult tasks. Procedural guidance has multiple advantages: it is minimal, leverages teachable moments, and can be ability-specific. This dissertation instantiates this insight

of procedural guidance through a sequence of increasingly complex social computing systems: *Gut Instinct* for curating ideas, *Docent* for generating hypotheses, and *Galileo* for citizen-led experiments.

Gut Instinct hosts online learning materials and enables people to collaboratively brainstorm potential influences on people's microbiome. *Docent* explicitly teaches people to create hypotheses by combining personal insights and online learning with task-specific scaffolding. Finally, *Galileo* reifies experimentation in the software, provides multiple roles for contribution, and automatically manages interdependencies. Multiple evaluations—controlled experiments and field deployments with online communities including American Gut participants—demonstrate that procedural guidance enables people to transform intuitions to hypotheses and structurally-sound experiments. By enabling people to draw on lived experience, this dissertation harbingers a future where people can convert their intuitions to actionable plans and implement these plans with online communities. This dissertation concludes by discussing opportunities for complex work using social computing platforms.

Chapter 1

Social Computing for Complex Work

Social computing platforms enable people to connect and share but provide little to no support for deeper work. This dissertation provides a novel approach for complex work by introducing procedural guidance in social computing. Systems instantiating this approach integrate theories from learning, collaboration, and interface design. Supporting people in personally meaningful activities such as generating and evaluating scientific theories provides multiple advantages: people can answer their own questions, the world can learn something new, and our future society can potentially have more diverse stakeholders and contributors.

Social computing platforms have revolutionized how people connect, communicate, and share. Friends and family stay in constant touch about both significant and mundane life events. Strangers from around the world discuss ideas about topics of mutual interest. Increasingly, these opportunities to connect and share have also translated to more active doing: people fund ideas that traditional business places might balk at [74]. Others have used social platforms to bring attention to important political, and economic questions [156, 55]. By altering how we communicate and chat, social platforms have cemented a central place in our professional, personal, and leisure activities. However, the benefits of social computing are not distributed equally.

Every internet user has a voice but some amplify them better than others: people's

online informaking seeking behavior is a significant predictor of their existing social capital [60]. Widely accessible research papers and articles provide useful learning materials; however, evidence-based rational discourse on online fora are an exception, not the norm. Social computing platforms have vastly succeeded at keeping people engaged with careful interface design but they barely support *citizen-led enquiry*. These examples suggest that simply connecting people is not enough for successful citizen-led work. Absent tools to build expertise, social computing seems less of a transformative panacea. How might people succeed at their goals using online systems?

People have strong personal motivations and contextual insights; they possess a remarkable ability to identify patterns and create theories from their lived experiences [59]. While people have an amazing breadth and depth of ideas, they lack the expertise to implement these ideas. To create knowledge, they need mental scaffolds for organizing complex work, domain knowledge to compose and execute the steps, and ways to ask for help. Experts benefit from conceptual knowledge, professional training, pre-existing organizational structure for collaboration, and direct access to resources. Currently, citizens lack these resources. By performing personally meaningful work, people can answer their own questions and their ideas can catalyze creating new knowledge. As a result, both individuals and the world at large have missed out on this this opportunity for useful work. How might online systems support citizen-led knowledge work?

This dissertation advances the design of social computing systems by integrating learning and collaboration for complex work such as generating and evaluating scientific theories. Over 600 people from 30 countries have self-organized to generate theories about the human microbiome and test them by running experiments. *Gut Instinct* emodies this insight and introduces a collaborative citizen science platform for people to transform lived experience into scientific theories. This dissertation raises the question:

how can global communities create knowledge that meets their goals without waiting for experts to lead?

1.1 Goal: Scale and Deepen Citizen Contributions

People's curiosity and needs provide endless possibilities to perform useful work. People design, build, and track to better understand and improve their health [41]. However, traditional social computing designs don't funnel this motivation to useful work.

On numerous online fora, people share their intuitions, observations, folk theories, and even results from trying different approaches to improve their health, e.g. from simple ideas like "giving up drinking coffee to improve quality of sleep" to testing different dietary approaches. Current online forum designs prioritize discussion—sharing personal details in long, free-flowing text—over goal-directed structure and learning [155]. While learning resources like Massive Open Online Courses (MOOCs) abound, they adopt traditional classroom pedagogy through lectures focused on conceptual knowledge. Synthesizing the social computing literature highlights three challenges: poor signal-to-noise from crowds due to lack of training [44]; inefficient collaboration with-

Figure 1.1: The Gut Instinct platform enables anyone to transform their intuitions to hypotheses and then design and run experiments to test them [131, 132, 134]. Gut Instinct integrates conceptual learning embedded via short lectures and software-guided procedural learning to enable designing and reviewing experiments. Participants from around the world join experiments, follow instructions, and provide data in response to automated data collection reminders.

out careful attention [138]; and poor results (or no results at all) unless experts lead. Desirable social computing techniques will reliably enable a wide variety of people to contribute more than they naturally could and manage the dependencies among a large set of tasks.

1.1.1 Challenge: Designing support for collaboration and knowledge requirements

While people are connected online and collectively have access to many resources, they need ways to collaborate effectively. In a large distributed community, there's often someone who happens to have important relevant knowledge, usually drawing on a relevant but distant domain. Such distributed efforts are a type of lead-user innovation [162]. Having many people work on the same problem increases the odds that one will break through. Drawing on secondary expertise as inspiration can be an important agent of creativity because almost by definition, the combination is rare [14]. While many hands make light work, novices need clear contribution opportunities.

Citizens have a different background than professional scientists; they have unique personal experiences but lack the years of domain training. Novices are also *uninfected* by all the knowledge that enables experts to innovate. 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.

To summarize, supporting complex work with social computing requires two key features: 1) careful collaboration primitives, and 2) procedural support for deeper individual contributions.

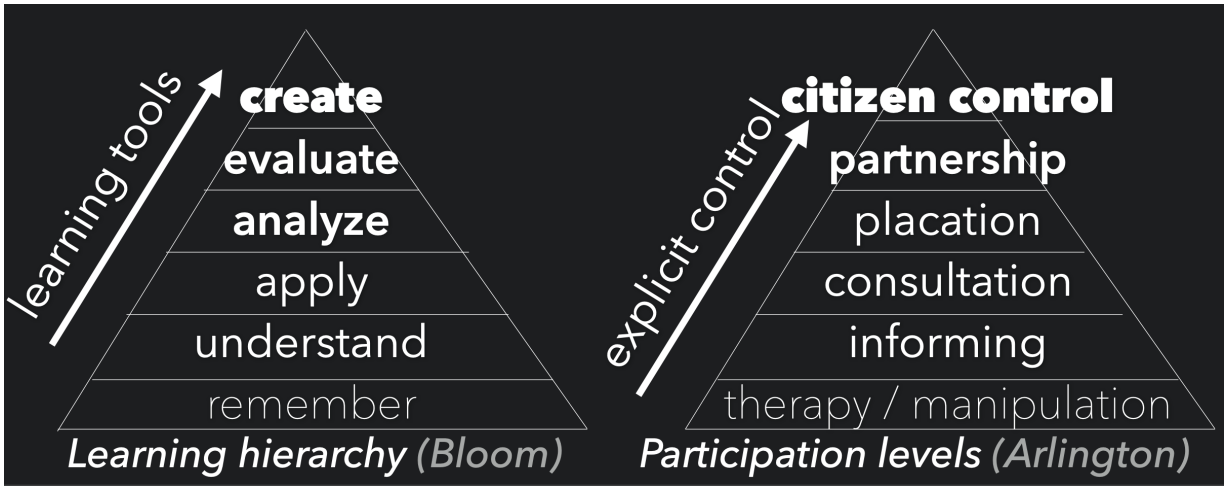

Figure 1.2: Maslow's hierarchy of learning provides a research trajectory for learning tools. Arlington's hierarchy of contribution provides goals for social computing systems

1.1.2 Scientific experimentation: An instance of complex knowledge work

Many people are interested in understanding and improving their health. Millions of peple from all over the world share their insights. Why haven't people run experiments to test these ideas? Scientific experimentation features technical requirements and contextual choices that are inscrutable for a lay individual yet necessary for success [118]. While professional scientists and commercial ventures run experiments every day, with notable exceptions [34, 41], empirical papers from non-professionals are vanishingly rare. This biases the questions asked, studies run, and knowledge created [72]. People have questions about their health, but lack the expertise and resources to scientifically investigate them. Broadening the pool of experimenters could help people investigate their curiosities, develop solutions to improve health and performance, and assist institutional researchers. To create computational systems that leverage people's strengths and mitigate the lack of training, this dissertation focuses on scientific domains that are nascent, highly contextual, and personally motivating.

1.2 Thesis Statement and Contributions

This dissertation investigates the question: how might online platforms enable people perform complex work that is personally meaningful work? Underlying these investigations is the thesis:

> *Procedural guidance in social computing catalyzes personally meaningful & useful scientific work*

This dissertation's primary contribution is the idea of intergrating learning in social computing tfor groups of novices to perform complex, creative activities. The thesis achieves this integration by building a sequence of interactive prototypes that enable people to collaboratively generate and test hypotheses. In the process, the prototypes divide complex work into distinct activities: self-sourcing the design and crowdsourcing people's inputs and data. Every prototype advances social computing further as a domain for deep, personally meaningful work. Beyond introducing learning abstractions, this dissertation carefully designs the affordances in the system to enable different users for different needs.

This dissertation makes three types of contributions: theoretical perspectives / techniques, real-world systems, and outcomes including empirical results.

1.2.1 Theoretical techniques

Improving work quality in social computing suggests deepening individual contributions and broadening participation by providing different contribution mechanisms. The former requires better learning tools and the latter requires better collaboration tools and dependency management. Consequently, this thesis' theoretical contributions include 1) principles to integrate guidance for complex tasks, and 2) ways to divide complex tasks into multiple roles or affordances.

Principles to integrate learning in social computing: Learning broadly comprises conceptual (declarative) and procedural knowledge. Conceptual learning—the primary focus of classroom teaching—involves understanding and interpreting concepts and the relations between concepts [9]. In contrast, procedural learning teaches "action sequences for solving problems" [140]. To contribute usefully, people need to have a good working model of both the concepts and procedures for an activity. This dissertation enables knowledge acquisition in two ways: 1) reifying conceptual bits in the software; and 2) providing procedural guidance with examples, checklists, and templates. These techniques aim to move individual contributions to higher levels on Maslow's hierarchy (Figure 1.2). Specifically, *Docent*'s Learn-Train-Ask workflow (Chapter 4) embeds procedural guidance about hypotheses while *Galileo*'s design workflow (Chapter 5) demonstrates the efficacy of reifying the genre of between-subjects experimentation in the software itself.

Ways to decompose a complex activity: Individuals, groups, and machines possess complementary strengths. Individuals might be highly motivated on a topic of personal interest; their lived experiences provide ideas that might be potentially novel. Groups of people might be less motivated but provide another set of eyes and complementary knowhow and insights from lived experiences; their efforts can help check slips from the creator and provide novel inputs. Finally, computers can implement things consistently to reduce biases but they cannot interpret open-ended instructions fairly in different contexts the way people can. Given these complementary strengths and limitations, this dissertation suggests the following process: 1) support an individual in creating an initial design on a personally meaningful topic; 2) improve it with others' feedback to create an actionable plan; and 3) implement the plan with the help of automated software. For all these steps, the system manages the interdependencies. This dissertation aims to push collaboration up Arlington's hierarchy (Figure 1.2).

1.2.2 User interface and system design

When trying to leverage guidance and roles for complex citizen work, two challenges emerge. First, the internet is a diverse community: people might have poor models of reasoning and frame their ideas and intuitions in weird ways. The software should be understandable to a broad range of participants. Second, complex activities are challenging: the interface should make it easy for people to focus on the current step and provide resources to get unstuck. Designing such a system requires walking a fine line: too much information might overwhelm people while too little make people struggle. These techniques need to be baked in simple, interactive interfaces.

To demonstrate the efficacy of these ideas, this dissertation introduces *Gut Instinct*, a collaborative citizen science platform for people to transform lived experiences into scientific theories. *Gut Instinct* frames the task of hypothesis-testing as a crowdsourcing problem, develops techniques and platform that supports different roles with just-in-time learning, and provides efficient backend support to automate simple tasks. *Gut Instinct* divides multi-party collaboration into complementary tasks and supports them using different contribution mechanisms (like adding a question, editing a response) and roles (like experimenter, reviewer, participant). This provides people the flexibility to choose how much they'd like to contribute. Finally, *Gut Instinct* automatically manages multiple activities to reduce bias and experimenter/participant workload, such as randomized placement of people into conditions, maintaining anonymity, and collecting and cleaning data.

1.2.3 Outcomes

This dissertation demonstrates how we might draw on people's diverse background knowledge, interest, and micro-expertise to expand scientific knowledge and

push it in new directions. More specifically, the *Gut Instinct* platform instantiates these ideas enabling participants of the American Gut Project (the world's largest crowd-funded citizen science project) to generate and experimentally investigate hypotheses. 344 volunteers from 27 countries created 399 hypotheses about their health and the gut microbiome. Remarkably, microbiome scientists rated a fifth (75) of these hypotheses to have a scientifically valuable insight about a topic not covered by existing published work. Volunteers fleshed out 60 of these hypotheses into complete experimental designs. My entire work (code + data) is open source so others can edit, build, and experiment.

This dissertation has also enjoyed sufficient support in multiple research communities: Innovation researchers at MIT, online and offline fermentation and self-tracking communities, and citizen science groups. Finally, parts of the system have been taught in classrooms including CSCI 499: (Computing for Social Good) at USC.

This work explores how online learning and process training systems, combined with peer collaboration, can help people learn similar skills that can be useful in scientific and design domains.

Chapter 2

Related Work

This chapter summarizes research in citizen science, lead-user innovation, and social computing that inform the design of systems in this dissertation. Citizen scientists have successfully solved expert-defined problems as sensors or algorithms. However, public involvement in scientific endeavors fail to provide a true participatory experience that is citizen-led and personally meaningful. Lead-user innovation provides a complementary setup. Lived experience, a tight feedback loop, and strong personal motivation enable people to create different and sometimes better products than experts; however, lead users rarely have access to training, conceptual knowledge, and pre-existing organizational structure for collaboration. Finally, social computing and crowdsourcing platforms support sharing potentially novel ideas but converting these to actionable plans requires expert guidance. This dissertation provides ways to integrate learning in social computing to enable deeper contributions from citizen scientists and lead users without expert involvement.

2.1 Science Misses People's Lived Experience

Science is increasingly networked, multidisciplinary, and open [131]. For instance, *LIGO*'s pathbreaking discovery of gravitational waves brought together over 100 researchers from over 100 institutions across 18 countries (ligo.org/about). Scientists increasingly share data and results faster (arxiv.org). Large scientific projects, like the

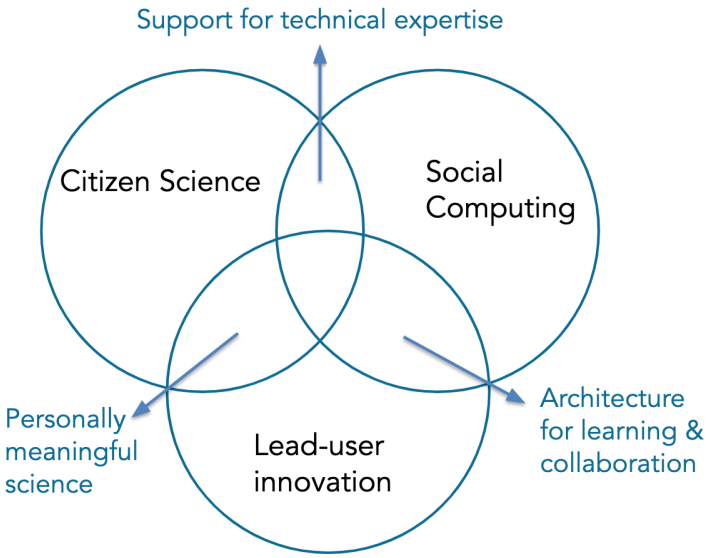

Figure 2.1: This dissertation draws from and contributes to citizen science, lead-user innovation, and social computing

Human Genome Project, took to agile science by sharing methods, data, and insights to collaboratively speed discoveries. Scientists also form global collaborations to accelerate research in nascent scientific domains, like the Earth Microbiome project (earthmicrobiome.org). At its best, institutional science has benefitted immensely from large-scale global collaboration. Complementing this success, many online projects enable people to help scientists [128]: annotating scientific papers [64]; labeling galaxies [78]; and providing microbiome samples [122], CPU cycles (worldcommunitygrid.org), or personal data (openhumans.org). Efforts to further expand participation in scientific research are bearing fruit: *Lab in the Wild* recruits anyone with an internet connection for behavioral studies [136]; and *All of Us* aims to recruit one million Americans from all strata of society (allofus.nih.gov). Such collaborative efforts from experts and citizens suggest a new model for scientific work.

Often, when citizens participate in science, it is as *embedded sensors* that are aggregated by experts. Public involvement in scientific endeavors continues to be largely limited to performing tasks just beyond the reach of computers. A classic example is

Audubon's Christmas bird count, run since 1900 [10]. Online examples include reporting flower blooms in *Project Budburst* [18]; recording wildlife activity [51]; identifying galaxies from satellite imagery in *GalaxyZoo* [177]; and biochemistry games: finding protein structures in *Foldit* [34], synthesizing RNA molecules in *EteRNA* [107], and aligning nucleotide sequences in *Phylo* [80]. Distributed data contributions from people around the world—browsing online [37], using activity trackers, and joining scientific projects—have enabled valuable insights on topics including obesity [4], aesthetic preferences [137], sleep [54], and the human microbiome [121]. At their best, these citizen science platforms yield novel insights. For example, *Foldit* players discovered protein structures that helped scientists understand how the AIDS virus reproduces [36]. Why have such collaborative efforts succeeded?

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 [152]. 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 [21]. This volunteer-led discovery demonstrates the need for fostering independent perspectives while simultaneously cultivating sufficient knowledge for meaningful domain contributions. Such collaboration requires strategic isolation: roviding just enough scaffolding to keep biases independent, while not stifling original ideas for bottom-up knowledge creation.

2.1.1 Opportunity: Can people be scientists rather than just sensors?

Citizens have successfully solved expert-defined problems as sensors or algorithms with a row-filling model of contribution. In the quest to get people to track, measure, accumulate, or sort both digital and analog data, citizen science has overlooked

the massive opportunity of leveraging people's unique advantages: our skills as reflective, creative thinkers who generate theories about the world, including ourselves. People can offer more than just their data and perceptual skills: they create theories, right or wrong, about a wide range of topics including emotions [77], motivation [117], or diet. These may be observational theories [82], folk theories passed in a family/culture across generations [59], or ideas brainstormed in online communities [1]. Perhaps, these intuitions can provide a starting point for independent, participatory experience that also assists the scientific community.

When are such personal experiences worth paying attention to? For every intuition proven right, many more may be closer to snake oil — e.g., the widespread belief in the utility of probiotics despite limited evidence [16]. The global internet increases the proliferation of both powerful and questionable ideas: sharing speculation is fast while evaluation remains slow. Moreover, people develop intuitions of cause and effect that may or may not be correct. Current online forum designs prioritize discussion — sharing personal details in long, free-flowing text — over structure, succinctness, learning, and potential scientific utility [155].

Advances in precision medicine have demonstrated the need to engage people in uncovering and sharing insights [8]. People are highly motivated to improve their health outcomes, more so if they suffer from a condition that severely affects their quality of life, naturally forming communities. For example, patients from the Amyotrophic Lateral Sclerosis (ALS) community on *Patients Like Me* (patientslikeme.com) organized a study to track effects of Lithium on their symptoms [166]. This is not surprising; lead users excel at tackling *need-intensive* problems where they can use their lived experiences to identify problems, try solutions, and readily observe the effects [163]. Other organized communities like *Quantified Self* hope to uncover lifestyle patterns that may improve their productivity and health outcomes. The word 'self' belies the fact that such movements

are highly collaborative: amateurs frequently share experiences and invite feedback on online fora (patientslikeme.com) and blogs (ibsgroup.org). Millions follow these ideas and some incorporate these intuitions in their lives. What kinds of scaffolds and structure may help people generate better ideas and implement them as actionable plans that enable researchers to identify promising insights?

Most scientists develop their skills through an apprenticeship- based graduate school experience. Apprenticeships emphasize hands-on experience with individualized, taskspecific feedback [145]. Scientists possess a wealth of declarative knowledge about their domains (e.g., how to set up a randomized controlled trial), and also procedural knowledge —some narrow, some broad —towards getting things done (e.g., improving fMRI signal intensity by having participants consume cocoa beforehand [56]). This dissertation explores how online learning and process training systems, combined with peer collaboration, can help people learn similar skills that can be useful in scientific domains.

2.2 Lead users Succeed When They Know What to Do and How to Do It

Lead-user innovation is both an inspiration and an application area for this dissertation. Lead users are users of a product (or service) who experience advanced needs unmet by existing products [163]. 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. Lead users also collaborate online to build software (github.com), create novel hardware & reference designs (openaps.org), and share personal data (quantified-

self.com, openhumans.org). Some go further still, e.g., the transcranial direct-current stimulation community draws ideas from scientific papers to attempt self-experiments (reddit.com/r/tDCS). In a few exceptional cases, lead users have authored scientific papers, e.g., Open Artificial Pancreas creator Dana Lewis discussed the benefits and challenges of first-generation automated insulin delivery at the 2016 American Diabetes Conference [41].

Why do people do this? Curiosity, personal learning, and social comparison are three reasons [136]. A massive interest in personal genomics (over 1 million 23andme participants) and the human microbiome (13,000 *American Gut* participants) demonstrates people's yearning for self-understanding. Users of these platforms send data, answer survey questions, and discuss on fora. Some even use online lectures to understand concepts of genes, phenotypes, and microbiota [2, 93].

2.2.1 Opportunity: Providing lead users the expertise to tackle complex knowledge work

Sometimes, having a different background than experts can be beneficial. Shared knowledge is great when it's right, but blocks progress when wrong. When false assumptions limit experts, at least some novices are likely to be *uninfected*. The converse also holds, and much more often: novices are also uninfected by all the knowledge that enables experts to innovate. Lead users have an advantage when the key ingredient is experience intensive; experts retain the advantage for *solution-intensive* innovations [32]. In a large distributed community, there's often someone who happens to have important relevant knowledge, usually drawing on a relevant but distant domain. Having many people work on the same problem increases the odds that one will break through. Drawing on secondary expertise as inspiration can be an important agent of creativity

because almost by definition, the combination is rare [14].

Community-driven approaches to understand personal health and well-being largely reside outside the realm of institutional science and medicine. While some fads and beliefs are questionable at best, on occasion communities break new ground that may provide widespread value, such as fecal transplants to alleviate *Clostridium difficile* infection symptoms [19]. Some doctors recommend that patients track their symptoms and reflect upon them to find insights. Putting people in charge can help them find significant relief for ailments like chronic migraine [57] and provide researchers and clinicians with useful patient data (smartpatients.com). Insights from N=1 studies have helped crack scientific puzzles about the working of the mind [157], heart, and microbes [165].

Personal needs and challenges can be highly motivating but performing complex work still requires multiple rounds of trial and error. People need to know the genre of work and implement it correctly. Professionals have the advantages of training, conceptual knowledge, pre-existing organizational structure for collaboration and support, and direct access to resources. Lead users either seek these resources from others or need to create them. Providing a correct and complete model for complex, structured activities might reduce efforts and improve the quality of results. This dissertation reduces the gap between lead users' ideas and implementation by providing templates for genre work using just-in-time training and a collaboration platform to find others. This dissertation focuses on enabling people transform their idea to a controlled experiment as opposed to self-tracking or informal iteration which is the focus of most current citizen-led work in health.

2.2.2 Case: Scientific experimentation is difficult

While public contributions have supported institutional science; it's rare for citizens to design their own experiments. A number of health and behavioral research projects enlist citizens as helpers (e.g., *HabitLab* [43]). *CivilServant* enables online communities' moderators to test policy ideas; moderators share these ideas with researchers who transform them to study designs [51]. Through the *PatientsLikeMe* website (patientslikeme.com), citizens and scientists created a study investigating whether consuming lithium alleviated ALS symptoms [64]. While an initial scientific study had provided positive benefits, both this citizen science study and a subsequent university study did not find benefits. *Tummy Trials* asked participants to generate health questions, introducing a protocol for self-experimentation combining ideation and self-tracking [36]. In all these cases, citizens rely on experts to provide sound experimental design.

Why is experimentation hard? 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. Providing feedback on experiment designs requires knowing the success criteria and how to help improve. Feedback can be provided by experts [45, 145], peers [17, 100], software [42, 71], or even oneself [17, 145]. While feedback from novices can potentially improve both structure and content, it can also emphasize superficial issues over the underlying structure [27]. Finally, successfully running an experiment requires managing multiple processes such as random assignment, anonymizing participant details, and sending instructions and reminders for data collection.

2.3 Social Computing and Crowdsourcing Architectures for Complex Work

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 [89]. Leveraging existing expertise is one approach for complex knowledge work. One strategy directly employs experts' just-in-time feedback to improve crowd work [45]. Workflows manage experts for open-ended work like developing interactive prototypes [139]. *Flash Organizations* uses automated hiring, a hierarchy with a central leader, and optional team leaders for collaborative projects like product design [158]. Another strategy creates roles that enable more experienced crowd members to orchestrate the work. *Ensemble* supports leaders in guiding and constraining crowd contributions [84]. 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.

Carefully-constructed interfaces can aid novices with task-specific expertise to solve problems that only experts previously could. *Foldit* introduced 3D game for specifying low-energy protein structures via direct manipulation [34]. Making a challenge visually salient is an effective way to on-board novices. For tasks that don't have as a crisp visual analogue as protein folding, people need better conceptual support. Prior work has explored collaborative hypothesis generation and testing on pre-existing data sets [113, 169]. This dissertation offers a complementary contribution: enabling citizens to generate data on topics of personal interest.

One way to make complex tasks manageable is to divide them into distinct

phases. Touchstone demonstrates the power of a semi-automated workflow integrating experiment design, testing, and analysis [115]. Crowdsourcing has similarly innovated by creating distinct phases: break larger tasks into microtasks; algorithms specify the division, dependency, and agglomeration activities while workers perform small tasks supported by task-specific guidelines [102]. From these systems, our work draws the idea of dividing experimentation into multiple tasks—some self-sourced, others crowdsourced; and introduce just-in-time domain expertise to perform these tasks.

2.3.1 Learning resources at the right time

Providing just-in-time supports, step-by-step instruction, and showing helpful supportive information are core ideas in instructional design [87]. Crowdsourcing systems leverage interactive guidance for specific tasks. For example, *CrowdLayout* and *Cicero* provide guidelines and static rules that workers use these to reason about their choices and improve network layouts [25, 148]. Others like *CrowdSCIM* and *Crowdclass* scaffold pre-task interventions [106, 164]. While learning resources are distributed across the internet, they are rarely integrated with the task.

Creative, open-ended work has rich pedagogical value. Online work, like online learning, requires appropriate scaffoldings, such as rubrics [17, 101], decision trees [106, 175], tutorials [6], and quick expert guidance [45]. Similar to general critique of pure discovery learning [120], simply asking participants to *figure it out* would be poor pedagogy. Hence, this dissertation introduces a guided discovery learning approach as Mayer advocates: expert-curated learning materials help participants start, with discovery following. Such integration offers a problem-based learning experience with context and motivation for the material students learn [144]. In principle, these real-world problems also provide a yardstick for measuring learning.

This dissertation introduces support during the task itself for those with little-to-

no mental model of the knowledge domain. Like the *Shepherd* review-writing system [45], this dissertation provides just-in-time support. There are two key differences: 1) this work scaffolds the entire creation process, not just the post-draft feedback stage, and 2) it does not draw on expert time – the knowledge is implemented in the software itself.

2.4 Microbiome Research: A Petri dish for Personally Meaningful Scientific Work

The human microbiome is the collection of all microbes and their genetic components in and on our bodies. It is highly personal: each of us hosts a different collection of microbes, and this collection is influenced by our environment, diet, health, lifestyle, and genetics. A major scientific effort is to better characterize and understand this diversity and the causal factors for it (hmpdacc.org). Understanding the human microbiome requires insights into people's lifestyles. Microbiome science is nascent, highly contextual, and personally motivating. However, research has only scratched the surface of understanding the microbiome and using it to improve our wellbeing. Engaging diverse participants at scale can potentially yield exciting new results.

The *American Gut Project* (americangut.org) offers a crowdsourced opportunity for people to get a microbiome sampling kit [94]. AGP participants contribute their samples for bacterial marker gene sequencing and analysis [43]. Participants then receive a summary of their results with all their raw data. Anonymized data is publically available.

To date, more than 13,000 people have participated. Participants submit both a physical sample and fill out a survey. AGP seeks to build a comprehensive map of the human microbiome, and identify its healthy and unhealthy components. Analysis has revealed lifestyle-microbiome correlations of dog ownership and beer or vegetable

consumption, among others.

People hold the key to understanding the gut microbiome

The structure of the human microbiome is influenced by many factors, including age, genetics, diet, and xenobiotic and antibiotic use [61]. The gut microbiome in particular plays an important role in metabolism and immune system development, and some microbiome dysbioses have been associated with diseases such as obesity, inflammatory bowel disease, type I and type II diabetes, autism, multiple sclerosis, and malnutrition [29]. The human microbiome is impossible to understand without information about its host [43] and many influence factors remain unknown. Teaching people about the gut microbiome and having them guess associations between the microbiome and health and disease states can potentially accelerate the process of discovering links between diet, disease, and lifestyle factors and the gut microbiome.

Currently, the topics for scientific investigation are handpicked by a small group of scientists. Can opening up the scientific process to the world yield additional insights? How can people's situated knowledge supplement institutional science? This dissertation provides systems and techniques for novices to complement experts in creating new knowledge about the microbiome.

This chapter provided an overview of related research; chapters dedicated to specific systems discuss additional research that informs that design of those systems. This dissertation explores integrating learning in social computing for complex, creative work with two goals in mind: efficacy of the resulting systems (i.e. usability, correctness, and existential evidence) and generality of the underlying techniques upon which the tools are built.

Chapter 3

Collaboratively Generating Ideas

Learners worldwide collectively spend millions of hours per week testing their skills on assignments with known answers. Might some of this time fruitfully be spent posing and exploring novel questions? This chapter investigates an approach for learners to contribute scientific ideas. The Gut Instinct system embodies this approach, hosting online learning materials and invites learners to collaboratively brainstorm potential influences on people's microbiome. A between-subjects experiment compared the performance of participants who engaged in just learning, just contributing, or a combination. Participants in the learning condition scored highest on a summative test. Participants in both the contribution and combined conditions generated novel, useful questions; there was not a significant difference between the two. Though participants in the combined condition both learned and contributed, this setting did not exhibit an additive benefit, such as better learning in the combined condition.

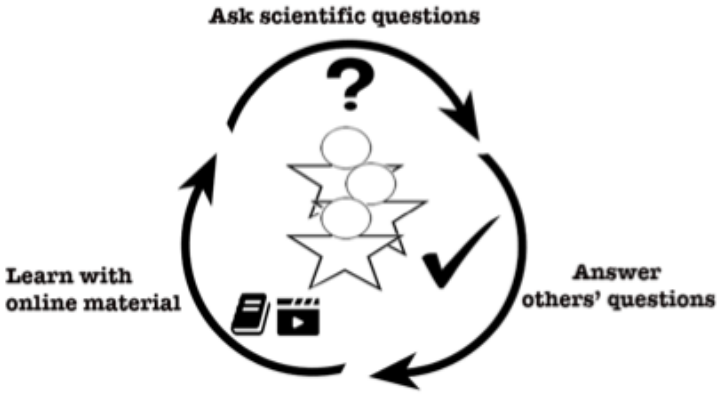

Figure 3.1: A dual objective: integrating citizen science and online learning

3.1 The Promise of Citizen Science with Learners

People worldwide have theories about their health, environment, interpersonal interactions, and myriad other topics [59]. Some of these folk theories encapsulate generalizable insights and wisdom; many others are completely false; and some are in between [82]. How might we harvest and assess such intuitive theories to extend human knowledge, especially in domains where science is limited? Worldwide, students collectively spend millions of hours a week testing their skills on assignments with known answers [146]. This community could be a potentially powerful resource. Repurposing even a small fraction of this effort towards scientific inquiry could pay significant dividends.

Our intuition is that scientific crowdsourcing will most usefully contribute to domains where science is nascent and/or highly contextual. Knowledge of the human microbiome is both. While everyone has a gut full of microbes, its causal influences remain largely unknown. The Human Microbiome Project and other studies have begun revealing its diversity and impacts [31, 32]. The world could benefit greatly from a more comprehensive understanding of the microbiome, what influences its composition, and the impact our gut has on our health. Understanding how people live may help build causal models. For example, rheumatoid arthritis patients have altered gut and oral bacteria [176]. Might changing their gut reduce their symptoms? As in many scientific domains, people's initial intuitions about what affects their gut are often poor. Does this improve with education? Could learners collectively advance human understanding in this domain? This chapter explores the potential of coupling online citizen science with learning materials to create scientific questions (Figure 3.1).

The main contribution of this chapter is *demonstrating that a crowd of online non-expert learners can collaboratively perform useful scientific work*. To investigate its efficacy

in practice, we have built a web system, Gut Instinct, which brings together learners to perform useful collaborative brainstorming on a citizen science project while developing expertise. A between-subjects experiment compared three variations of Gut Instinct: a contribution focus, a learning focus, and a combined condition. Participants did indeed perform useful creative work. For example, they generated 10 distinct questions that mirror recent scientific discoveries [95]. However, the combined condition did not show additive benefits.

3.1.1 Leveraging Crowdsourcing successes

Collectively aggregating many people's responses can produce faster, better, and more reliable results—at much larger scale—than lone individuals can, at least when errors and biases are independent events [153]. Canonical crowdsourcing tasks have clear right or wrong answers – like whether two images represent the same product, whether an image region contains a feature, or what street number is written on a sign.

Distributing labor redundantly across multiple workers also guards against individual shortcomings [149]. For example, workers using the Soylent crowd-powered document editor found a typo late in a paper that eluded all eight authors and six reviewers [13]. Why? In later pages, fatigue can reduce attention to detail. Because individual crowd workers saw only a small piece of the document, their collective attention to detail remained constant throughout. This illustrates how a collection of novices offers complementary contributions to experts, often in small but nonetheless useful ways.

3.1.2 Dual objective functions in learning and crowdsourcing

Combining university classes in psychology with editing Wikipedia articles led to improvement in the scientific content of over 800 Wikipedia articles while students

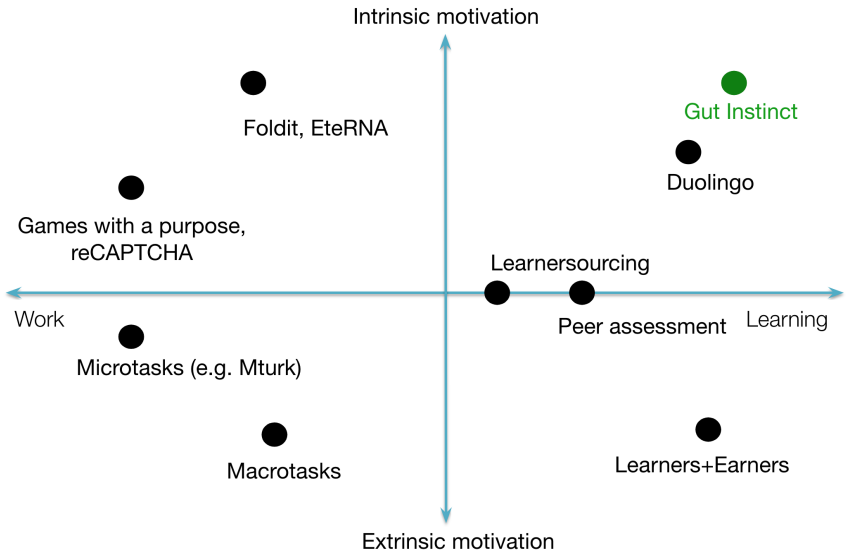

Figure 3.2: Crowd systems/techniques place different emphasis on work and learning. Some, like Mechanical Turk [5], emphasize work over learning. Crowd approaches also vary in their motivation. Games like Foldit [34] leverage participants' motivation to perform altruistic work while having fun. Gut Instinct helps participants learn about the gut microbiome while contributing towards the altruistic purpose of helping researchers better understand it.

learned about the topic they edited [52]. Similarly, Kim et al. asked learners to create how-to video segments as part of an online curriculum [83]; the student-created videos then became a learning resource for the next cohort.

Some crowdsourcing offers a dual objective: user-facing goals include fun (e.g., Peekaboom [160]), authentication (reCAPTCHA [161]), and learning (Duolingo [66]). Under the hood, these tasks simultaneously label images, transcribe text, and translate phrases. Such crowd work can also bootstrap machine learning [11]. This paper is distinct from prior work (Figure 3.2) in leveraging people's individual lived experience, knowledge, context, and folk theories, rather than treating people as interchangeable respondents.

3.2 Hypotheses

This chapter investigates an approach for a community of learners to collaboratively create scientific theories. Learning is any endeavor that seeks to increase a participant's knowledge. In this submission—like many MOOCs—watching videos is the main form of learning, & quizzes are the main assessment. Work is any endeavour where the outcome has value. In this submission, authoring & answering questions are the main work forms. This study operationalized engagement as time spent. We hypothesized that doing useful work on real-world problems helps learning, and vice versa. Specifically:

H1. Learning improves quality of work on relevant problems. Learning, by definition, improves performance on similar tasks. Strangely, transfer to novel tasks (like creating new & different questions) is famously uneven [14]—and sometimes detrimental. H1 tests whether learning would improve work (e.g., novel question creation) because it marries lived knowledge (about diet, health, etc.) with a conceptual framework about the gut's role.

H2. Working on relevant real-world problems improves learning. H2 tests whether working improves learning because it increases motivation & provides an immediately relevant *host context* for new knowledge.

H3. Working while learning improves learners' engagement with the learning material. For similar reasons, we hypothesized that working alongside learning would increase engagement because the two endeavors both *get the wheels turning* in hopefully complementary ways.

We test these hypotheses in the context of brainstorming potential causal relationships in the human gut microbiome.

3.3 The Gut Instinct System

Gut Instinct is a collaborative system with a dual objective: help people learn about the gut microbiome, and catalyze the creation of a list of factors that may be associated with gut microbiome differences (Figure 3.3). People anonymously post questions about lifestyle and health for peers to answer. Learners both ask & answer questions, there are no distinct workers. These questions and discussions provide researchers cues to build associations between lifestyle and the microbiome. Gut Instinct is a web application built with Meteor (http://www.meteor.com). The front-end uses Angular (http://www.angularjs.org) and is stylized with Materialize (http://www.materialize.css).

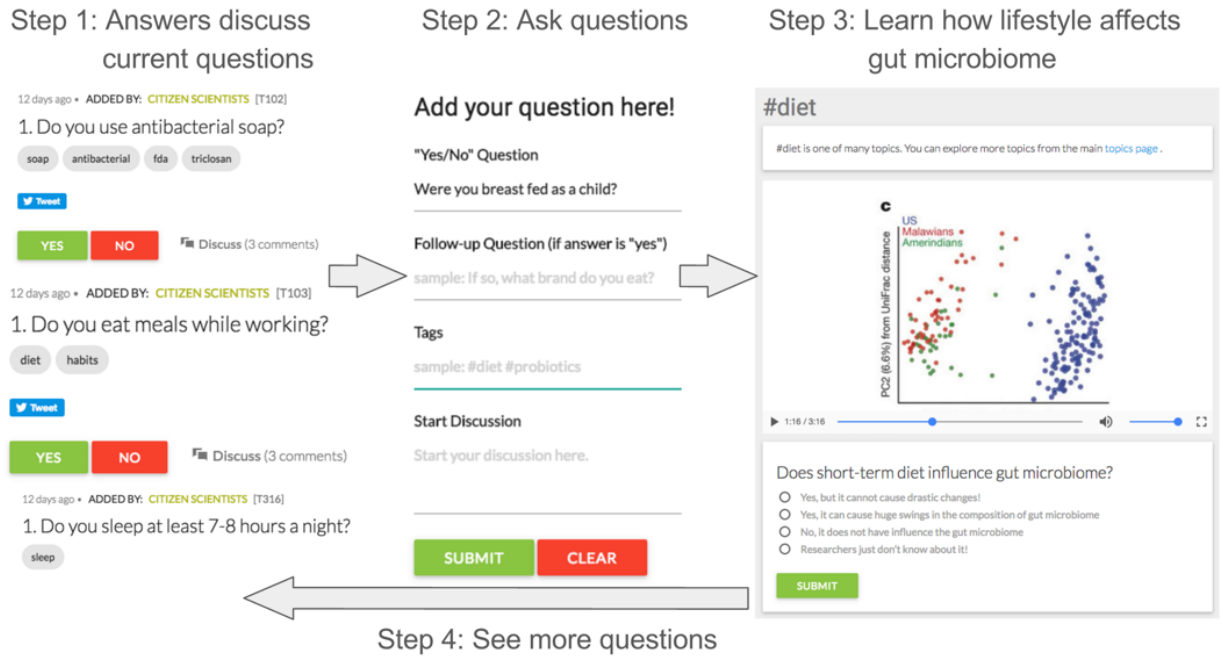

Figure 3.3: Gut Instinct is a web system to learn about the gut microbiome and create causal theories about gut microbiome (a) A discussion board where learners add their questions and discuss them with other learners (b) "Add question" box for people to add their own questions, (c) A tutorial video showing how gut microbiome varies across countries with different food habits [174]

3.3.1 Curating content based on topics

Gut Instinct provides expert-approved learning material including online lectures, science articles and research papers. Participants add articles they feel are useful, which can be fact-checked by experts. The gut microbiome is an active area of research with new results being generated rapidly. A popular MOOC provides an introduction to science the gut microbiome including its relation to some lifestyle choices [93]. Popular online articles about the microbiome are split between providing correct, useful information and clickbait articles without scientific validity.

Gut Instinct organizes the learning material based on topics such as diet or antibiotics. A topic-based classification of learning material provides two advantages: (a) People can deeply focus on the topics that interest them, and (b) Topics related to specific lifestyle aspects can trigger specific questions. The topics pages include videos and articles about diet, antibiotics, probiotics, physiology, and genetics based on vetted content from online sources. Quick multiple-choice questions with detailed feedback at every topic page help people test their understanding (Figure 3.4). Overall, these elements of the interface form the learning part of the system.

Figure 3.4: A question on topics page for diet to test understanding of the learning material

3.3.2 GutBoard: Discussing and answering questions

The GutBoard provides a discussion board with user-generated questions tagged by topics (Figure 3.3a). People can browse questions, answer them, or participate in discussions. GutBoard presents unanswered questions first. The most popular questions (in terms of discussion comments) bubble to the top of the board.

3.3.3 Adding questions

Gut Instinct provides different tutorials, articles, and expert examples to help users contribute. Gut Instinct requires that questions have a two-part structure: a yes/no question followed by an open-ended elaboration. For example, the yes/no question "Do you take any meal replacements such as protein powders?" might be followed by "Do you take them on a daily basis?" This structure addresses two problems we witnessed with pilot users: (a) Some questions were actually multiple different questions, confusing readers (b) Readers had to read every question in full to understand what was being asked, even if the topic was not relevant to them. With this structure, every question has a single focused topic. Participants can also start a discussion about the question and provide relevant tags. "Add Question" box in Figure 3.3b shows the interface.

Nudges to think creatively and to stay on task

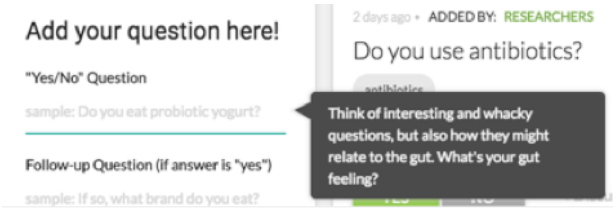

Figure 3.5: An example of a nudge used in Gut Instinct to remind people of their role as a citizen scientist in raising interesting questions about the gut microbiome

Gut Instinct employs several best practices for increasing high-quality contributions [75, 138]. It provides cues to teach participants to generate good questions. All parts of the Add Question box contained sample questions to help participants frame their questions that could be useful to them and to gut microbiome researchers (Figure 3.5). To reduce user confusion, GutBoard was seeded with expert questions that set norms for the nature of questions. To provide a clear call to action, GutBoard was the default landing page and the only place to add or view questions. Every page had a tour that users could invoke anytime to learn its interface.

3.4 Experiment: Work, Learning, & Combined

Figure 3.6: Three conditions for experiment. (a) Contribute condition where participants read some general articles about microbiome and added questions and answered others' questions (b) Learn condition where participants saw curated topic videos (e.g. about diet) and answered practice problems from a Coursera class [93] (c) Combined condition where participants saw curated topic video, and added questions and answered others' questions

A between-subjects experiment compared the work and learning performance of participants across three different conditions: *Contribute*, *Learn* and *Combined* (Figure 3.6). In the Learn condition, participants were provided learning material and some practice

problems, both curated from the Coursera microbiome class [93]). In the Contribute condition, they had access to brief pop-science articles to know basic details about the gut microbiome, and GutBoard for creating questions. In the Combined condition, subjects had access to both learning material from Coursera and the GutBoard. The GutBoard content was common to both conditions that used it (Contribute and Combined).

Method

Participants were randomly assigned to one of the three conditions. Each comprised an individual lab session followed by web study, during which participants were asked to use the tool for 3 days. During this period, participants asked and answered each others' questions in the tool.

Lab: A researcher introduced the condition-appropriate Gut Instinct site. Participants were told there was no lower or upper limit on how much time to spend using the system. Each session comprised the following steps: (1) accessing the consent form, (2) seeing GutBoard/problems, (3) accessing topic videos/articles, and (4) participating in a short interview. The interview asked participants about their knowledge of the gut microbiome before using the system, and their experience using the system. The interview was tailored to the participant's behavior: for example, if a participant did not click on Google Scholar references inside Gut Instinct but opened up a browser for web search, the interviewer would ask why.

Web usage: Once all participants had completed the lab portion, the web application was opened to all participants for collaborative usage for three days. Gut Instinct sent email notifications about activity on the site, along with feedback on some questions raised on GutBoard such as providing links to research studies about effects of eating blueberries on the gut microbiome.

After web usage, two independent raters (experts in human microbiome) rated

the questions on novelty & usefulness using the following workflow: (1) calibrate: rate 3 questions independently and discuss; (2) rate: independently rate all participant generated questions; (3) combine: discuss ratings where different & develop a common score. The discussion in step 3 was valuable for adding to the set of rules for rating such open-ended questions.

Table 3.1: Demography information for 44 participants. Some participants did not complete portions of survey

Nationality	Indian = 22	Non-Indian = 22
Gender	Female = 7	Male = 37
Age	18-20 = 1	26-30 = 19
	21-26 = 14	31-35 = 5
Ethnicity	Indian = 18	Caucasian= 11
	Asian/Pacific Islander= 5	Hispanic/Latino = 2
	Others/Not said = 4	
Current	Undergraduate = 3	Ph.D. = 29
educational status	Masters = 7	Postdoc = 2

Participants

44 participants were recruited from a Southern California university (Table 3.1). Participants were novices in terms of their knowledge of the human microbiome. Random assignment balanced gender and nationality across conditions. There were equal numbers of women—and equal numbers of men—in each condition. Where not evenly divisible by 3, one condition had one more or fewer.

Measures

Dependent variables comprised work (number of questions contributed, novelty and usefulness measured by blind, independent raters); learning (score on summative test); and engagement (time spent during lab session, and number of discussion comments during web usage). Qualitative measures included how participants used the tool, where they got stuck, how they collected information, which questions they engaged with, and a post-experiment survey.

Results

Analysis of variance estimated the effect of working, learning and the work-learn interaction. Two condition comparisons used a Mann-Whitney U test with the corresponding independent variable (learning or working).

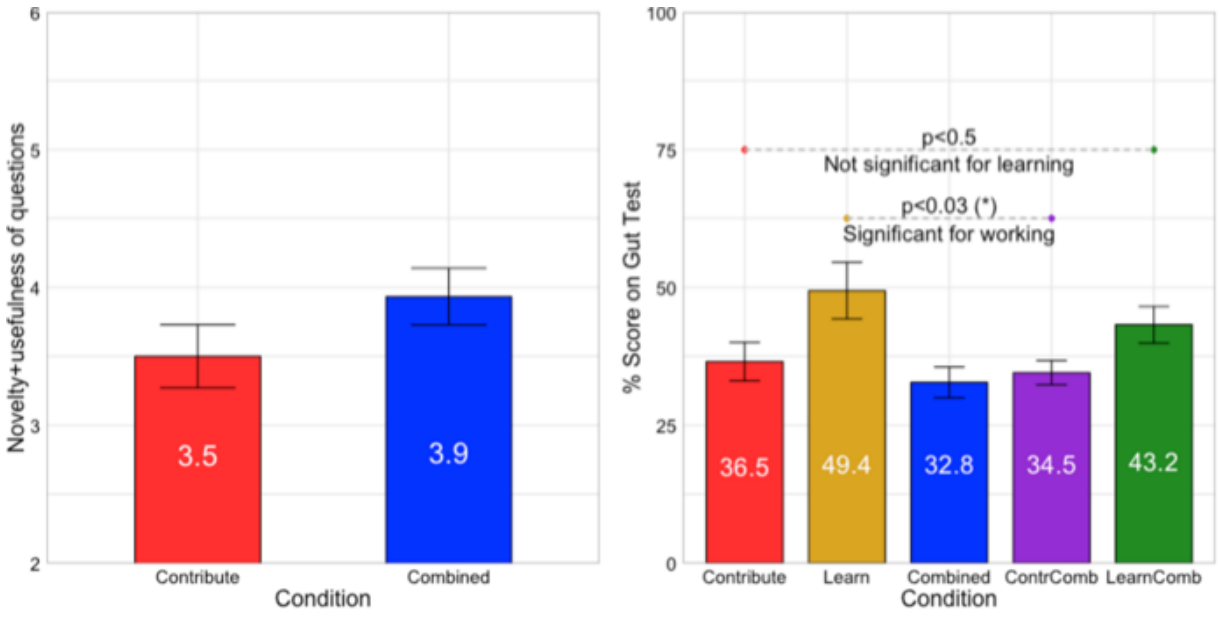

Figure 3.7: a) Participants in Contribute and Combined conditions created questions of similar quality b) Participants in Learn condition performed the best on a summative test. Learning did not show a significant effect on score but working did

Work: Did access to Coursera learning material (Combined) impact quantity and

quality of questions relative to not having access (Contribute)? The Combined participants generated questions of similar novelty and usefulness (M = 3.5) as Contribute participants (M = 3.93), Mann–Whitney U = 79, n1 = 14, n2 = 15, $p < 0.23$ two-tailed (Figure 3.7a). Figure 3.8 shows two examples of questions rated by experts. Ten of the 29 questions mirrored questions found on the American Gut survey. Half of the participants' questions (14 of 29) asked about diet. Participants in Combined and Contribute conditions generated a total of 14 and 15 questions, respectively, averaging one question per participant (see Figure 3.10).

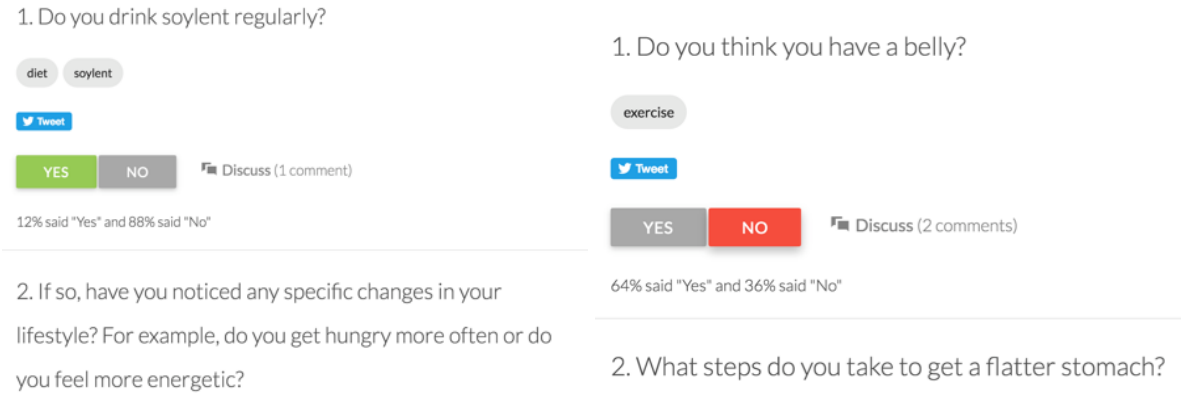

Figure 3.8: An example of a good and bad question added by participants. Soylent question was scored 5/6 (2 on novelty and 3 on usefulness) while the belly question was rated 2/6 (1 on novelty and 1 on usefulness)

Learning: Did participants instructed to ask questions (Contribute & Combined) score differently than those who were not (Learning)? Did access to learning videos (Learning & Combined conditions) impact quiz scores relative to not having access (Contribute)? A two-factor ANOVA estimated these effects, finding significantly lower scores for those requested to ask questions. By contrast, access to learning materials did not yield a significant difference in quiz score.

The Learn participants performed scored higher (M = 5.93) on Learning test than participants in Combined (M = 4.38) or Contribute (M = 3.93) conditions. An analysis of variance showed that this effect was significant for working, $F(1, 39) = 5.22$, $p < 0.03$, but

not for learning, F(1, 39) = 0.46, p < 0.5 (Figure 3.7b). The effect size for working was small (Cohen's effect size d = .11).

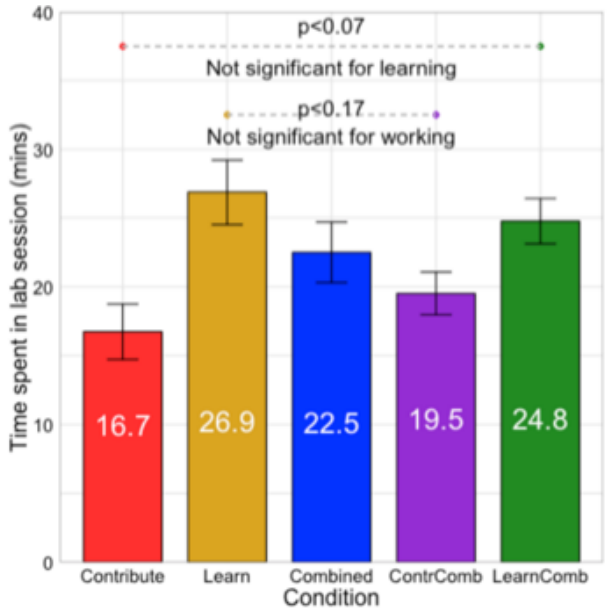

Figure 3.9: There were no significant differences in time spent in lab session across the conditions

Engagement: The mean length of lab session was 20 min (varying from 9-40 min). Learn participants spent marginally more time (M = 26.9 min) in the lab session than participants in Combined (M = 22.5 min) or Contribute (M = 16.7 min) conditions. An analysis of variance showed that this effect wasn't significant for either Learning F(1, 41) = 3.40, p < 0.07 or working F(1, 41) = 1.95, p < 0.17, (Figure 3.9). Combined and Contribute participants contributed 35 discussion comments each; Learn participants contributed 10 discussion comments.

38 of 40 correspondents reported prior use of online courses, varying from occasional use of online learning material to taking more than five online classes. Preliminary analyses found no effects for gender and nationality (Indian or non-Indian), so these were excluded from further analyses. Table 5.1 summarizes results from the experiment.

Table 3.2: Summary of results from experiment

Measures (mean values)	*Combined*	*Contribute*	*Learn*	*p*
No difference in quality or quantity of questions across Combined or Contribute				
Quality of questions (2-6 scale)	3.5	3.93	-	$< .23$
# of questions/# of participants	14/14	15/15	-	-
Working reduced test scores				
Test score (max: 12 pts)	4.38	3.93	5.92	L $< .5$, W $< .03$
Learning or contributing did not have a significant effect on time spent in lab				
Time taken in lab session (min)	22.5	16.7	26.8	L $< .07$, W $< .17$
# of discussion comments	35	35	10	-

3.5 Discussion

These results suggest that some learners create useful research questions based on their lifestyle but its effect on better learning is unclear.

Why did Learn participants perform better on tests? Learn participants had a clear objective: learn about the gut microbiome, practice problems related to it, and take a summative test. By contrast, participants in the other two conditions had to both generate novel questions and take the summative test. They may have placed less emphasis on the test. Generating questions and taking test on a novel topic might have required greater effort than what the participants wanted to put in. Additionally,

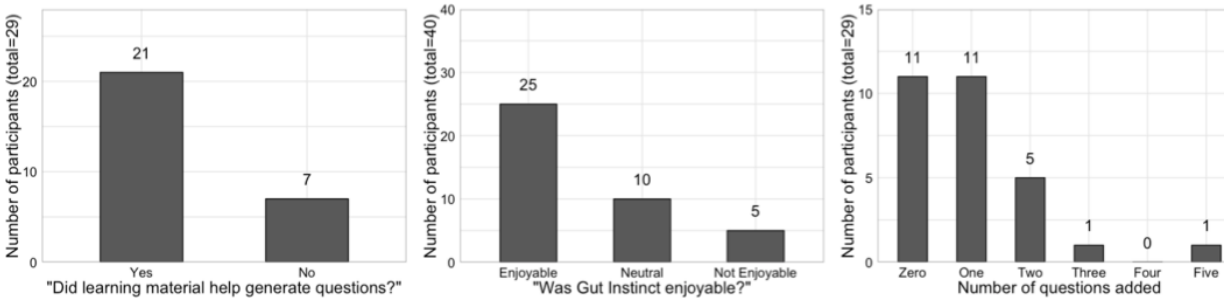

Figure 3.10: Most participants reported that the learning experience was helpful and the system was enjoyable. 65% of participants asked questions; the mode was 1)

difficulty of creating questions may have lowered participants' confidence in taking the test.

Personalized learning and need for feedback Participants were curious to know the microbiome science behind disclosed aspects of their lifestyle. One participant commented, "After answering the question, I would expect to see some succinct information about where my lifestyle stands with respect to scientific wisdom." Participants also asked for a section curated for them by the tool, or a section where they could save items of personal relevance.

Need for self-directed learning Online learning material provided useful information about a complex topic like the microbiome hoping that it might spur participants to find and use other similarly trustworthy sources of their liking. In the lab, participants used web search to find relevant resources. Most participants reported that they did not search at home.

Learning did not improve quality of work Combined participants did not generate questions of higher quality than those without learning material (Contribute condition). Crowdclass [106] found similar results where workers who simply classified images did better than those who learned about decision trees and subsequently classified images as an assignment. How do online learning materials and useful work tie to each other? Gut Instinct explored one design point where learning and working were provided specific

components in the tool to reduce participant confusion. An alternate approach could be to have a work-biased design where learning material would be tailored to participants' questions or a learn-biased design where participants could add questions only in the specific context of learning materials. For instance, people could raise questions at different point of a topic video [108] rather than using a separate part of the tool.

Difficulty of generating questions A remarkable and concrete measure of participants' insights is that ten of their questions mirrored those asked by the American Gut survey [95]. Unsurprisingly, other participants reported difficulty creating good questions. Asking questions is a valuable metacognitive experience that can be scaffolded by examples of good questions from experts.

Gut Instinct sent email asking people to contribute, reminding them of the importance of their task, and showing successful examples of citizen science work. Such reminders prompted a temporary increase in questions increased or discussion contributions [98] but did not lead to a sustained stream of questions and discussions. Some participants complied with the letter of the request but not the spirit by taking a sample question and tweaking it slightly.

What kind of innovation can we expect from citizen science Half of participants' questions were about diet. Diet offers both a clear influence mechanism and immediate personal relevance. While a compelling video about effect of diet on the microbiome likely helped, a video alone appears insufficient: for instance, the topic of genetics also had a video, but no participants asked questions about genetics. Moreover, many diet questions are perceived as less personally disclosive than genetics questions.

That participants asked many questions about diet and none about genetics is consistent with patterns of where lead users innovate, and where they don't [162]. Lead-user innovation works best for *need intensive* problems where people's lived experiences provide the key ingredient, e.g., a snowboarder who cuts their boots to improve fit. These

innovations arise through trial and error, and solution efficacy is readily observable. Lead-user innovation is less common with *solution intensive* problems, where highly technical knowledge, access to equipment, and/or significant financial capital are critical.

Does browsing displace contributing? Participants spent most of the lab session browsing discussions and learning material. By our observation, later participants spent more time using the system in the lab. Despite spending more time browsing discussions, we think later participants added fewer questions. Participants mentioned that browsing and answering questions felt like *contributing* without putting in a lot of effort. Participants also reported that they had to break a mental barrier to publicly post a discussion comment or question.

Limitations

Participants could login as little or as many times as they wished. One participant commented that even though she had some ideas to add, she was conscious of disclosing information about her personal life (participants were anonymous). It may be that using the tool in an experiment made them more cautious of what they added or commented. As a web application, participants assumed comparable facilities to forum/ discussion sites like Quora. This exemplifies a challenge of testing research prototypes: the absence of production-level features can change participants' impressions and possibly their behavior.

3.6 Science with Learners: Promise & Challenges

This chapter investigated the merits of combining learning and contributing. While experimental results did not show the hypothesized additive benefits, we still believe this combination has potential. Is it intrinsically self-contradictory to ask learners

to contribute scientific ideas? Not necessarily. In addition to the diversity benefits that the global community brings, those with brand new knowledge can, for example, give useful feedback to peers [101]. Furthermore, the newly-aware sometimes articulate useful insights that familiarity has blinded experts to [73]. Drawing on the results, related literature, and our intuitions, here are avenues that might find additive benefits where this experiment did not.

Aligning objectives The chapter's experiment gave participants two objectives: take a summative test and generate ideas for lifestyle-microbiome relationships. While both relate to the same general topic—the microbiome—the "doing" of each was quite different. For example, the question that the fewest participants answered correctly asked which type of bacteria population would be affected by a behavior change. While the test emphasized specific biological facts like this, participants' GutBoard questions were much more general. Consequently, it is not surprising that success on one didn't catalyze success on the other. Conversely, given the mild negative correlation, it seems likely that time spent on one might have taken away from time spent on the other. More tightly aligning the test of learning with the work activities could yield the additive benefits we seek. (We say *test of learning* because participants may indeed have learned more in ways not measured.)

We also hypothesize that an additive benefit is more likely when the knowledge and/or motivation generated by one activity transfers to the other. While this may seem obvious in retrospect, the loose-connection problem observed here may be relatively common. We hope the results warn against this risk.

Make learning & work personally relevant Many American Gut Project (AGP) participants exhibit a strong intrinsic motivation to learn more about why they have a particular microbiome [43]. The students who participated in this experiment may not have equivalently strong motivation. Motivated users may increase the quality of

citizen science work. For instance, AGP participants could organize a focused effort around a specific health issue like Type 1 diabetes or Inflammatory Bowel Disease (IBD). Similar to how Wikipedia editors co-ordinate efforts [99] using Gut Instinct with more differentiated roles like question generation, question ranking, and literature search might lead to further distinguishing work.

Learning & working: integrate & provide clear criteria We believe that integrating learning and work will be mutually beneficial when learning new material immediately opens up the possibility of contributing useful work and contribution solicitations include relevant learning material. This extends problem-based learning [144] and just in time learning [15] to the scale of Internet. For example, browsing StackOverflow before fixing programming questions leads to better work, and lateral learning [116]. Similarly, global-scale distributed contributions like peer review have enabled massive online courses to offer creative, open-ended assignments through peer review [101]. Such active learning approaches seek a dual objective of content learning and metacognitive growth [39].

Reflection and curiosity play a similar orienting role: having people guess the answer to a task-relevant question before performing the main task led to better performance on the task when hints were revealed to maintain the curiosity of the learners [3, 105]. Similarly, the surprise that arises from making a guess that's revealed to be wrong generates a *teachable moment* for learners. How might we use these lessons for online learners to teach themselves about specific domains while performing useful work?

Other fields for this approach Many other fields may benefit from the diverse contexts that online citizen scientists offer. For example, 96% of psychology experiments used participants from Western industrialized countries [72]. Recent attempts have started to collect and analyze data about people all across the world by offering them fun-based

rewards in lieu of collecting data about their online interactions [136]. Success of such initiatives hints at a motivated set of online participants who could also benefit from learning about cultural psychology concepts in more depth while undertaking relevant scientific work.

This chapter investigated techniques for integrating learning and citizen science for the benefit of both. For us, the most striking result is that users contributed many causal questions of sufficient novelty and importance that they only recently have emerged in the literature. It is possible that other of the causal questions will be borne out in the future. This study also illustrates the challenges of double-bottom-line work. Specifically, these dual objectives can be in tension rather than being additive. The chapter describes the Gut Instinct system and suggests strategies that may help the dual objectives enhance each other. Looking forward, we hope the approach introduced here will find value in other domains especially where the science is nascent and/or contextual information is key. The knowledge of science impacts a diverse planet; in the future, this diverse community may importantly contribute to it.

This chapter, in part, includes portions of material as it appears in *Gut Instinct: Creating Scientific Theories with Online Learners* by Vineet Pandey, Amnon Amir, Justine W. Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, and Scott R. Klemmer in the proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2017). The dissertation author was the primary investigator and author of this paper.

Chapter 4

Collaboratively Generating Hypotheses

People's lived experiences provide intuitions about health. Can they transform these personal intuitions into testable hypotheses that could inform both science and their lives? This chapter introduces an online learning architecture and provides system principles for people to brainstorm causal scientific theories. The Learn-Train-Ask workflow (Figure 4.1) guides participants through learning domain-specific content, process training to frame their intuitions as hypotheses, and collaborating with anonymous peers to brainstorm related questions. 344 voluntary online participants from 27 countries created 399 personally-relevant questions about the human microbiome over 4 months, 75 (19%) of which microbiome experts found potentially scientifically novel. Participants with access to process training generated hypotheses of better quality. Access to learning materials improved the questions' microbiome-specific knowledge.

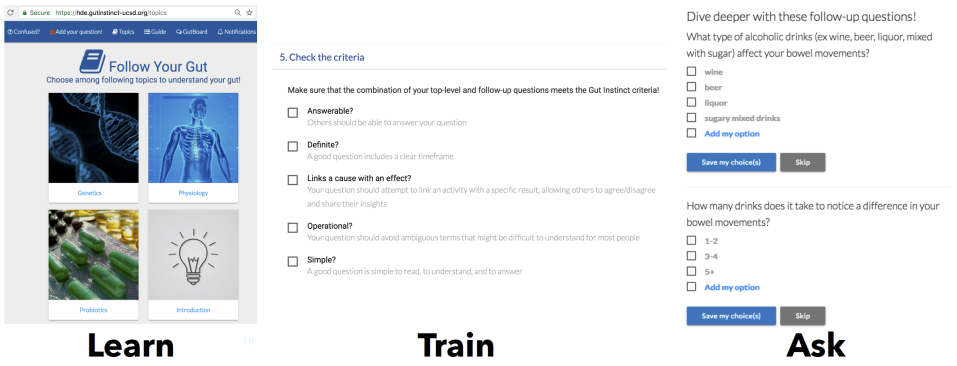

Figure 4.1: The Learn-Train-Ask workflow enables novices to ask useful questions collaborate with online peers

4.1 Can People be Scientists Rather than Just Sensors?

Thinking like a scientist involves generating useful questions, operationalizing them as hypotheses, and testing them with experiments. While people generate (implicit) intuitions from lived experiences, transforming tacit knowledge into explicit questions is not easy. People don't always realize the extent of their knowledge and even when they do, asking questions that can be answered by others to yield clear, actionable insights is hard. For instance, people often bury questions in long entries (Figure 4.2). Transforming intuitions into falsifiable questions is a key skill for scientists and designers alike. How can people create questions that are novel (contain new information), useful (relate to and potentially extend existing scientific knowledge), easy to answer, and specific (relate to only one topic)? Such questions can potentially accelerate research in nascent scientific domains, such as the human microbiome.

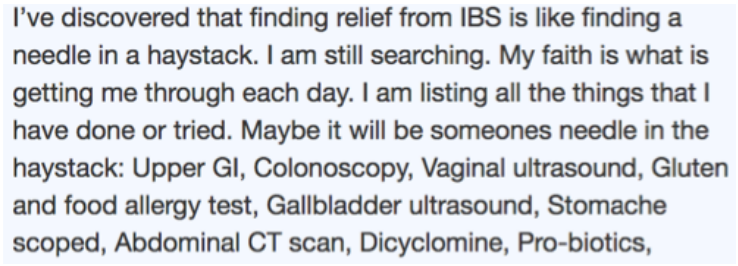
I've discovered that finding relief from IBS is like finding a needle in a haystack. I am still searching. My faith is what is getting me through each day. I am listing all the things that I have done or tried. Maybe it will be someones needle in the haystack: Upper GI, Colonoscopy, Vaginal ultrasound, Gluten and food allergy test, Gallbladder ultrasound, Stomache scoped, Abdominal CT scan, Dicyclomine, Pro-biotics,

Figure 4.2: This post to a Mayo Clinic forum shows how people seeking advice online combine many ideas into one post..

This paper contributes (1) the Learn-Train-Ask method for people to perform personally meaningful scientific work by sharing personal insights and receiving feedback from others, and (2) its embodiment in Docent — a novel crowdsourcing system for causal scientific questions. Docent enables novices to ask useful questions by learning domainspecific content, undergoing process training to develop task-specific skills, and collaborating with online peers. A between-subjects study evaluated this new method by measuring the quality of participants' questions to test causal scientific theories

about the human microbiome. 344 voluntary online participants from 27 countries — including participants from the American Gut Project, Open Humans, Coursera, and Reddit — signed up to share personally-relevant questions about the human microbiome. Participants created 399 questions, 75 (19%) of which microbiome experts found novel. Participants with access to process training generated hypotheses of better quality. Access to learning materials improved the questions' microbiome-specific knowledge. These results highlight the promise of performing personally-meaningful scientific work using massive online learning systems.

4.2 The Docent Social Computing System

The Docent social computing system enables people to create specific, personal hypotheses by providing: content learning & process training; a guided question-asking interface; and an online collaboration platform. Docent was designed via multiple iterations of early pilot studies. Docent's pilot participants were 50 lead users of the American Gut Project & Health Data Exploration workshop (hdexplore.calit2.net). Early participants used the website to provide in-person feedback about the interface. As the system matured, later participants provided both explicit online & in-person feedback along with usage data that led to a number of improvements. For instance, a pilot session led to the idea to enable editing others' questions to improve clarity (especially for non-native English speakers).

The Docent web application is built with Meteor (meteor.com), and extends Gut Instinct [131] that also leverages learning materials, but does not address causal theory generation. The front-end uses BlazeJS (blazejs.org) and is styled with Materialize (materialize.css).

4.2.1 Learn-Train-Ask: From intuitions to hypotheses

Docent embodies three main principles. The first is two-way integration of learning and asking questions for improved conceptual understanding of the microbiome. Novel, domain-specific work (such as asking questions for micro-biome discovery) needs to integrate a novel idea with existing knowledge, perhaps even using specific terms/metrics in the process (e.g. Bristol stool scale for quality of bowel movement). To forge a two-way link between learning and asking questions, Docent provides online lectures and feed-back on the questions people create using scientific material. For instance, for a question about the effects of probiotics on mood among people suffering from gastrointestinal diseases, Docent would provide feedback using lectures about probiotics, gastrointestinal diseases, and the gut-brain axis. The two attributes that training questions seek to model are a) that others can answer them, and b) that each addresses a single topic. Training helps participants get a feel for how precise questions should be: overly vague and overly specific terms both reduce a question's utility. To link many ideas with one cause, a sequence of questions can iteratively refine a hypothesis space. For instance, a question linking probiotics use to bowel movements might begin by asking how frequently people consume probiotics and in which form, following up by asking about bowel movements. Third, Docent provides clear success criteria: creating useful questions. Docent converts question-asking and answering into an engaging social interaction by enabling people to participate in multiple ways, such as by asking questions, adding follow-ups, editing questions to improve clarity, or responding to questions. A new user can access the entire Docent system only after adding a question.

4.2.2 Learn content: Integrate concepts with insights

Docent's learn module teaches people about the gut microbiome using online lectures, lifestyle questions, feedback from scientific material on questions, and guessing potential mechanisms (Figure 4.3). We describe each below.

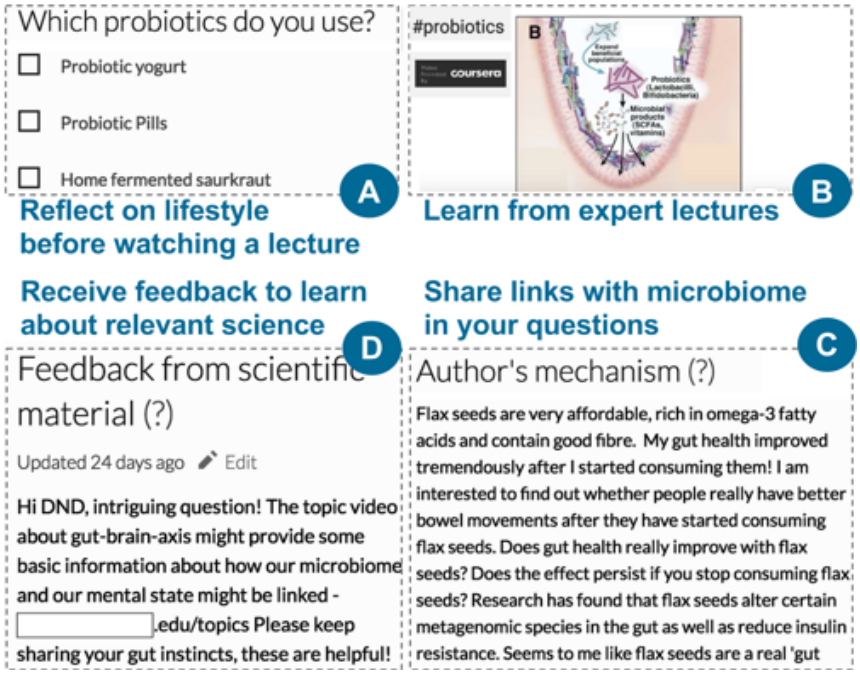

Figure 4.3: User flow of Docent learning module. Participants reflect on their lifestyle by (A) answering a personal question before (B) watching an online lecture about probiotics and the microbiome. Participants (C) must propose a mechanism when adding a question and (D) they receive feedback from scientific material on their questions. Probiotics was one of 16 topics for which Docent provided 5min long expert lectures.

Online lectures to improve conceptual understanding: The human microbiome is nascent, yet fast-growing. There is a lot of room for new contributions, but few people have up-to-date and accurate knowledge, even to the extent that it exists. People ask more questions when the learning material causes inconsistencies in their understanding of a topic [123]. Since few people know about the microbiome, Docent uses introductory learning material curated from Gut Check, a Coursera MOOC [93], rather than scientific papers that may be too abstruse. Apart from introductory material about the microbiome,

Docent also provides learning material about specialized topics, such as gastrointestinal diseases & the gut-brain axis, to engage people in guided discovery learning based on their interests and health conditions (Figure 4.3B).

Personal questions to improve reflection: Prompting participants to explicitly reflect on the learning topic can increase curiosity and question-quality [3, 104]. Before watching a lecture about the microbiome, Docent invites people to answer questions that make them reflect on the connection between the learning material and their lifestyle (Figure 4.3A).

Guessing a mechanism to reflect on question and knowledge: Docent asks people to guess mechanistic explanations for how the microbiome can play a role in answering their questions (Figure 4.3C). This is intended to help users learn by connecting personal observations with existing knowledge [147].

Feedback from scientific material: Rapid, relevant feedback improves quality [68]. The first author provided links to scientific papers and web content in response to people's questions. To integrate questions with Docent's learning material, they also received feedback on their questions using links to the Coursera MOOC lecture hosted on Docent (Figure 4.3D). Participants also added scientific papers to questions.

4.2.3 Process training: From intuitions to scientific questions

Docent uses three components to help people ask useful questions: a training guide, expert examples, and a question-asking checklist (Figure 4.4).

Training guide to identify useful features: New Docent participants must add a question before accessing the entire system (learning, training, and the GutBoard). Before asking a second question, however, a user needs to complete the training guide: people learn about five features of successful questions; train by identifying these features in two sample questions (Figure 4.4A,B); and then immediately ask a question. Training

draws on successful techniques from crowd work and peer assessment by using gold standards [88] and rubrics [7]. When adding a question, a checklist reminds participants about the features of good questions (Figure 4.4D).

4.2.4 Ask well-framed questions

Good questions specify a cause and associated effects.

Two-step question format to separate cause and effect: Docent questions comprise two parts: a top-level question identifies a cause (e.g., frequency of consuming probiotics). A follow-up question links the cause to a specific insight from the user (e.g., effects of consuming probiotics). The creator can add multiple follow-ups to link a cause to many effects.

Templated options to reduce common errors: Poor and/or vague options can discourage responses, erode esprit de corps, and model bad behavior that others follow. To counter this, Docent provides popular templated multiple-choice options. These templates are editable, but providing templates helps people be specific.

Cues to improve question quality: These cues comprise alert messages when people add long or short options; notes about details needed in their questions; and restricting people from adding a question without providing a potential mechanism or comment.

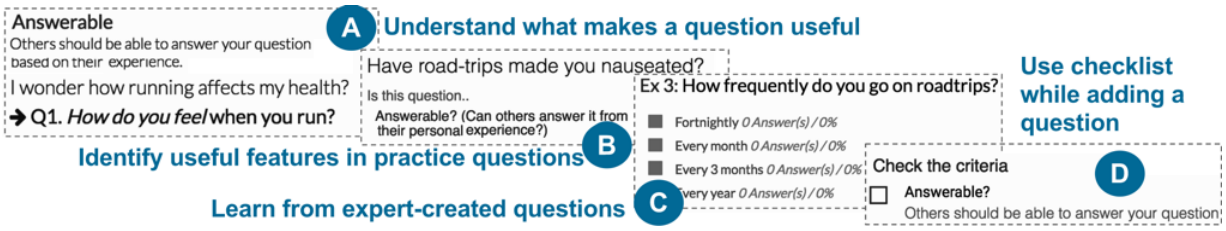

Figure 4.4: The Docent training process. People (A) learn what makes a question useful, (B) practice on sample questions, and (C) read expert-curated questions. (D) When asking a question, a checklist reminds people to ensure their question meets the criteria for useful questions. Answerability was one of 5 features for which Docent provided training.

4.2.5 GutBoard: Crowd responses, discussion, expert feedback

The GutBoard is designed for quick question traversal, easy response, and collaboration. Only the top-level question from each question is displayed: if a user is not interested in, say probiotics, they can simply skip that set. However, to access follow-ups, people need to answer the top-level question by selecting from the existing options or adding their own. To focus people's attention on specific questions, the first author starred promising questions that people can access from the Starred tab (Figure 4.5). Starring signifies that a question is likely of high quality or broad interest, and helps focus participants' answering efforts on them. Docent also enables people to bookmark questions of interest, so they can visit them again.

To de-incentivize lurking and increase engagement, the GutBoard shows only one question at a time; we call this sequential access. When people could see multiple questions simultaneously (parallel access), they skimmed through many questions without interacting with them.

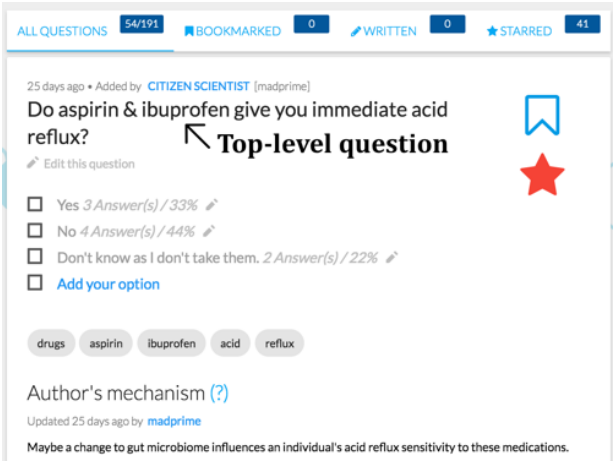

Figure 4.5: After answering a top-level question, participants can see follow-up questions, add new options, new follow-up questions, edit the question, guess potential mechanisms, and bookmark. A star (added by an expert) shows that this question may be promising for further enquiry by people.

4.3 Hypotheses: Effect of Learning and Training on Question Quality

Learn-Train-Ask scaffolds collaborative scientific question generation. We tested the following hypotheses in the context of brainstorming potential causal relationships in the human gut microbiome.

H1. Access to learning improves question's content. While people's questions are based on their experiences and curiosity, aligning them with what is already known about the microbiome can uncover novel insights. Alternately, learning can make people reflect less on their personal knowledge and more on institutional knowledge, reducing the novelty of their questions. A question is deemed to have good content if it is insightful (exhibits microbiome-specific knowledge) and novel (contains potentially new knowledge for microbiome science).

H2. Just-in-time training improves question's structure. We hypothesize that training helps people create useful questions for receiving feedback as well as for generating insights for researchers. A question is deemed to have good structure if it is easy for other participants to answer and focuses on one topic.

4.4 Study: Scaffolds for Better Questions

A between-subjects study compared the participants' question quality across four conditions: Learn, Train, Neither and Both. In the Learn condition, participants saw online lectures, answered personal questions, guessed mechanisms, and received feedback from scientific material on their questions (provided by the first author). The Train condition provided participants access to a training guide, expert examples, and checklist. In the Neither condition, participants did not have access to either the Learn

or Train step, while in the Both condition they had access to both. All participants had access to the question-asking and GutBoard collaboration module (with required condition-appropriate adjustments — e.g., participants in Learn and Both conditions were asked to guess the mechanism for their question, while participants in the other two conditions were asked to add a discussion comment). The GutBoard content was unique to the participants for a specific condition, to ensure that participants were not influenced by behavior in other conditions.

Method

Participants were randomly assigned to one of the four conditions. Each gave participants access to a condition-specific Docent. Each condition began with an introductory tour describing the significance of microbiome research and the importance of their contributions towards making discoveries. At the end of the tour, participants in every condition had to add one question (using identical question-asking modules) before moving on. Docent sent regular email reminders about site activity.

Participants

Recruitment: Participants were recruited via online publicity. To invite people especially interested in their microbiome, the American Gut Project emailed 550 participants. Docent was promoted on the American Gut Project's and their collaborators' Facebook and Twitter pages. Docent was added as a project on Open Humans (openhumans.org) — a platform where people donate their personal data for scientific research and participate in scientific experiments. Docent was posted on multiple subreddits pertaining to health and lifestyle (e.g., reddit.com/r/keto) and added as an optional assignment to the Gut Check Coursera MOOC [93]. Participation was voluntary and un-

paid; participants were entered in a raffle for an American Gut Microbiome kit (provided for $99 on American Gut's crowdfunding page (fundrazr.com/campaigns/4Tqx5)) on survey completion.

Measures

Dependent variables comprised structure, content, and creativity of questions (Table 4.1). American Gut researchers with multiple years of post-PhD expertise independently rated all 399 questions. (330 questions were rated by three; 69 were rated by two). The average ICC measure for 3 raters was 0.48 with 95% CI[0.42,0.54], ($F(328,656) = 3.73$, $p<.001$). Raters agreed that evaluating novelty was difficult since the nascent microbiome literature is rapidly growing.

Table 4.1: The five question quality criteria (rated as 0: no, 0.5: maybe, 1: yes). The 5-point sum represents overall quality.

Criteria	Operationalized as
Structure	**Answerable**: Is it a question about the participant?
	Specific: Does it ask about only one topic?
Content	**Insightful**:Does the question & discussion link to existing knowledge of the microbiome ?
	Novel: Is there a chance the world will learn something?
Creativity	Is it reasonably interesting/creative?

Raters were instructed to assign points for structure if it asked participants about a specific topic that they could answer. For example, *How often do you consume fermented foods?* was rated as both answerable by participants and specific, while, *Does our modern agricultural system affect our microbiome?* was neither answerable nor specific. Question content was the sum of insightfulness and novelty. Insightfulness addressed the quality

of the microbiome content in the questions. Novelty was assessed as the potential to create new knowledge in the microbiome field and operationalized as the lack of research papers about the specific question. For example, *Does consuming bone broth improve digestion?* was rated as both insightful and novel, while, *Can microbiome cure cancer?* was neither insightful nor novel. Broad questions related to well-studied topics or those fishing for links with the microbiome, such as the difference between generic vegetarian and meat-based diet were not deemed novel. A question was considered creative if it suggested an interesting idea without necessarily drawing it from personal experience.

Table 4.2: 344 participants completed the baseline exercise, continuing to the condition-specific intervention (*Learn*, *Train*, *Both*, or *Neither*). Participants with both learning and training generated a significantly larger number of question points than the other three conditions

	No Learn	*Learn*
No Train	**25.5** question points from (**13** of) **91** people (M= 0.28)	**31.5** question points from (**14** of) **90** people (M= 0.35)
Train	**26.5** question points from (**7** of) **74** people (M= 0.36)	**50** question points from (**16** of) **89** people (M= 0.56)

Results

344 participants completed the baseline exercise, continuing to the condition-specific intervention (Learn, Train, Both, or Neither). Participants who asked follow-on questions were scored as described above; those who did not were scored as 0. Table 4.2 shows, for each condition, the total number of participants, the number that provided questions, the total question points, and the average score across all participants. The overall quality of the questions generated in the combined Train+Learn condition is the

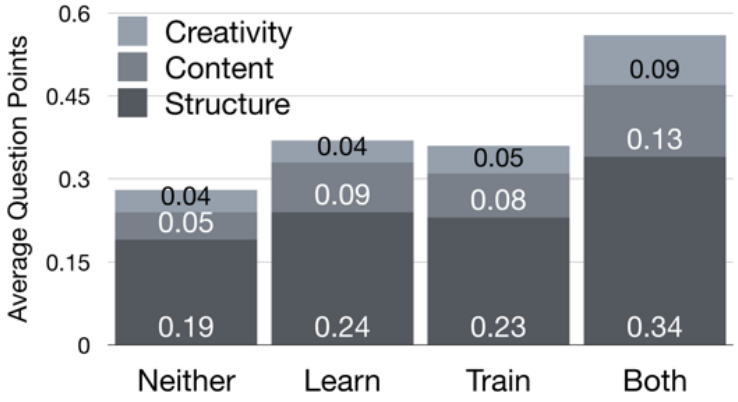

Figure 4.6: Training improved the overall question quality (p <.05) but learning did not (p < 0.1). H1: Training marginally improved question structure (p <.06). H2: Learning improved question content (p <.05).

highest (0.56 points) and appears to stand out from the others. We used a permutation test (a kind of bootstrapping method) to assess the statistical reliability of this apparent interaction, namely, whether the combined Learn-Train condition produces better questions than is expected from the independent main effects of the Learn and Train conditions. Indeed, a permutation test with 10,000 replications found that the observed differences in question points are different than the expected differences (generated by the main-effect marginals) as they fall outside the 95% confidence interval [-24.5, 24.5], p<.05. The three score components (structure, content, and creativity) were pairwise weakly correlated (r=0.32, p< 0.02; r=0.19, p< 0.17; r=0.33, p< 0.01).

Total question points: Training improved overall question quality (M=0.31, vs. M=0.47); a permutation test with 10,000 replications found that the observed difference in question points are different than the expected differences as they fell outside the 95% CI[-19.5, 19.5], p ¡.05. Learning did not improve overall question quality (M=0.32, vs. M=0.46); a permutation test with 10,000 replications found that the observed differences in question points are not different than the expected differences as they did not fall outside the 95% CI[-29.5, 29.5], p < 0.1. (Figure 4.6)

Structure: H1: Did access to training material (Train and Both conditions) improve

the structure of questions relative to not having access (Learn and Neither conditions)? Training marginally improved question structure (M=0.21, vs. M=0.29); a permutation test with 10,000 replications found that the observed difference in question structure points are different than the expected differences as they did not fall outside the 95% CI[-9.5, 9.5], p $<$.06.

Content: H2: Did learning material (Learn and Both conditions) improve the content of questions relative to not having access (Train and Neither conditions)? Learning improved question content (M=0.06, vs. M=0.11); a permutation test with 10,000 replications found that the observed difference in question content points fell outside the 95% CI[-9, 9], p $<$.05.

Creativity: Did training or learning material (Learn and Both conditions) impact the creativity score of questions relative to not having access (Train and Neither conditions)? Neither training (M=0.04 vs. M=0.07; 10,000-replication permutation test 95% CI[-4, 4], p $<$ 0.08) nor learning (M=0.05 vs. M=0.07; 10,000-replication permutation test 95% CI[-4, 4], p $<$ 0.2) improved the creativity score.

Table 4.3: Examples of questions created by participants

Quality	Sample Question
High	Have you ever eaten raw pumpkin seeds to eliminate parasites? (Structure: 2, Content: 1, Creativity: 1)
Medium	Do you get constipated when stressed? (Structure: 2, Content: 0.5, Creativity: 0.5)
Low	Does day of the week influence good vs. bad microbiota? (Structure: 1, Content: 0, Creativity: 0)

4.5 Discussion

4.5.1 The effect of learning and training on questions

As a check of random assignment, participants' required pre-intervention question was of comparable quality in all conditions. Participants in the Both condition scored higher total question points after the intervention. Training en-forced a tutorial when asking a question, and the add-question module presented heuristics for asking better ques-tions. This tight integration may have enabled people to focus their questions on a specific topic and frame their questions to be answerable by others. Moreover, the presence of learning material might have provided a useful setup and improved participant engagement leading to greater number of questions, and question points.

The study found a significant effect for learning on content ratings and a marginally significant effect for training on structure ratings. This asymmetry could be substantive: that learning improves content, but training only lightly improves structure. Alternatively, it could be a statistical mirage: the lower inter-rater reliability for content might show an effect if there isn't one. Inter-rater reliability was higher for structure (M=0.65; 95% CI[0.59,0.69] ($F(328,656) = 6.59$, $p < .001$)) than content (M = 0.11; 95% CI[0.04,0.18], ($F(328,656) = 1.37$, $p < .001$)). We hypothesize that content learning more clearly helped because domain knowledge provided insights and, potentially, ideas for questions, whereas the benefits for training heuristics were less clear. Some participants mentioned that understanding the learning material deeply wasn't their goal, which is corroborated by our experience designing and building the notes feature. Pilot feedback led us to create a time-annotated collaborative notes section alongside lecture videos. People could add notes about the lectures, raise clarifying questions with specific points in the video and answer others' questions. Collaborative, time-annotated notes below lecture videos have shown to improve social interaction and learning [109]. However, people

hardly used the notes features. After limited uptake, we removed these notes.

Participants watched 2.5 of 15 lectures on average. Moreover, in the Both condition, the combination of training and learning materials might have provided both useful content and sufficient structure for novices to utilize well. These results suggest that citizen scientists improve their work when presented with specific, just-in-time training. Self-guided question improvement may be valuable more broadly, as poor questions and question bloat are common problems in many social computing systems [173].

4.5.2 Which topics did the questions deal with?

The best questions had three features: they shared a clear insight from the participants' life (frequently elaborated upon in the discussion section of the question), enabled others to answer them from their lifestyle, and linked to known microbiome research (Table 4.4). Common question themes included probiotics; fermented foods; the consumption of fruits and vegetables in different forms; medicine usage; activities like exercises; stool quality & consistency. The three most popular lectures viewed discussed diet, antibiotics, and probiotics, hinting that either people were inspired by the lectures or at the very least, the lectures may have satisfied some of participants' curiosity about the links between their lifestyle and the microbiome. 50% of participants with learning mentioned that the lectures influenced their questions. Personal health was a big motivator; 78% of questions pertained to diseases (e.g., Irritable Bowel Syndrome), general health and well-being (obesity) or medication. 90% of survey respondents were motivated by personal health to ask questions. People created questions that linked activities with observable results (e.g., evacuation of bowel before colonoscopy with frequency of bowel movements after the procedure), but also raised questions that were driven more by curiosity about the microbiome: these questions inquired about their American Gut results, or the effect of a certain lifestyle choice (e.g., fasting) on microbiome, or

microbiome's effect on health (e.g., anxiety). 37% of all questions contained "hypotheses" i.e. they identified relationships between clearly identified variables (e.g., "Does eating probiotic foods reduce sugar cravings?"), while 46% only contained curiosity about the microbiome (e.g., "Hydrocolonic therapy change gut microbiome?" [sic]). Some of these questions were difficult even for experts to answer, since they are topics of active research (e.g., brain-gut axis [119]).

4.5.3 How novel are the questions?

75 of the 399 questions were found to be novel by the American Gut Project researchers. Novelty was defined as "Is there a chance the world will learn something?" The probiotic-sugar question above is novel because no published work addresses it directly. Other work on the sugar-microbiome relationship establishes plausibility [69]. Such questions meet Docent's primary objective: to uncover insights about topics where people's lived experiences provide them more knowledge than lab experts. Docent-like citizen science platforms can leverage people's lived experience to identify novel questions that experts have missed out and to evaluate these questions.

4.5.4 Emergent behavior, engagement, and growth

Docent offers more avenues for active collaboration than traditional web fora. People can create questions, answer and edit questions, create follow-ups, and guess potential mechanisms. Participants added a total of 2424 answers, 74 follow-up questions, 466 new options and 358 mechanism/ discussion comments. Discussion comments fell into three types, sorted by popularity: (a) sharing personal insights; (b) sharing potential mechanisms for the question; and (c) providing links to related online resources. People edited others' questions 119 times. Most edits were done by leaders and collaborators

Table 4.4: Distinct roles emerged as people added, edited, and answered questions

Role	Actions
Leader	Add questions, answer & edit others' questions, add follow-ups, discuss
Helper	Add & answer questions, add follow-ups
Participator	Answer questions
Lurker	Add questions but no collaborative work
Dropout	Add a question; never return

who attempted to clarify the question. None of the edits were reverted by the authors hinting that the edits were acceptable to them. Different ways to contribute creates different informal roles and behavior patterns in social computing systems, from leaders who perform all the activities to lurkers who may watch but not actively engage in collaborative activities (Table 4). Figure 4.7 shows the work distribution by roles in the Both condition.

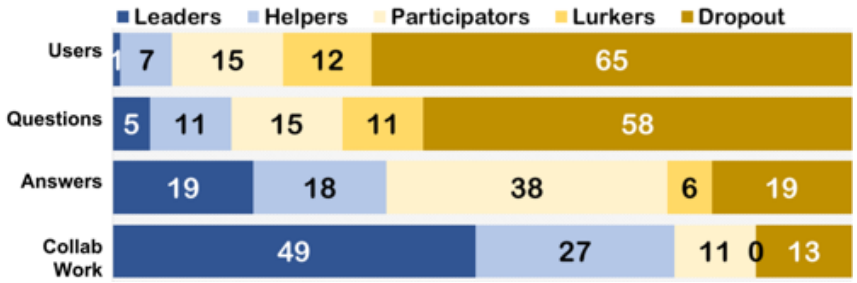

Figure 4.7: Distribution of role types for the 68 participants in the Both condition who added at least one question. The Leader formed 1% of participants but contributed 19% of the answers and 49% of all collaborative activity (adding follow-ups, editing others' question, adding options). Dropouts formed 65% of participants and added one question each (57% of all questions) but contributed to only 13% of collaborative work.
.

Prior citizen science platforms have demonstrated lurker and dabbler behavior [50]. Since people perform work on citizen science platforms, they require a prominent "circuit of engagement" [143]. Systems research comprises many choices; some we

evaluate, others follow from prior work, & some are hunches. To counter lurker behavior, Docent employs three techniques: First, Docent encourages members' self-selection using a strategy that hides Docent's content (lectures, training, and the GutBoard) until people add one question. Making even small contributions makes people feel more vested in the effort as a part of a community and removes their fear of performing a novel activity [138]. Extending these ideas can be useful for future work. Second, the GutBoard's continuous question updates provides social translucence [138]. Third, Docent sends regular email and social media updates to engage the community.

From Asking Questions to Building a Community

Docent's 20 email templates cover three areas: 1) user-activity specific e.g., reminder when someone added a follow-up question to a user's question; 2) condition-specific e.g., weekly emails about activity on the platform for the user's condition; 3) general reminders e.g., creating a username, or adding a question if participants had not done so already. Activity emails were sent each time a user's question received an edit, follow-up question, option, discussion comment, or when experts starred the question. Docent sent weekly general updates and links to tutorials.

A renowned microbiome expert recorded answers to popular questions which were subsequently mailed to participants and uploaded on social media channels. Docent maintained an active profile on Facebook and Twitter (240 and 224 followers respectively) by providing updates about platform activity, researchers' feedback on people's questions, and microbiome-relevant scientific articles and studies. Despite attempts to engage people, Docent saw high dropouts. Of 1630 participants, 907 (55%) took up usernames; and 344 participants (21%) added at least one question.

Two Optimizations Significantly Improve Scaling

To efficiently handle hundreds of participants, Docent initially renders only the first two questions while the remaining questions are rendered in the background. Docent also reduces page load times by storing markers in the browser's local storage for frequently accessed details, e.g., last-seen question. This enables Docent to fetch and show the next question on subsequent login rather than having to pull all questions and then traverse to the last-seen question.

4.5.5 Diversity & social behavior

Of the 344 participants who added a question, 219, who provided location information, hailed from 27 countries. 174 participants (80%) were from US or UK who asked 76% of questions. This is not without reason —Docent's online publicity was focused on the American Gut Project and its offshoot the British Gut Project. However, people in these countries may have higher socio-economic, educational, and technological status than the average global (or even, national) citizen (80% of our survey respondents had at least an undergraduate degree; an American Gut Project kit costs $99). MOOCs face a similar challenge: educated and affluent learners complete online classes at higher rates [91]. Science and humanity will likely benefit from diverse global participants [72], as in Lab in the Wild [136].

Diversity brings another design challenge. Diverse participants interpret prompts like Likert scale differently. People might also use terms that are not obvious to others. For instance, one participant asked about the frequency of "hoovering your home" which likely was lost on some participants. Since Docent participants hail from dozens of countries, terms need to be understood broadly. Participants could potentially flag such questions for clarification.

Willingness to share private information: Only 6% of survey respondents said they did not feel comfortable sharing personal insights. This is promising; however, there were questions that some may feel embarrassed to answer — e.g., questions pertaining to bowel movements, flatulence, and sexual activity. For these cases, questions and options can be rephrased — e.g., "Do you suffer from bouts of flatulence?" can be changed to "In the past week, how often have you suffered from flatulence?" enabling people to provide some useful information rather than entirely avoiding such questions. Moreover, specific communities' motivation can be focused on generating specific insights [23]. Docent already has many questions raised by participants suffering from different ailments; such patient groups may have specific insights as well as greater motivation to share them in exploratory projects [79].

Different strokes for different folks: Docent users were volunteers. Multiple survey respondents mentioned that busyness impeded their platform usage. We hypothesize that encouraging moderators may increase platform stickiness. We plan to create guides and have people try different roles which can boost creative thinking [154]. With a diverse participant set — health-hackers, MOOC learners, even some advanced microbiome students — people's attention can be put to specific tasks that they want to contribute to. For instance, MOOC learners may be more interested in unearthing mechanisms for people's hypotheses. Such differentiated roles, including different levels of editor, have contributed to the success of social computing systems like Wikipedia [99].

Validating hypotheses shared by participants: One early benefit of our work is that American Gut researchers are using the best Docent questions to potentially add/revise the metadata catalogue in the American Gut Project. Moreover, with the right online support, citizens can design and run experiments to test some of the hypotheses (e.g. probiotics reduces sugar cravings). Scaling causal reasoning could transform many domains. One interested party is the non-profit Open Humans platform where people

volunteer their personal data (e.g., microbiome/genomic data) and provide access to researchers to use their data.

This chapter investigated integrating learning and training with online scientific work. Experts rated 75 of 399 questions as potentially scientifically novel. Participants with access to process training generated hypotheses of better quality. Access to learning materials improved the questions' microbiome-specific knowledge. The Learn-Train-Ask method can be applied towards next steps in the scientific process, namely designing and running experiments. This study also illustrates the challenges of designing a social computing system that engages voluntary participants in performing personally-relevant scientific work. Such online experiences also naturally provide a problem-based learning setup for better learner engagement. Intui-tions gathered from a large online crowd can significantly scale up scientific inquiry by augmenting scientific expertise with insights and know-how drawn from the lived experiences of diverse individual people. We believe that dual-objective online systems that combine learning with personally meaningful work can enable people to meet their needs.

This chapter, in part, includes portions of material as it appears in *Docent: Transforming Personal Intuitions to Scientific Hypotheses through Content Learning and Process Training* by Vineet Pandey, Justine W. Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, and Scott R. Klemmer in the proceedings of the ACM Conference on Learning at Scale (L@S 2018). The dissertation author was the primary investigator and author of this paper.

Chapter 5

Collaboratively Running Experiments

People have creative scientific questions and folk theories; yet most lack the expertise to investigate them. How might people transform their questions into experiments that inform both science and their lives? This chapter introduces the Galileo system; it provides procedural support using three techniques: 1) experimental design workflow that provides just-in-time training; 2) review workflow with scaffolded questions; and 3) automated routines for data collection. We present three empirical investigations: a between-subjects experiment with students, a study with online volunteers across 16 countries, and a deployment with online volunteers across 8 countries. Procedural support yielded higher-quality experiments than watching lecture videos. People generated structurally-sound experiments on personally meaningful topics and ran them with participants from around the world.

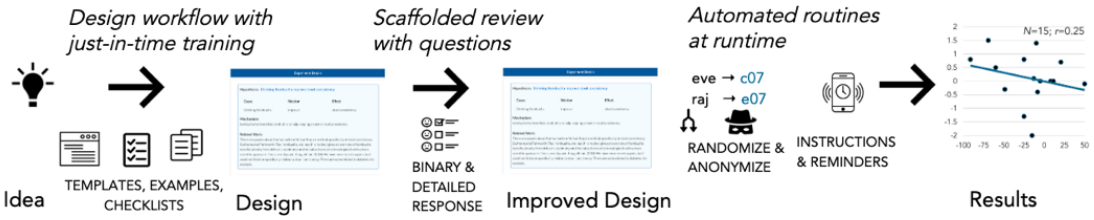

Figure 5.1: Galileo enables anyone to design and run experiments to test their intuitions. Experiment creators can invite anyone to review and participate in the experiment. Participants from around the world join experiments, follow instructions, and provide data in response to automated data collection reminders.

5.1 Experience to Experiments: 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 [30, 112]. A frequent error is mistaking correlation for causation [124]. 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?

Scientific experimentation features technical requirements and contextual choices that are inscrutable for a lay individual yet necessary for success [118]. While professional scientists and commercial ventures run experiments every day, with notable exceptions [34, 111], empirical papers from non-professionals are vanishingly rare. This biases the questions asked, studies run, and knowledge created [38, 72]. People have questions about their health, but lack the expertise and resources to scientifically investigate them. Broadening the pool of experimenters could help people investigate their curiosities, develop solutions to improve health and performance, and assist institutional researchers.

The main contribution of this chapter is *demonstrating that online volunteers can collaboratively perform scientific experimentation*. It does so in two main ways: 1) procedural

support for acquiring domain expertise using three techniques: experimental design workflow that provides just-in-time training, review with scaffolded questions, and automated routines for data collection; and 2) the Galileo social computing system that instantiates procedural support for citizen experimentation (Figure 5.1).

Three empirical investigations tested Galileo's approach. First, a controlled between-subjects experiment with 72 participants found that procedural support yielded significantly higher-quality experiment designs than lecture videos. Second, a deployment across 16 countries found that people generated structurally-sound experiments on personally meaningful topics. Third, in a field deployment, online users from three communities—kombucha, Open Humans, and beer—across 8 countries demonstrated that people designed, iterated on, and ran week-long experiments.

5.2 The Galileo Experimentation Platform

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

Despite a predetermined goal and a formalized process, experimentation requires making contextually-appropriate decisions [118]. Good experiment design is inherently user centered; designers need awareness of others' interpretation of their ideas and asks. Providing feedback on experiment designs requires knowing the success criteria and how to help improve.

Finally, successfully running an experiment requires managing multiple processes such as random assignment, anonymizing participant details, and sending instructions and reminders for data collection.

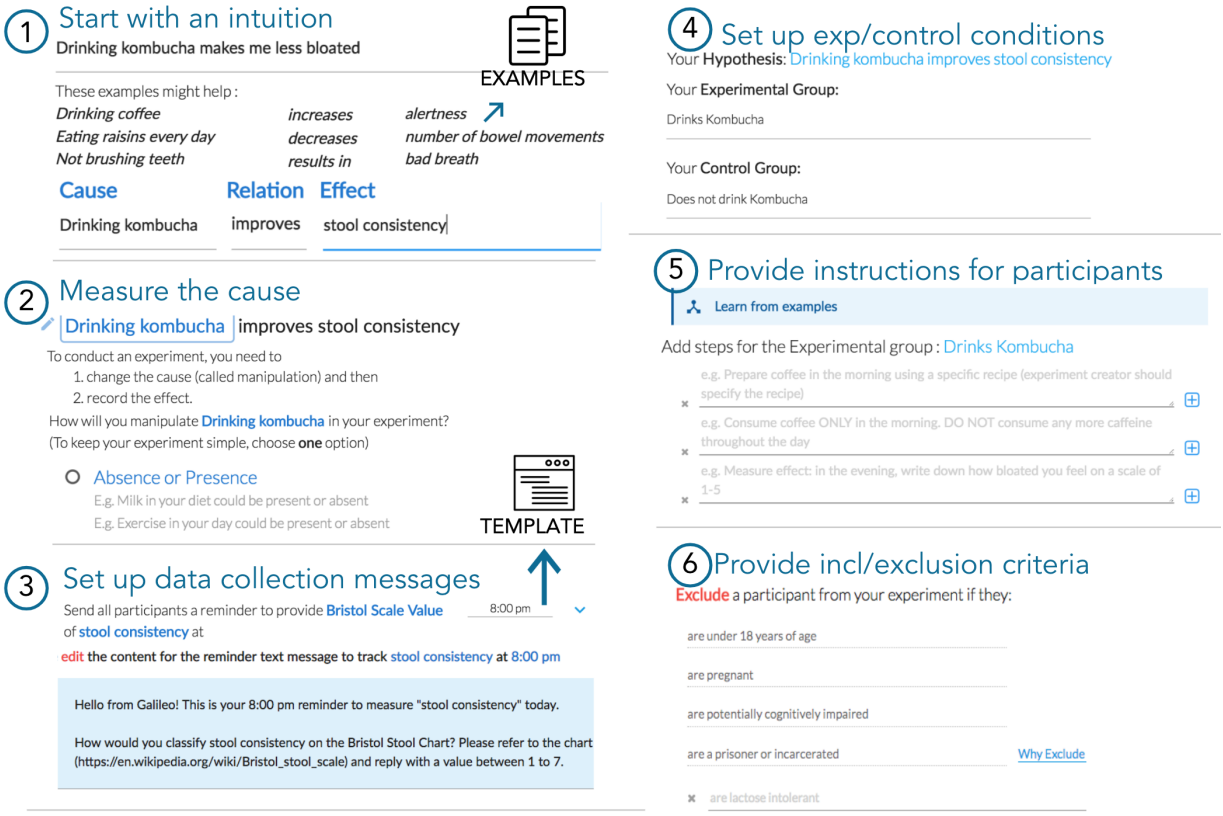

Figure 5.2: Galileo's design module helps people transform intuitions into experiment designs. It walks people through 1) converting an intuition to a hypothesis, 2,3) providing ways to manipulate/measure cause and effect, 4-5) specifying control and experimental conditions, and (not shown) providing inclusion/exclusion criteria.

5.2.1 Design-Review-Run: From intuitions to investigations

Galileo requires three roles for each experiment: designer, reviewer, and participant. Galileo offers procedural support for each: 1) a design workflow provides just-in-time training, 2) review with scaffolded questions, and 3) automated routines for runtime activities like data collection. Users form and refine with the help of contextual support and learning resources from the system.

5.2.2 Design an experiment from an intuition

People have many, often poorly-framed, hypotheses. Galileo's design workflow helps people harvest and sharpen them (Figure 5.2). Examples illustrate possible choices and how they relate; templates provide structure; and embedded videos explicate technical issues. Such procedural support can improve on-task performance [132]. 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.

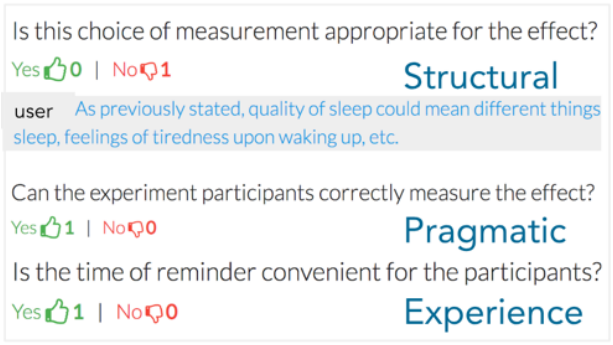

Figure 5.3: Reviewers walk through an experiment providing binary rubric assessments. A No response prompts reviewers to provide concerns and suggestions.

5.2.3 Review the design via feedback from others

Galileo 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 designer edits the experiment to address any issues. For research purposes, Galileo logs version changes. Reviewers provide both binary assessment and written responses to specific questions (Figure 5.3). These questions cover structure (e.g., accounting for confounds), pragmatics (e.g., measuring the real-world cause/effect), and participant experience (e.g.,

data reminder time). Reviewers are ineligible to be participants in the same experiment. Similarly, creators may not review their own experiment.

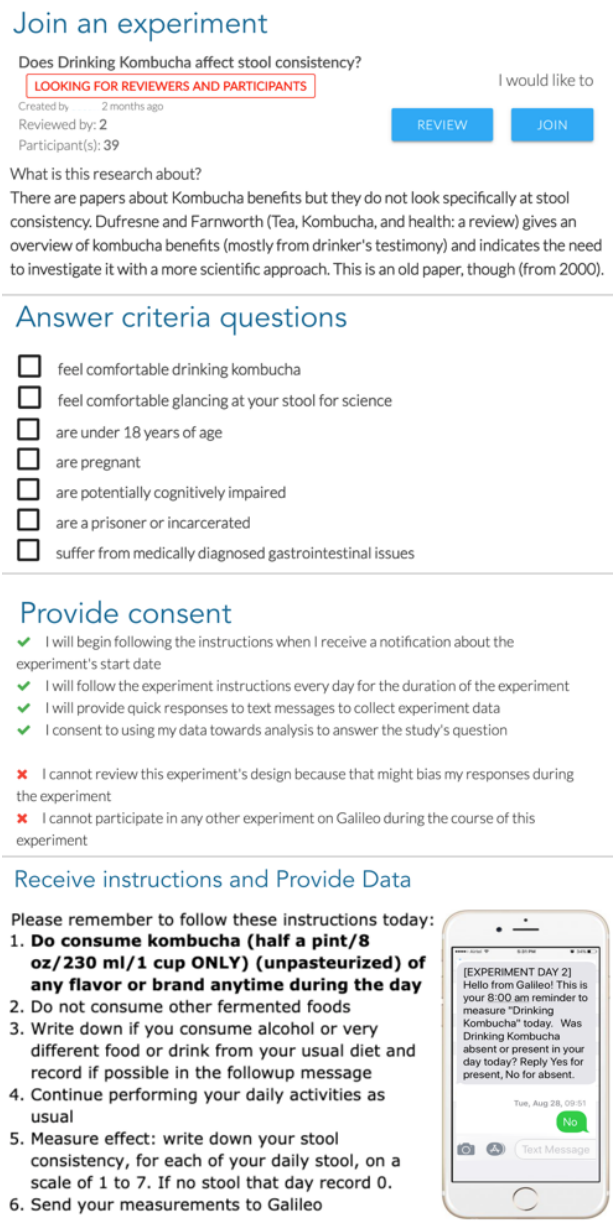

Figure 5.4: Join workflow for participants. 1) Participants can view a list of experiments. When they elect to join one, they 2) answer inclusion/exclusion criteria, 3) consent to following the provided steps, and 4) receive instructions. Participants receive daily, condition-specific requests, and respond with data and/or clarifying questions.

5.2.4 Run an experiment using procedural support

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

1. Randomized placement of people into conditions [118].
2. Maintain a per-experiment participant map ([usernames]→ [exp_id]) for anonymity
3. Collect and clean data (sending data collection messages and reminders at time zone appropriate times, parsing the responses, updating participant and experimenter views).
4. Prompt experimenters to perform tasks when conditions are met (e.g., setting the start date when enough participants have joined or reminding participants with missing data).

Participation comprises following instructions (e.g., drink kombucha) and providing self-report responses to platform queries (Figure 5.4). 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 real name or username. When the experiment ends, participants receive a summary of results. Participants can anonymously discuss the experiment at the end, so the experimenter can learn from their feedback.

The experimenter's dashboard lists tasks: answer clarifying questions, remind/thank participants, or look at trends in data (Figure 5.5). Experiments have a minimum participation count; there's no upper limit to the number of participants. People who sign up after a cohort begins are added to a waitlist. 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).

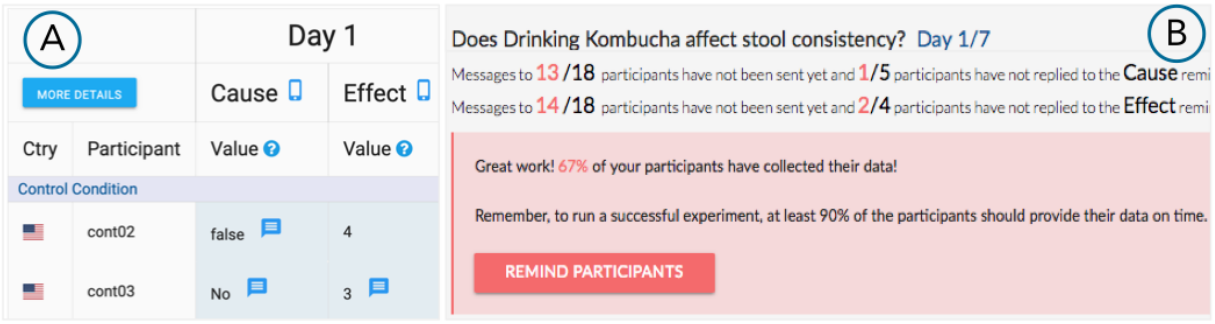

Figure 5.5: Galileo takes care of many experimenter responsibilities such as random placement of people, sending instructions and reminders, and cleaning and displaying data in both participant and experimenter dashboard. The dashboard enables experimenters to A) remind those with missing data; and B) see participants' data; and clarify questions raised by participants.

5.2.5 Designing the platform

More than 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 gathered feedback on the usefulness of the interface items and resources. Students in an undergraduate Psychology class (Introduction to Research Methods) also used Galileo in a 90-minute classroom deployment to rapidly design and review each others'experiments and receive feedback. We provide three examples of how pilot studies informed Galileo's design:

1. *Embedded written training over videos*: Early versions provided short, online lecture videos as the learning materials. Most users did not watch them end-to-end to

extract the step-relevant insight(s). In response, each step's content now offers written examples, which are easier to skim and refer back to.

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.

3. *Ease of glancing at participants' data*: Pilot users ran six trial experiments. The idea of a runtime dashboard (Figure 5.5A) 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.

5.2.6 Integrating procedural support 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.

Embedding Just-in-Time Support

Complex activities overwhelm learners' working memory because of their many interrelated pieces [48]. Recalling work from previous steps and frequent context-

switching are especially taxing [63]. Experts mitigate memory demands by integrating multiple elements into conceptual chunks [24]. A well-chunked interface can still require knowledge that novices lack. Galileo provides missing knowledge by providing learning materials in the interface. This *in-situ* embedding has three advantages: it is minimal [22], leverages teachable moments [70], and can be ability-specific [35]. 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 support.

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 [110] 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.

Three studies evaluated Galileo's approach: a controlled experiment to test procedural support's efficacy in the design workflow; a field study to test whether people can

create structurally-sound experiments based on personal intuitions; and a deployment to test whether people can design and run experiments.

5.3 Study 1: Experiment Comparing Procedural Support to Videos

To investigate whether procedural support helps novices design experiments, a between-subjects experiment tested the following hypothesis: ***Procedural support yields higher quality experiment designs than lecture videos***.

Procedural training, when successful, helps people solve unique problems with similar structure. It is perhaps best studied in K-12 mathematics instruction [141]. We hypothesize that participants who use interactive procedural support create better experiment designs than those who watch videos on the topic.

Method

Participants were randomly assigned to one of two conditions: *Videos* or *Galileo* (Figure 5.6). The *Videos* condition provided a playlist of videos about experiment design from a Coursera MOOC that operationalized the task-specific concepts [171]. The *Galileo* condition provided participants access to Galileo. Both provided the same content for creating 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.

The study asked participants to compose an experimental design for a personal intuition of their choosing. Each condition provided informational resources and a means to document their design (videos with a text document, or procedural support

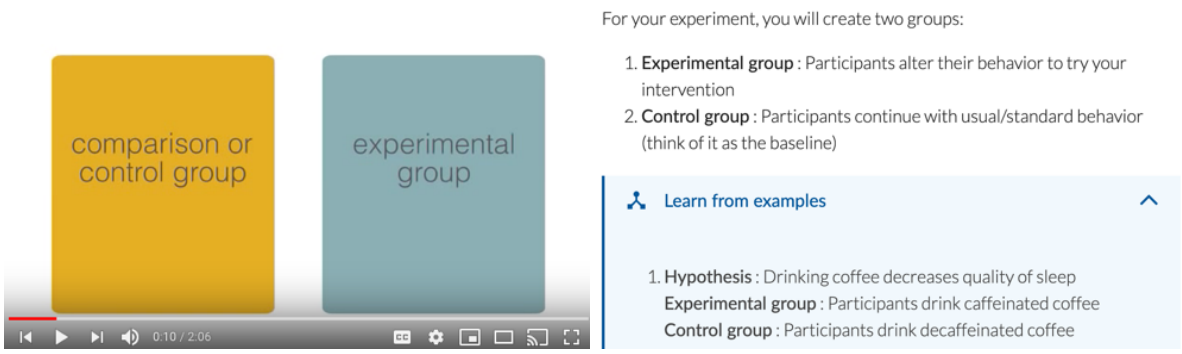

Figure 5.6: Two conditions for experiment. In the *Videos* condition, participants accessed videos about experiment design. In the *Galileo* condition, participants accessed Galileo tool (which included the videos if participants wanted to see them). Both conditions provided the same content.

with inline text fields). Participants were told that there was no lower or upper time limit. Each session comprised the following steps: consent, 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 asked why. An independent rater (a professor who teaches experiment design) blind to condition rated each participant's experiment using the rubric (Table 5.2).

Participants

Recruitment: 72 participants were recruited from a Western US Research University (Table 5.1). 11 had no prior experience with experiment design; 61 had taken a course or equivalent. Expertise was counterbalanced across conditions.

Table 5.1: Demography information for 72 participants (all undergraduate students). Some participants did not complete portions of the survey.

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

Measures

The independent variable is access to Galileo/videos. The study scored experiments via a 13-question rubric (Table 5.2), and recorded time taken. A blind-to-condition expert (a regular instructor of large, undergraduate courses on experiment design) provided the scores. Qualitative measures included how participants used the tool, where they faced challenges, and a post-experiment survey. A non-parametric Mann-Whitney test assessed the effect of condition on design quality.

5.3.1 Access to Galileo improved the quality of experiment design

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 5.7). Of

Table 5.2: Rubric for design-quality criteria for structure

***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? 2 points
	Do the conditions differ in manipulating the cause? 1 point
***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?
***Overall*: 1 point**	Can the overall experiment be run as is?

the 36 designs rated in the top half, 29 were from *Galileo* condition. *Galileo* participants performed better on five out of six sections (all except hypothesis). 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 = 29.0 mins) conditions; Mann–Whitney U = 734, n1 = n2 = 36, p = 0.33 two-sided.

Discussion

As Galileo aims to improve tasks like experimental design, Study 1's primary dependent variable was quality (as opposed to learning gains). Online video resources represent a common status quo: contemporary and bite-sized yet still static resources. This comparison enabled us to observe how Galileo's procedural support changed design outcomes. *Videos* participants followed one of two strategies: 1) watch all the videos at once and then begin writing the experiment; or 2) begin designing the experiment and use the videos to fill in the gap when stuck. Like cramming, all-at-once watching floods

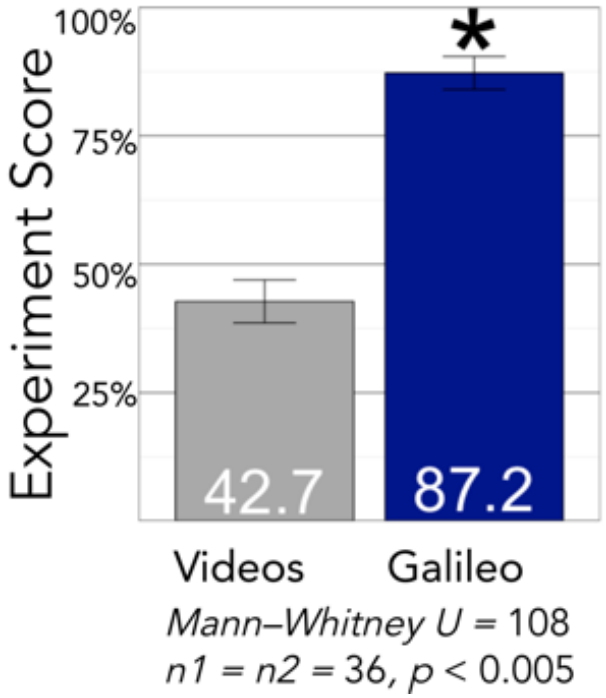

Figure 5.7: Access to Galileo improved the quality of experiment design

the mind, perhaps making it difficult to use seen ideas [97]. By contrast, the search-when-needed approach interrupts people's flow, replacing the attention on design with a task of locating needed information. *Videos*' lower score and our observations, in conjunction with the literature, suggest contextually-integrated approaches like procedural support increase people's useful adoption of information.

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 [85].

Participants in both conditions expressed a lack of confidence in their chosen cause/effect measures. Some spent over 15 minutes searching for measures: one found

a formal sleep-quality scale from Stanford researchers. Participants in 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”.

Limitations

This experiment found procedural support to yield higher-rated designs than watching videos. Important direction for future work will be to compare different approaches to procedural support, and exploring additional measures (e.g., novelty).

5.4 Study 2: People Design & Review Experiments Online

The first study evaluated the efficacy of procedural support for designing experiments. The second study 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 5.2), review existing experiments (see Figure 5.3), or join an experiment (see Figure 5.4).

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 [121]. 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 [93]. Participation was voluntary and unpaid.

Table 5.3: Rubric for design-quality criteria for Structure, Content, and Novelty

Structure	*Described in Table 5.2*
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?
Novelty	Is there a chance the world will learn something: absence of published research for this question?

Measures

Measures comprised structure, content, and novelty of experiment designs (Table 5.3) and usefulness of reviews. Raters with training in experiment design independently rated participants' work, then discussed them to form a shared view of assessment. Next, each independently rated all experiments. The final score is the mean of their independent ratings. Moderate reliability was found between the two raters' measurements [96]; m(ICC) = .62, 95% CI [.45, .75], ($F(64,64)= 4.33$, $p<.001$.

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

Results

5.4.1 People design structurally-sound experiments, and draw 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.

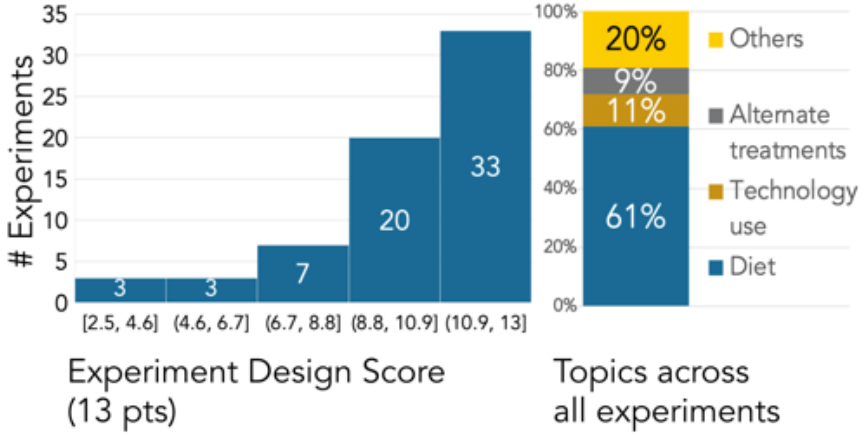

Figure 5.8: A) Most experiments were structurally-sound, scoring high on the structure rubric. B) Most experiments drew from personal experiences

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 5.8). 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."

17% of 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 citizens and microbiome researchers alike.

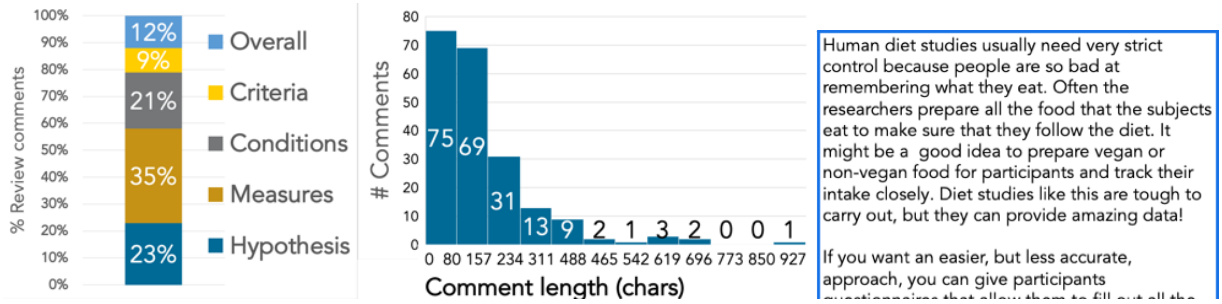

Figure 5.9: Summary of review A) Review comments were broadly distributed across all components of experimental design. B) Review comments ranged from 3 chars "yes" to 871 char long descriptions. C) The longest review comment described multiple problems with an experimental design while providing numerous actionable suggestions.

5.4.2 Reviewers use domain knowledge to improve designs and advocate for participant experience

158 review comments (77%) were rated useful; incorporating them would improve the experiment. Average comment length was 140 characters ranging from 3 characters (*yes*) to 871 characters (Figure 5.9B,C). Most comments were direct responses to a 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"?*)

5.5 Study 3: People Design, Review, & Run Experiments

The previous two studies found that people generated novel, structurally-sound experiments. Might they successfully run experiments with others? Participants from three communities—Kombucha, Open Humans, Beer—designed and ran experiments (Figure 5.10).

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 [28]. While there is

widespread belief that kombucha "benefits the gut", there is little published empirical evidence for these claims [49]. The experimenter hypothesized that kombucha supplies beneficial probiotics that help maintain normal stool consistency, and designed a between-subjects experiment.

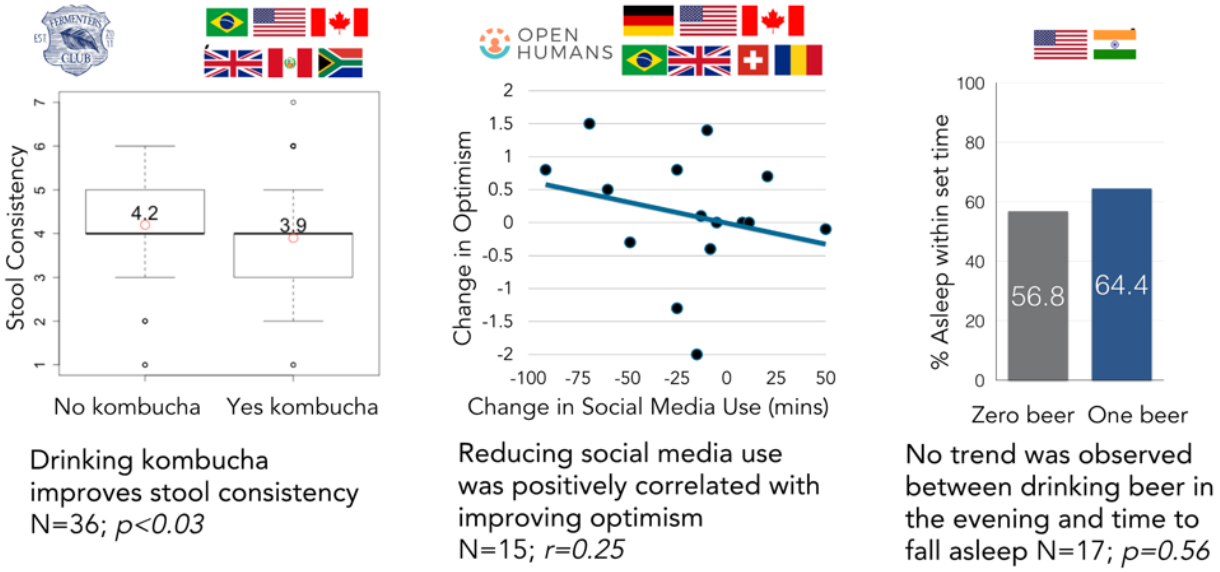

Figure 5.10: Three communities—Kombucha, Open Humans, Beer—designed and ran experiments; each ran for a week. The flags represent participants' nationality.

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 [37], 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 [135]. Alcohol helps people fall asleep but disrupts the REM cycle [47]. Still, it can be more convincing to see the evidence oneself. The experimenter (a graduate student) tested the effect of beer on sleep time with a between-subjects

experiment.

5.5.1 Before the experiment: Design, review, pilots, and finding participants

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 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 [168]. 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 (Fig-

ure 5.3). At the end of review, the three experiment designs used appropriate measures, provided a minimal-pairs design, tracked confounds, and provided appropriate criteria for participation.

Pilots: Three lessons emerged. First, some participants were loath to look at their stool. Since viewing one's stool is necessary, the experimenter added an inclusion criterion 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 the experimenter abandoned the study.

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

5.5.2 During the experiment: Retention and data collection

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 11). 78% of dropouts occurred in the first 48 hours. The reasons participants reported for dropping out included lack of interest, holidays, and work

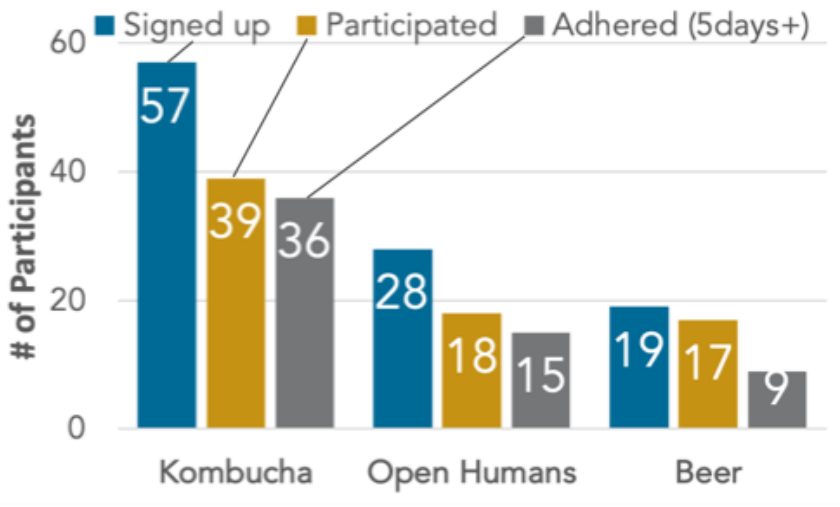

Figure 5.11: Dropout and adherence Rates across the three experiments. After signing up, a smaller fraction of people participated in Kombucha (68%) and Open Humans (63%) experiments than Beer (90%). However, those who participated reported greater adherence in Kombucha (92%) and Open Humans (83%) compared to Open Humans (50%). Reasons for non-adherence included being busy, annual leave, and brewers needing to check on the taste of Kombucha.

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 5.11). 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.

Data Collection: Most American participants selected text solicitations (86%);

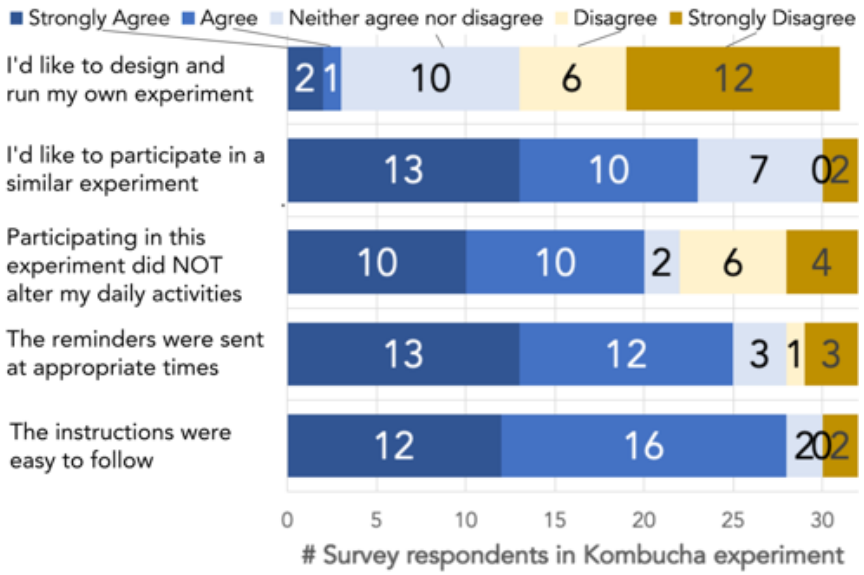

Figure 5.12: Participants in the kombucha experiment reported an overall positive experience expressing an interest to participant in another similar experiment (23/32). Most found the instructions easy to follow (28/32) and the reminders sent at appropriate times (25/32).

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

Clarifying questions: The experiment requested that all participants adhere to the protocol as much as possible without harming their health. Participants could ask the experimenter (via the platform) if confused. Participants' clarifying questions focused on measurements (e.g., measuring stool consistency once during the day or multiple times) and specific lifestyle choices (e.g., consuming probiotics while drinking kombucha?). Participants in kombucha experiment reported an overall positive experience (Figure 5.12).

5.6 Reflection

Our results point at three challenges in democratizing experimentation: 1) the three experimenters had advanced degrees, 2) two of the three completed experiments were underpowered, and 3) experiment participants demonstrated varying levels of adherence.

5.6.1 Do successful citizen-led experiments require prior expertise?

While an advanced degree is not a prerequisite, having one confers an advantage. This is unsurprising; contributions to open access web platforms are rarely uniform across educational levels. MOOCs are disproportionately completed by learners from more-affluent and better-educated neighborhoods [67], and 73% of citizen scientists and Wikipedia contributors have advanced degrees [126, 167] .

Why were all three experiments run by people with advanced degrees? One reason could be self-selection: those without prior expertise in experimentation might have opted out of running an experiment. This is weakly supported by data: all 36 participants in the kombucha experiment enjoyed the experiment and wanted to participate in more experiments (Figure 14). However, only two participants wanted to design and run follow-up experiments; both have an advanced degree. While simply asking people to contribute might work for traditional citizen science projects, experimentation might be a bigger leap. We suggest two improvements. First, reduce effort by providing ready to run experiments; common health and lifestyle topics such as coffee consumption and sleep might be good candidates. Running a sample experiment enables people to pilot the platform before testing their ideas while also potentially making them comfortable with the idea of experimentation itself. Second, support a growth mindset [46]: e.g. the platform can emphasize that anyone can learn how to run an experiment.

Another reason for experimentation by those with advanced degrees could be their awareness of potential participants. All three experimenters had access to people who were interested in similar topics; e.g., the Open Humans experimenter received both participants and feedback for their idea from the group's slack community. Such *affinity spaces* are known to provide potential participants as well as social support [58]. To tackle this, the design workflow can nudge the creator to start their experiment design by thinking of topics relevant to their social connections.

5.6.2 Guidance techniques to enable citizens to recruit others

Two of the three completed experiments were underpowered. Citizen experimenters learned what many scientists know: recruiting participants is time-consuming. This suggests that a good experimental design is not enough and recruiting is the next challenge for citizen scientists on their way to develop meaningful knowledge. While the absence of shared knowledge with experts can sometimes give novices' work a boost (e.g. identifying green pea galaxies on Galaxy zoo), it's less useful when the lack of knowledge is a hindrance. Tools for training and collaboration can help by clearly conveying the importance of getting enough participants; enabling experimenters estimate what "enough" is; and providing sources and strategies to recruit participants.

Citizen experimenters aren't as ardent about sufficient participation numbers as professional scientists. One important piece of technical knowledge is performing power analysis before running the experiment. Additionally, following the lead of data journalists [65], conveying results through real-world effect sizes—such as additional years you'll live—might be useful. Moreover, the experimenter need not find all the participants by themselves. Akin to a Clinical Research Coordinator, a separate recruitment role can help the experimenter rope in others to help out. Participants signed up for an experiment can also assist by suggesting others ala snowball sampling.

How might we help increase participation? Common reasons why people join *expert-led* experiments include [129]: 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 [33].

Adherence, however, remains a challenge. The opportunity to contribute to science is exciting (e.g. kombucha experiment participants mentioned this as a motivation). While altering one's lifestyle for a day might not be very difficult for many people, doing the same for a week (or more) might be tedious enough to entirely avoid participating, drop out after signing up, or not adhere to the instructions.

Drawing on findings from social computing and crowd-funding [74, 79], we suggest four remedies to improve both participation and adherence numbers: 1) increase participant trust by sharing more information about the experiment's goals, approximate effort expected, and the experimenter's biography; 2) implement *activation thresholds* to make social reciprocity explicit for group activities and to reduce potentially wasted efforts [26]; 3) leverage participation from communities with already strong ties and common goals; 4) allow people to pre-register for topics of interest so they might join relevant experiments created at a later date [12].

Our study did not provide experimenters or participants monetary compensation. Consequently, people's motivation is more intrinsic, which has benefits [127] (e.g. telling people the importance of their work improves performance [23]), but also empirically shows a high dropout rate. Compensation may help some citizen science experiments.

5.6.3 Design implications for knowledge work

We suggest three heuristics for systems that chunk complex knowledge work: separate roles, provide interactive guidance, and facilitate iteration. These ideas extend

minimalism to design learning experiences [159].

1. *Identify roles*. People, especially novices, often struggle to get started. Role-based approaches confer three benefits: clean delineation of responsibilities improves chances of task completion, clustering similar tasks reduces overhead and increases consistency; and people can decide their contribution levels. Procedural support operationalizes minimalism by co-locating tailored learning resources with specific steps.

2. *Provide procedural support for just-in-time domain-expertise*. A diverse audience might not interpret instructions consistently, or fail to translate textbook definitions to practice. Our results show that people are good at interpreting procedural support (like examples) for their use case. Creating such learning resources using the crowd or even learners themselves could reduce the effort needed [86].

   Checklists, cognitive aids, and tutoring systems exploit chunking as a means of onboarding new participants in a community of practice. For checklists, these chunks are usually static and expert-designed. A powerful benefit of interactivity is the opportunity for personalization. To reduce the time spent designing scaffolds and workflows, reuse lessons from other tools.

   Our first two heuristics focus on the authoring piece. Our third heuristic focuses on the reviewing piece. In contrast with cognitive aids and tutoring systems that "bake in" knowledge, our review step—like learnersourcing [86]—leverages the crowd for customized feedback. Structured reviewing—like Galileo—simultaneously discourages the lazy shortcut of superficial reviewing, and lowers the cognitive burden of providing deep, actionable feedback.

3. *Handle errors using iterations*. Most first drafts have errors. Feedback can be provided by experts [45, 145], peers [17, 100], software [42], or even oneself [17, 145].

Support iterations and pre-task training can counter concerns of superficial reviews. Why? Scaffolded questions and checklists help people reflect on their work at every step, especially when the system fails to automatically tackle inconsistencies for open-ended work.

5.6.4 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 range 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 [62], 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.

Because not all experiments are appropriate for people to run, some gatekeeping of citizen experiments might be necessary. 62 of 66 complete designs were posted online on Galileo for others to view; the primary author took 4 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 [41]. While the designer may be well-intentioned, consuming colloidal silver can cause irreversible damage such as skin discoloration, and the NIH has sued manufacturers for misleading claims [130]. 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.

This paper investigated citizen-led experimentation with novel procedural support. Three empirical investigations tested this approach. For us, the most striking result is that online volunteers collaboratively performed scientific experimentation without any expert help by drawing on their lived experience. Our work also illustrates the challenge of helping novices successfully execute a complex knowledge task. Specifically, finding and retaining participants and making the platform accessible to a broader audience emerged as key challenges. With systems that enable citizen-led experimentation, people can potentially match scientists' knowledge with their lived experiences to create insights both for themselves and for the scientific community. More generally, we hope that our work suggests ways to build systems that provide just-in-time domain expertise for knowledge work. Such systems can enable novice-led work that is personally meaningful, and situated in people's lived experiences.

This chapter, in part, includes portions of material as it appears in the submitted paper *Galileo: Procedural Support for Citizen Experimentation* by Vineet Pandey, Tushar Koul, Chen Yang, Daniel McDonald, Mad Price Ball, Bastian Greshake Tzovaras, Rob Knight, and Scott R. Klemmer. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Conclusion

This dissertation demonstrates that procedural guidance works well for scientific experimentation. This chapter provides future steps.

6.1 Systems & Domains

Domain experts make creative contributions like writing articles, curating museums, leading teams, and more. As in science, the number of experts in many domains is relatively small and their training relatively homogenous. Can procedural guidance support other genre of work?

6.1.1 Systems for end to end scientific work

Experimentation provides one method to create knowledge across the natural and behavioral sciences. Other ways to empirically evaluate hypotheses—case-control or cohort studies—require different support [125]. Furthermore, designing and running an experiment is one step among many in creating new knowledge. Scientists perform a range of activities including analyzing study data and communicating the results (e.g.

by writing a paper). One key challenge in such complex work is coming up with the initial design(s) that can be refined.

Case study: writing

EteRNA participants used system-provided templates to write up their results and share with others [107]. How might procedural systems assist? As is common for complex work, experts possess knowledge of the success criteria, mental scaffolds to help with writing, and access to other experts for feedback [81]. Showing specific knowledge to novices in the context of the work might be useful. Scientific writing follows different styles; let's consider two contrasting examples: the methods section and the discussion section of a paper.

The methods section provides specific details about how certain research was conducted. It describes the study hypotheses, choices of measures, method of enquiry, and all relevant decisions taken while running the study. Others should be able to perform these steps and (hopefully) find the same result. By using templates as the procedural guidance tool, a system can help people exploit the standard structure of the methods section and avoid standard mistakes. The discussion section of a paper is far less templated though since it summarizes multiple topics including the key ideas, the methods, and the results. A procedural guidance system for writing the discussion section can use multiple techniques; it can 1) identify the research question from a previous section; 2) use rubrics to prompt the writer to reflect on their claims; 3) show examples from other discussion sections; and 4) use checklists and peer feedback to improve clarity. The key insight here would be to help people explore the set of questions that a discussion section needs to answer. Such suggestions are preliminary. Rapid iterations immensely benefitted this dissertation's research.

6.1.2 Domains for citizen-led scientific investigations

This dissertation used microbiome research as a petri dish. Microbiome science is nascent, personal, and motivating. Other health related domains—like nutrition and Transcranial Direct Current Stimulation (tDCS)—are a good match. Transfering this dissertation's techniques to other domains raises design questions. First, different scientific domains might accept different methods of creating knowledge; e.g. some might rely strictly on Randomized Controlled Trials while others might prefer observational studies owing to the difficulty of randomization. Second, research communities create standard measures for popular outcomes of interest. Supporting standard measures provides three benefits: 1) it reduces citizens' efforts in coming up with a new measure, 2) it improves reliability and reproducibility, and 3) it helps people compare their results with prior research. For example, tDCS' effect on cognitive performance intrigues online communities, using standard Cognitive Ability Tests to recod the effects might help [114]. The correct implementation of standard measures can be especially useful for domains where self-reports are a primary way of collecting data; using confidence ratings and multiple questions could support citizen experimenters in collecting useful data. Implementing specific measures lends itself to interesting interface design challenges as well: people should understand the ask and provide correct data, all with minimum overhead.

6.1.3 Designing efficient procedural support

Computational problem-solving focuses on four key processes: abstraction, decomposition, generalization, and pattern matching [170]. The dissertation presents systems that use examples, checklists, and templates to embed procedural support. A promising avenue for future work might be to use procedural support to help people

with pattern matching for higher order tasks.

Useful procedural support for people needs to be simple, actionable, and potentially domain-specific. Examples or checklists that are too long and not directly linked to the task will see people struggle. Since people are better at identifying useful features than generating them [151], two ideas emerge for designing similar systems. First, start with "good enough" ideas, observe how people identify the useful features, and iterate to develop guidance techniques that lead to a more consistent and correct interpretation across people. Second, textual instructions provide one low-effort way to embed procedural support; providing examples using expert-created videos can be useful as well. Complex tasks such as laboratory work could benefit from short, specific tutorials.

6.1.4 Sources for procedural support

The systems described in this dissertation leveraged insights from experimentation in psychology [118] and design guidelines for social computing [138]. Wikihow provides a corpus of instructions for a wide range of activities from gardening to writing letters (wikihow.com). Online fora support crowd-generated resources that are distributed and unstructured: people share details of their goals, their attempts (including instructions), and even their evaluation of different techniques. Curating procedural resources from books and online resources can bootstrap online systems. Learnersourcing has demonstrated that learners can generate content that can be useful for others [83] while other systems have created new lexical categories from seed terms by mining fiction text [53]. Curating online resources has other advantages too: identifying structure in people's posts, research articles, or books identifies specific features that can bootstrap AI systems.

Experts know the rules and ways of domain-specific work [56]. How might we

leverage experts' knowlegde and experience to make their practices available more widely? For this dissertation research, microbiome experts were wary of providing feedback on citizens' work due to two reasons: 1) the time and effort invested, and 2) the potential of nudging citizens into accidently harmful work. To leverage and reuse their strategies, experts can lead by demonstration. Experts perform a task as part of their regular workflow; an annotated recorded version of the workflow can be programmatically reused by others [40]. Such a macro recording and annotation approach can be more passive or proactive and the annotation can be performed by demonstrators or annotators.

6.2 Patterns: Learning tools for end-users

Learning has always been lifelong. Rapid change and the ready availability of online resources make it even more so. This dissertation seeks to place learning experiences at the right time for people to use them. In the learning sciences, Bloom's taxonomy shows a hierarchy of ways of engaging knowledge, from remembering facts to evaluating theories (Figure 1.2). Traditionally, this diagram invites a discussion of classroom learning objectives. Such an order implies a potential research trajectory. Procedural guidance systems can double up as learning tools to provide a petri-dish to test important questions in learning science. Might such systems improve people's understanding of the domain and the task?

How well do goal-driven learning approaches translate online? Problem-based learning suggests starting with a problem that provides the context for learning new techniques [76]. Students construct a solution and—in many cases—the problems themselves. Discovery learning follows a similar model: starting with learners' curiosities and then providing the right mental tools to structure the discovery process. While sup-

porting people in storytelling in online classrooms improves engagement [133], learner motivation in online environments differs from traditonal classrooms [92]. Classrooms use test scores as external benchmarks but online learners might be motivated by their goals and care less about scores. Assessing competence at a task (or related tasks) can be one way to assess learner performance. This approach has the additional advantage of providing learners more time on tasks similar to the ones they're interested in.

The interaction between procedural guidance and social computing raises several research questions. Leveraging similarities between Bloom's taxonomy of learning and the hierarchy of social computing roles provides a potentially rich area of enquiry. Student interactions in online classrooms demonstrate similarity to role-taking on social networking platforms [90]. Legitimate peripheral participation [20] proposes that onboarding people with simple, low-risk tasks improves their participation and contributions. Organizing tasks in increasing order of learning complexity and supporting them with procedural guidance can potentially move students up the knowledge as well as engagement hierarchies (Figure 1.2).

6.3 Methods: Building a Science of Social Computing Systems

This dissertation identifies questions of prototyping, co-design, and emergent behavior as important issues and proposes a combination of theory, prototyping tools, and benchmarks.

6.3.1 Prototyping

How can we rapidly build, debug, and improve social computing systems? For instance, evaluating social computing systems is time-consuming: such systems embed multiple ideas; and usage phenomena are scale dependent and emergent.

All three systems presented in this dissertation have multiple features. Evaluating many features increases the designer's workload and requires more participants. Therefore, social computing systems benefit from more holistic evaluation feature testing. One approach might be to categorically separate measures for system evaluation (e.g. do people collaboratively create better questions using Docent?) from feature evaluation (e.g. does Docent's edit feature help people improve another user's question). A clear separation might help system designers sort the evaluation components in order of importance, assign different quality thresholds (e.g. controlled experiments vs observational evaluation), and communicate overall evaluation effectively to the research community.

Finding enough participants has been one bottleneck in developing systems for this dissertation. Every design-build-deploy cycle requires multiple iterations with groups of people. Friends and labmates aren't good proxy for real world users: social relations and prior knowledge of the research might bias participation. Paying crowdworkers doesn't work either because extrinsic motivation can skew results [23]. Furthermore, using non-representative population can increase threats to external validity of the study.

This dissertation research leveraged participation from American Gut and multiple communites. Asking community leaders to be early users provides multiple advantages: 1) they represent the system's intended users; 2) they have more experience with the community's working and goals; and 3) gatekeeping reduces chances of harm for others. However, this increases the workload for community leaders willing to help

out; creating low-effort channels for feedback can help.

6.3.2 Emergent behavior

How do participants organize and succeed in community-driven systems? How does such behavior evolve over time? Studies with small sample sizes can be poor predictors of emergent behavior in new systems. Furthermore, results are not even-keeled across users: some do more than others, many drop out, and take different roles [20]. One approach would be to develop a more stratified understanding of the results. Many health studies attempt to identify factors that influence people's individual responses. Social computing researchers can follow a similar model: rather than testing the efficacy of ideas with population-level measures, they might ask "Did this idea work for some people but not for others? If so, why?". Accumulating such insights across multiple research efforts can complement principles derived from psychology and organizational behavior.

6.3.3 Collaborating with domain experts

This dissertation features contributions from multiple communities—such as kombucha enthusiasts, Open Humans. The research papers feature 27 co-authors from five fields including microbiology, cognitive science, learning psychology, and systems. Many diverse efforts, including Precision Medicine Initiative (allofus.nih.gov), Zooniverse (zooniverse.org), and Foldscope Microcosmos (foldscope.com), might benefit from using this dissertation's principles to diversify and deepen citizen contributions. However, building such a network requires effort that feel tangential to research. How do experts across multiple domains contribute towards building systems that support domain-specific enquiry?

Working with multiple domain experts brings great value and learning opportunities but also multiple challenges. Developing a shared vocabulary helps. One approach is to use prototypes to ground the conversation across different domain experts. Concrete prototypes invite specific feedback from domain-experts that helps the system designer understand higher-level principles. Regular meetings can help catch early errors and also add to the trust [142]. For example, an early prototype with chat ideas floundered at the prototyping stage itself; experts mentioned that the effort of looking through people's chats for insights made this idea a non-starter. Templating support for automatically creating multiple prototypes for specific atomic tasks (asking questions, adding responses) can improve the rate of iteration.

6.3.4 Supporting global participation

This dissertation aims to complement global data collection with global distribution of expertise. Most Gut Instinct participants are from rich educated countries: 80% of Docent questions were from people in the developed world and all 3 experimenters had advanced degrees. People not represented on the platform across the world might have different ideas. How might such systems support a more diverse participation?

Efforts to scale and diversify participation can build on ideas that are *common* across cultures. For example, disease awareness months might provide a common timeframe for a global audience to collaborate on relevant issues. The ice bucket challenge raised awareness and donations for Amyotrophic Lateral Sclerosis (www.alsa.org). Understanding and building on *differences* in cultural norms is important too. Studies about human psychology have been traditionally run with a limited demography: overwhelmingly Western, educated, and residents of rich, industrialized, democratic countries [72]. Recent research has explored different cultural norms across device use and sensitive health topics. Lab in the Wild demonstrates that people across cultures

evaluate webpage designs in starkly different ways [137]. TeachAIDS has improved awareness of AIDS in India with a culturally-sensitive design that provided using locally tailored videos [150]. Furthermore, complex socio-economic factors can shape participation as well. E.g. in traditonally hierarchical societies, novices might be concerned about challenging experts. These examples suggest that successful, diverse participation might start with identifying what works in different cultures and amplify these ideas online. Finally, even when people might be motivated, they might lack the time/remuneration to learn new things and implement them in their lives. Might payment help? To reduce reliance on payments, one approach could be to pay people to start using the system and then remove the payment as participation becomes more stable [138].

6.4 Implications and Limitations

The systems developed as a part of this dissertation support people in generating hypotheses and running experiments. What are the implications of this research for social computing? What broader technical and societal transformations might we foresee?

6.4.1 Collaboration between novices and experts

Experts provide feedback and lead crowd efforts [45]. How might people support experts in complex knowledge work? Experts need help with multiple activities—such as participant recruitment—where citizens might have complementary skills and contexts. Citizens' efforts can also help experts refine the design space. For example, many novel ideas in health haven't been studied before; therefore, the effect size of such interventions is unknown and difficult to guess. Citizen-led experiments can help experts take better informed guesses, hopefully improving the odds of finding significant results. Citizens could also try reproducing current scientific research in fields with "reproducibility

crisis" such as psychology. Such collaborative efforts raise novel questions. Might working with novices help experts uncover their blind spots? How might such teams of experts and novices work through disagreement? Furthermore, how might credit be allocated in such settings?

6.4.2 Focus on processes over titles

Scientific research increasingly leverages larger teams with diverse expertise [172]. Since Gut Instinct users collaboratively designed and ran experiments, can they be called scientists now? Oxford dictionary defines a scientist as "a person who conducts scientific research or investigation; an expert in or student of science, esp. one or more of the natural or physical sciences". For all their useful contributions, this dissertation does not consider Gut Instinct users—experiment designers, reviewers, participants, hypotheses generators—to be scientists. One key reason is a lack of evaluation of users' conceptual skills; understanding how key concepts are linked is important for mastering complex knowledge work. For example, *absent* system support, can people design an experiment from scratch or recognize well-constructed experiments from poorer ones? Answering such questions can also uncover the limitations of procedural guidance systems.

The scientist-or-not question is also tied to the process of doing science: did Gut Instinct participants perform scientific work? This dissertation's evaluation demonstrates this to be true. However, there are many aspects to scientific work that does not lend itself well to workflows. For example, science is a contact sport [103]: ideas are gleamed from talks, discussed with others, and drawn from personal experiences. Famously, watching someone throw a plate in the air inspired Richard Feynman to pursue a question; answering this question won him the Nobel Prize in Physics in 1965. Creating such serendipitious encounters for inspiration and feedback can be powerful. Embedding social processes in online systems for complex work can help us draw ideas

about designing such interactions.

This dissertation research provides a vision and prototype systems for complex work by drawing on insights from interactive systems, social computing, and learning theory for enabling people to perform personally meaningful work. By doing so, this dissertation intends to democratize expertise and provide ways to meaningfully embed computation in society.

Bibliography

[1] 23ANDME. Something to Chew On. https://blog.23andme.com/23andmeresearch/something-to-chew-on/, 2016.

[2] 23ANDME. Genetics 101. 23andme.com/gen101/, 2017.

[3] ALEVEN, V., MCLAREN, B., ROLL, I., AND KOEDINGER, K. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education 16*, 2 (2006), 101–128.

[4] ALTHOFF, T., SOSIČ, R., HICKS, J. L., KING, A. C., DELP, S. L., AND LESKOVEC, J. Large-scale physical activity data reveal worldwide activity inequality. *Nature 547*, 7663 (2017), 336–339.

[5] AMAZON. Mechanical Turk. mturk.com, 2016.

[6] ANDERSEN, E., ROURKE, E. O., LIU, Y.-E., SNIDER, R., LOWDERMILK, J., TRUONG, D., COOPER, S., AND POPOVI, Z. The Impact of Tutorials on Games of Varying Complexity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)* (ACM, New York, NY, USA, 59-68. DOI: http://dx.doi.org/10.1145/2207676.2207687, 2012), pp. 59–68.

[7] ANDRADE, H. L., AND DU, Y. Student perspectives on rubric-referenced assessment. *Educational & Counseling Psychology Faculty Scholarship* (2005).

[8] ARONSON, S. J., AND REHM, H. L. Building the foundation for genomics in precision medicine. *Nature 526* (oct 2015), 336.

[9] ARSLAN, S. Traditional instruction of differential equations and conceptual learning. *Teaching Mathematics and its Applications: An International Journal of the IMA 29*, 2 (2010), 94–107.

[10] AUDUBON. Audubon Science. Using data to realize the best conservation outcomes, 2016.

[11] BERNSTEIN, M. S. Crowd-powered Systems, 2012.

[12] BERNSTEIN, M. S., BRANDT, J., MILLER, R. C., AND KARGER, D. R. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology* (2011), ACM, pp. 33–42.

[13] BERNSTEIN, M. S., LITTLE, G., MILLER, R. C., HARTMANN, B., ACKERMAN, M. S., KARGER, D. R., CROWELL, D., AND PANOVICH, K. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)* (2010), pp. 313–322.

[14] BODEN, M. A. *"The Story so far". The Creative Mind: Myths and Mechanisms.* Routledge, 2004.

[15] BOLTON, M. K. The Role Of Coaching in Student Teams: A "Just-in-Time" Approach To Learning. *Journal of Management Education 23* (1999), 233–250.

[16] BONIFAIT, L., CHANDAD, F., AND GRENIER, D. Probiotics for oral health: myth or reality?. *Journal of the Canadian Dental Association, 75(8).* (2009).

[17] BOUD, D. *Enhancing learning through self-assessment*. Kogan Page, London, 1995.

[18] BOULDER COLORADO, P. B. Project BudBurst: An online database of plant phenological observations, 2016.

[19] BRANDT, L. Fecal transplantation for the treatment of Clostridium difficile infection. *Gastroenterology & hepatology 8*, 3 (2012), 191–194.

[20] BRYANT, S. L., FORTE, A., AND BRUCKMAN, A. Becoming Wikipedian : Transformation of Participation in a Collaborative Online Encyclopedia.

[21] CARDAMONE, C., SCHAWINSKI, K., SARZI, M., BAMFORD, S. P., BENNERT, N., URRY, C. M., LINTOTT, C., KEEL, W. C., PAREJKO, J., NICHOL, R. C., AND OTHERS. Galaxy Zoo Green Peas: discovery of a class of compact extremely star-forming galaxies. *Monthly Notices of the Royal Astronomical Society 399*, 3 (2009), 1191–1205.

[22] CARROLL, J. M., SMITH-KERKER, P. L., FORD, J. R., AND MAZUR-RIMETZ, S. A. The minimal manual. *Human-Computer Interaction 3*, 2 (1987), 123–153.

[23] CHANDLER, D., AND KAPELNER, A. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior and Organization 90* (2013), 123–133.

[24] CHASE, W. G., AND SIMON, H. A. Perception in chess. *Cognitive psychology 4*, 1 (1973), 55–81.

[25] CHEN, Q., BRAGG, J., CHILTON, L. B., AND WELD, D. S. Cicero: Multi-Turn, Contextual Argumentation for Accurate Crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), ACM, p. 531.

[26] CHENG, J., AND BERNSTEIN, M. Catalyst: triggering collective action with thresholds. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (2014), ACM, pp. 1211–1221.

[27] CHI, M. T. H., GLASER, R., AND REES, E. Expertise in problem solving. Tech. rep., PITTSBURGH UNIV PA LEARNING RESEARCH AND DEVELOPMENT CENTER, 1981.

[28] CHILTON, S. N., BURTON, J. P., AND REID, G. Inclusion of fermented foods in food guides around the world. *Nutrients 7*, 1 (2015), 390–404.

[29] CHO, I., AND BLASER, M. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics 13*, 4 (2012), 260–270.

[30] CHOE, E. K., LEE, N. B., LEE, B., PRATT, W., AND KIENTZ, J. A. 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* (2014), 1143–1152.

[31] CONSORTIUM, T. H. M. P. A framework for human microbiome research. *Nature 486*, 7402 (2012), 215–221.

[32] CONSORTIUM, T. H. M. P. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature 486*, 7402 (2013), 207–214.

[33] COOPER, C. B., SHIRK, J., AND ZUCKERBERG, B. The invisible prevalence of citizen science in global research: Migratory birds and climate change. *PLoS ONE 9*, 9 (2014).

[34] COOPER, S., KHATIB, F., TREUILLE, A., AND AL., E. Predicting protein structures with a multiplayer online game. *Nature 466*, 7307 (2010), 756–760.

[35] CORBETT, A. T., KOEDINGER, K. R., AND ANDERSON, J. R. Intelligent tutoring systems. *Handbook of human-computer interaction 5* (1997), 849–874.

[36] COREN, M. J., AND COMPANY, F. Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks, 2011.

[37] COVIELLO, L., SOHN, Y., KRAMER, A. D., MARLOW, C., FRANCESCHETTI, M., CHRISTAKIS, N. A., AND FOWLER, J. H. Detecting emotional contagion in massive social networks. *PLoS ONE* (2014).

[38] CRAWFORD, J. T., AND JUSSIM, L. *Politics of Social Psychology*. Psychology Press, 2017.

[39] CROUCH, C. H., AND MAZUR, E. Peer Instruction: Ten years of experience and results. *American Journal of Physics 69*, 9 (sep 2001), 970.

[40] CYPHER, A., AND HALBERT, D. C. *Watch what I do: programming by demonstration*. MIT press, 1993.

[41] DANA LEWIS. Real-World Use of Open Source Artificial Pancreas Systems.

[42] D'ANTONI, L., KINI, D., ALUR, R., GULWANI, S., VISWANATHAN, M., AND HARTMANN, B. How Can Automatic Feedback Help Students Construct Automata? *ACM Trans. Comput.-Hum. Interact. 22*, 2 (2015), 9:1–9:24.

[43] DEBELIUS, J. W., VÁZQUEZ-BAEZA, Y., MCDONALD, D., XU, Z., WOLFE, E., AND KNIGHT, R. Turning Participatory Microbiome Research into Usable Data: Lessons from the American Gut Project. *Journal of Microbiology & Biology Education 17*, 1 (2016), 46–50.

[44] DOROUDI, S., KAMAR, E., BRUNSKILL, E., AND HORVITZ, E. Toward a Learning Science for Complex Crowdsourcing Tasks. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (2016), 2623–2634.

[45] DOW, S. P., KULKARNI, A., KLEMMER, S. R., AND HARTMANN, B. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)* (2012), ACM, pp. 1013–1022.

[46] DWECK, C. What having a "growth mindset" actually means. *Harvard Business Review 13* (2016), 213–226.

[47] EBRAHIM, I. O., SHAPIRO, C. M., WILLIAMS, A. J., AND FENWICK, P. B. Alcohol and sleep I: effects on normal sleep. *Alcoholism: Clinical and Experimental Research 37*, 4 (2013), 539–549.

[48] ENGLE, R. W. Working memory capacity as executive attention. *Current directions in psychological science 11*, 1 (2002), 19–23.

[49] ERNST, E. Kombucha: a systematic review of the clinical evidence. *Complementary Medicine Research 10*, 2 (2003), 85–87.

[50] EVELEIGH, A., JENNETT, C., BLANDFORD, A., AND AL, E. Designing for dabblers and deterring drop-outs in citizen science. *32nd annual ACM conference on Human factors in computing systems - CHI '14* (2014), 2985–2994.

[51] FARIDANI, S., LEE, B., GLASSCOCK, S., RAPPOLE, J., SONG, D., AND GOLDBERG, K. A Networked Telerobotic Observatory for Collaborative Remote Observation of Avian Activity and Range Change. *IFAC Proceedings Volumes 42*, 22 (2009), 56–61.

[52] FARZAN, R., AND KRAUT, R. E. Wikipedia classroom experiment: bidirectional benefits of students' engagement in online production communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)* (2013), pp. 783–792.

[53] FAST, E., CHEN, B., AND BERNSTEIN, M. S. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 4647–4657.

[54] F.LUX. f.lux: sleep research, 2019.

[55] FOX, S., DIMOND, J., IRANI, L., HIRSCH, T., MULLER, M., AND BARDZELL, S. Social justice and design: Power and oppression in collaborative systems. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2017), ACM, pp. 117–122.

[56] FRANCIS, S. T., HEAD, K., MORRIS, P. G., AND MACDONALD, I. A. The effect of flavanol-rich cocoa on the fMRI response to a cognitive task in healthy young people. *Journal of cardiovascular pharmacology 47* (2006), S215–S220.

[57] GAWANDE, A. The heroism of Incremental care, 2017.

[58] GEE, J. P. Semiotic social spaces and affinity spaces. *Beyond communities of practice language power and social context 214232* (2005).

[59] GELMAN, S. A., AND LEGARE, C. H. Concepts and folk theories. *Annu Rev Anthropol* (2011), 379–398.

[60] GIL DE ZÚÑIGA, H., JUNG, N., AND VALENZUELA, S. Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of computer-mediated communication 17*, 3 (2012), 319–336.

[61] GILL, S., POP, M., DEBOY, R., AND ECKBURG, P. Metagenomic analysis of the human distal gut microbiome. *Science 312*, 5778 (2006), 1355–1359.

[62] GODLEE, F., SMITH, J., AND MARCOVITCH, H. Wakefield's article linking MMR vaccine and autism was fraudulent. *BMJ 342* (2011).

[63] GONZÁLEZ, V. M., AND MARK, G. Constant, constant, multi-tasking craziness: managing multiple working spheres. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2004), ACM, pp. 113–120.

[64] GOOD, B. M., AND SU, A. I. Crowdsourcing for bioinformatics. *Bioinformatics 29*, 16 (2013), 1925–1933.

[65] GRAY, J., CHAMBERS, L., AND BOUNEGRU, L. *The data journalism handbook: How journalists can use data to improve the news*. O'Reilly Media, Inc., 2012.

[66] HACKER, S. B. H. *Duolingo: Learning a Language while Translating the Web. Ph.D Dissertation*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2014.

[67] HANSEN, J. D., AND REICH, J. Democratizing education? Examining access and usage patterns in massive open online courses. *Science 350*, 6265 (2015), 1245–1248.

[68] HATTIE, J., AND TIMPERLEY, H. The power of feedback [transfer argument]. *Review of Educational Research 77*, 1 (2007), 81–112.

[69] HAUKIOJA, A., SÖDERLING, E., AND TENOVUO, J. Acid Production from Sugars and Sugar Alcohols by Probiotic Lactobacilli and Bifidobacteria in vitro. *Caries Research 42*, 6 (2008).

[70] HAVIGHURST, R. J. Human development and education. *Oxford, England: Longmans, Green.* (1953).

[71] HEAD, A., GLASSMAN, E., SOARES, G., SUZUKI, R., FIGUEREDO, L., AND ANTONI, L. D. Writing Reusable Code Feedback at Scale with Mixed-Initiative Program Synthesis. In *Learning@Scale 2017* (2017).

[72] HENRICH, J., HEINE, S. J., AND NORENZAYAN, A. Most People are not WEIRD. *Nature 466*, July 2010 (2010).

[73] HINDS, P. J. The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied 5*, 2 (1999), 205–221.

[74] HUI, J. S., GERBER, E. M., AND GERGLE, D. Understanding and leveraging social networks for crowdfunding: opportunities and challenges. In *Proceedings of the 2014 conference on Designing interactive systems* (2014), ACM, pp. 677–680.

[75] JENNETT, C., AND COX, A. L. Eight Guidelines for Designing Virtual Citizen Science Projects. *Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP '14)* (2014), 16–17.

[76] JOHNSON, A., KIMBALL, R., MELENDEZ, B., MYERS, L., RHEA, K., AND TRAVIS, B. Breaking with tradition: Preparing faculty to teach in a student-centered or problem-solving environment. *Primus 19*, 2 (2009), 146–160.

[77] JOHNSON-LAIRD, P. N., AND OATLEY, K. Basic emotions, rationality, and folk theory. *Cognition & Emotion 6*, 3-4 (1992), 201–223.

[78] JORDAN RADDICK, M., BRACEY, G., GAY, P. L., AND AL., E. Galaxy zoo: Motivations of citizen scientists. *Astronomy Education Review 12*, 1 (2013), 1–41.

[79] KARKAR, R., SCHROEDER, J., EPSTEIN, D. A., PINA, L. R., SCOFIELD, J., FOGARTY, J., KIENTZ, J. A., MUNSON, S. A., VILARDAGA, R., AND ZIA, J. 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* (2017), ACM, pp. 6850–6863.

[80] KAWRYKOW, A., ROUMANIS, G., KAM, A., KWAK, D., LEUNG, C., WU, C., ZAROUR, E., SARMENTA, L., BLANCHETTE, M., AND WALDISPÜHL, J. Phylo: A citizen science approach for improving multiple sequence alignment. *PLoS ONE 7*, 3 (2012).

[81] KELLOGG, R. T. Professional writing expertise. *The Cambridge handbook of expertise and expert performance* (2006), 389–402.

[82] KEMPTON, W. Two theories of home heat control. *Cognitive Science 10*, 1 (1986), 75–90.

[83] KIM, J. *Learnersourcing : Improving video learning with collective learner activity. Ph.D Dissertation*. Ph.d dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2015.

[84] KIM, J., CHENG, J., AND BERNSTEIN, M. S. 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* (2014), ACM, pp. 745–755.

[85] KIM, J., NGUYEN, P. T., WEIR, S., GUO, P. J., MILLER, R. C., AND GAJOS, K. Z. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), ACM, pp. 4017–4026.

[86] KIM, J., NGUYEN, P. T., WEIR, S., GUO, P. J., MILLER, R. C., AND GAJOS, K. Z. 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* (2014), 4017–4026.

[87] KIRSCHNER, P. A., AND VAN MERRIËNBOER, J. Ten steps to complex learning a new approach to instruction and instructional design. 2008.

[88] KITTUR, A., NICKERSON, J., AND BERNSTEIN, M. The Future of Crowd Work. *Proc. CSCW '13* (2013), 1–17.

[89] KITTUR, A., NICKERSON, J., BERNSTEIN, M., GERBER, E., SHAW, A., ZIMMERMAN, J., LEASE, M., AND HORTON, J. The future of crowd work. In *ACM Conference on Computer Supported Coooperative Work (CSCW 2013)* (2013).

[90] KIZILCEC, R. F., PIECH, C., AND SCHNEIDER, E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge* (2013), ACM, pp. 170–179.

[91] KIZILCEC, R. F., SALTARELLI, A. J., REICH, J., AND COHEN, G. L. Closing global achievement gaps in MOOCs. *Science 355*, 6322 (2017), 251–252.

[92] KIZILCEC, R. F., AND SCHNEIDER, E. Motivation as a lens to understand online learners: Toward data-driven design with the olei scale. *ACM Transactions on Computer-Human Interaction (TOCHI) 22*, 2 (2015), 6.

[93] KNIGHT, R., METCALF, J., AND AMATO, K. Gut Check: Exploring Your Microbiome. Coursera., 2016.

[94] KNIGHTLAB. American Gut - What's in your gut?, 2016.

[95] KNIGHTLAB. American Gut Project. Login., 2016.

[96] KOO, T. K., AND LI, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine 15*, 2 (2016), 155–163.

[97] KORNELL, N. Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition 23*, 9 (2009), 1297–1317.

[98] KOTTURI, Y., KULKARNI, C. E., BERNSTEIN, M. S., AND KLEMMER, S. Structure and messaging techniques for online peer learning systems that increase stickiness. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale - L@S '15* (2015), 31–38.

[99] KRIEGER, M., STARK, E. M., AND KLEMMER, S. R. Coordinating tasks on the commons: designing for personal goals, expertise and serendipity. *Proceedings of the {SIGCHI} Conference on Human Factors in Computing Systems* (2009), 1485–1494.

[100] KULKARNI, C., BERNSTEIN, M., AND KLEMMER, S. PeerStudio: Rapid Peer Feedback Emphasizes Revision and Improves Performance. In *Learning at Scale* (2015).

[101] KULKARNI, C., WEI, K. P., LE, H., CHIA, D., PAPADOPOULOS, K., CHENG, J., KOLLER, D., AND KLEMMER, S. R. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI) 20*, 6 (2013), 33.

[102] LASECKI, W., MILLER, C., SADILEK, A., ABUMOUSSA, A., BORRELLO, D., KUSHALNAGAR, R., AND BIGHAM, J. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology* (2012), ACM, pp. 23–34.

[103] LATOUR, B., AND WOOLGAR, S. *Laboratory life: The construction of scientific facts*. Princeton University Press, 2013.

[104] LAW, E., YIN, M., GOH, J., CHEN, K., TERRY, M., AND GAJOS, K. Z. Curiosity Killed the Cat , but Makes Crowdwork Better. *Chi 2016* (2016).

[105] LAW, E., YIN, M., GOH, J., CHEN, K., TERRY, M., AND GAJOS, K. Z. Curiosity Killed the Cat, but Makes Crowdwork Better. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2016).

[106] LEE, D., LO, J., KIM, M., AND PAULOS, E. Crowdclass: Designing classification-based citizen science learning modules. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP '16)* (2016).

[107] LEE, J., KLADWANG, W., LEE, M., CANTU, D., AZIZYAN, M., KIM, H., LIMPAECHER, A., GAIKWAD, S., YOON, S., TREUILLE, A., AND DAS, R. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences 111*, 6 (2014), 2122–2127.

[108] LEE, Y.-C., LIN, W.-C., CHERNG, F.-Y., WANG, H.-C., SUNG, C.-Y., AND KING, J.-T. Using Time-Anchored Peer Comments to Enhance Social Interaction in Online Educational Videos. In *SIGCHI Conference on Human Factors in Computing Systems* (2015), pp. 689–698.

[109] LEE, Y.-C., LIN, W.-C., CHERNG, F.-Y., WANG, H.-C., SUNG, C.-Y., AND KING, J.-T. Using Time-Anchored Peer Comments to Enhance Social Interaction in Online Educational Videos. *Proceedings of the ACM CHI'15 Conference on Human Factors in Computing Systems 1* (2015), 689–698.

[110] LEVY, L., JONES, B., ROBERTSON, S., AND PRICE, E. D. Health Mashups : Presenting Statistical Patterns between Wellbeing Data and Context in Natural Language to Promote Behavior Change. 1–27.

[111] LEWIS, D., AND LEIBRAND, S. Real-World Use of Open Source Artificial Pancreas Systems. *Journal of Diabetes Science and Technology 10*, 6 (2016).

[112] LI, I., DEY, A., AND FORLIZZI, J. A stage-based model of personal informatics systems. *Proceedings of the 28th international conference on Human factors in computing systems* (2010), 557.

[113] LUTHER, K., COUNTS, S., STECHER, K. B., HOFF, A., AND JOHNS, P. Pathfinder: an online collaboration environment for citizen scientists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2009), ACM, pp. 239–248.

[114] MACAN, T. H., AVEDON, M. J., PAESE, M., AND SMITH, D. E. The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology 47*, 4 (1994), 715–738.

[115] MACKAY, W. E., APPERT, C., BEAUDOUIN-LAFON, M., CHAPUIS, O., DU, Y., FEKETE, J.-D., AND GUIARD, Y. Touchstone: exploratory design of experiments. *CHI '07 Proceedings of the SIGCHI Conference on Human Factors in Computing System* (2007), 1425–1434.

[116] MAMYKINA, L., MANOIM, B., MITTAL, M., HRIPCSAK, G., AND HARTMANN, B. Design Lessons from the Fastest Q&A Site in the West. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)* (2011), pp. 2857–2866.

[117] MARKUS, H., AND KITAYAMA, S. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review 98*, 2 (1991), 224–253.

[118] MARTIN, D. W. *Doing psychology experiments*. Cengage Learning., 2007.

[119] MAYER, E. A., KNIGHT, R., MAZMANIAN, S. K., CRYAN, J. F., AND TILLISCH, K. Gut Microbes and the Brain: Paradigm Shift in Neuroscience. *Journal of Neuroscience 34*, 46 (2014), 15490–15496.

[120] MAYER, R. E. Should There Be a Three-Strikes Rule Against Pure Discovery Learning? The case for guided methods of instruction. *American Psychologist 59*, 1 (2004), 14–19.

[121] MCDONALD, D., HYDE, E., DEBELIUS, J. W., MORTON, J. T., GONZALEZ, A., ACKERMANN, G., AKSENOV, A. A., BEHSAZ, B., BRENNAN, C., AND CHEN, Y. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems 3*, 3 (2018), e00031–18.

[122] MCDONALD, D., HYDE, E. R., DEBELIUS, J. W., MORTON, J. T., GONZALEZ, A., ACKERMANN, G., AKSENOV, A. A., BEHSAZ, B., BRENNAN, C., CHEN, Y., DERIGHT GOLDASICHA, L., DORRESTEIN, P. C., DUNN, R. R., FAHIMIPOURG, A. K., GAFFNEY, J., GILBERT, J. A., GOGUL, G., GREEN, J. L., HUGENHOLTZ, P., HUMPHREY, G., HUTTENHOWER, C., JACKSON, M. A., JANSSEN, S., JESTE, D. V., JIANG, L., KELLEY, S. T., KNIGHTS, D., KOSCIOLEK, T., LADAU, J., LEACH, J., MAROTZ, C., MELESHKO, D., MELNIK, A. V., METCALF, J. L., MOHIMANI, H., MONTASSIER, E., NAVAS-MOLINA, J., NGUYEN, T. T., PEDDADA, S., PEVZNER, P., POLLARD, K. S., RAHNAVARD, G., ROBBINS-PIANKA, A., SANGWAN, N., SHORENSTEIN, J., SMARR, L., SONG, S. J., SPECTOR, T., SWAFFORD, A. D., THACKRAY, V. G., THOMPSON, L. R., TRIPATHI, A., VAZQUEZ-BAEZA, Y., VRBANAC, A., WISCHMEYER, P., WOLFE, E., ZHU, Q., AND KNIGHT, R. American Gut: an Open Platform for Citizen-Science Microbiome Research. *bioRxiv* (jan 2018).

[123] MIYAKE, N., AND NORMAN, D. A. To Ask a Question , One Must Know Enough to Know What is Not Known.

[124] MUNROE, R. Correlation, 2009.

[125] MURAD, M. H., ASI, N., ALSAWAS, M., AND ALAHDAB, F. New evidence pyramid. *BMJ Evidence-Based Medicine 21*, 4 (2016), 125–127.

[126] NATIONAL ACADEMIES OF SCIENCES, E., AND MEDICINE. *Learning through citizen science: enhancing opportunities by design*. National Academies Press, 2018.

[127] NATIONAL COUNCIL FOR VOLUNTARY ORGANISATIONS, U. Why Volunteer?, 2018.

[128] NIELSEN, M. *Reinventing discovery: the new era of networked science.* Princeton University, 2012.

[129] NIH. NIH Clinical Trials Research and You, 2015.

[130] OF HEALTH, N. I. Colloidal Silver — NCCIH, 2018.

[131] PANDEY, V., AMIR, A., DEBELIUS, J., HYDE, E. R., KNIGHT, R., AND KLEMMER, S. Gut Instinct: Creating Scientific Theories with Online Learners. In *2017 CHI Conference on Human Factors in Computing Systems (pp. 6825-6836). ACM.* (2017).

[132] PANDEY, V., DEBELIUS, J., HYDE, E. R., KOSCIOLEK, T., KNIGHT, R., AND KLEMMER, S. 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* (2018), ACM, p. 9.

[133] PANDEY, V., KOTTURI, Y., KULKARNI, C., BERNSTEIN, M. S., AND KLEMMER, S. Connecting Stories and Pedagogy Increases Participant Engagement in Discussions. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (2015), ACM, pp. 253–256.

[134] PANDEY, V., KOUL, T., YANG, C., MCDONALD, D., KNIGHT, R., AND KLEMMER, S. Galileo: Scaling Citizen-led Experimentation with a Procedural Training Platform. *In Preparation* (2019).

[135] PH.D., M. J. B. Alcohol and Sleep: What You Need to Know.

[136] REINECKE, K., ARBOR, A., AND GAJOS, K. Z. LabintheWild : Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (2015).

[137] REINECKE, K., AND GAJOS, K. Z. Quantifying visual preferences around the world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), ACM, pp. 11–20.

[138] RESNICK, P., AND KRAUT, R. *Building Successful Online Communities: Evidence-based social design*. MIT Press, Cambridge, MA, 2011.

[139] RETELNY, D., ROBASZKIEWICZ, S., TO, A., LASECKI, W. S., PATEL, J., RAHMATI, N., DOSHI, T., VALENTINE, M., AND BERNSTEIN, M. S. Expert Crowdsourcing with Flash Teams. *Proceedings of the 27th annual ACM symposium on User interface software and technology - UIST '14* (2014), 75–85.

[140] RITTLE-JOHNSON, B., AND ALIBALI, M. W. Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of educational psychology 91*, 1 (1999), 175.

[141] RITTLE-JOHNSON, B., AND ALIBALI, M. W. Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology 91*, 1 (1999), 175–189.

[142] ROCCO, E. Trust breaks down in electronic contexts but can be repaired by some initial face-to-face contact. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1998), ACM Press/Addison-Wesley Publishing Co., pp. 496–502.

[143] ROTMAN, D., PREECE, J., HAMMOCK, J., PROCITA, K., HANSEN, D., PARR, C., LEWIS, D., AND JACOBS, D. Dynamic Changes in Motivation in Collaborative Citizen-Science Projects. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (2012).

[144] SAVERY, J. R., AND DUFFY, T. M. Problem based learning: An instructional model and its constructivist framework. *Educational Technology 35*, 5 (1995), 31–38.

[145] SCHÖN, D. A. *The reflective practitioner: How professionals think in action*, vol. 5126. Basic books, 1984.

[146] SHAH, D. By The Numbers: MOOCS in 2015., 2015.

[147] SIMON, M. A., TZUR, R., HEINZ, K., AND KINZEL, M. Explicating a mechanism for conceptual learning: Elaborating the construct of reflective abstraction. *Journal for research in mathematics education* (2004), 305–329.

[148] SINGH, D. P., LISLE, L., MURALI, T. M., AND LUTHER, K. CrowdLayout. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (New York, New York, USA, 2018), CHI '18, ACM Press, pp. 1–14.

[149] SNOW, R., CONNOR, B. O., JURAFSKY, D., NG, A. Y., LABS, D., AND ST, C. Cheap and fast - but is it good? Evaluation non-expert annotiations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)* (2008), pp. 254–263.

[150] SORCAR, P. *Teaching taboo topics without talking about them: an epistemic study of a new approach to HIV/AIDS prevention education in India*. Stanford University, 2009.

[151] STAHL, G., KOSCHMANN, T., AND SUTHERS, D. Computer-supported collaborative learning: An historical perspective. *Cambridge handbook of the learning sciences 2006* (2006), 409–426.

[152] STARBIRD, K., MADDOCK, J., ORAND, M., ACHTERMAN, P., AND MASON, R. M. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 Proceedings* (2014).

[153] SUROWIECKI, J. *The Wisdom of Crowds*. Anchor, 2005.

[154] TEEVAN, J., AND YU, L. Bringing the Wisdom of the Crowd to an Individual by Having the Individual Assume Different Roles. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition* (New York, NY, USA, 2017), ACM, pp. 131–135.

[155] THOMAS, M. J. W. Learning within incoherent structures: The space of online discussion forums. *Journal of Computer Assisted Learning 18*, 3 (2002).

[156] TUFEKCI, Z., AND WILSON, C. Social media and the decision to participate in political protest: Observations from tahrir square. *Journal of communication 62*, 2 (2012), 363–379.

[157] V. S. RAMACHANDRAN, S. BLAKESLEE, AND SHAH, N. *Phantoms in the brain: Probing the mysteries of the human mind*. New York: William Morrow., 1998.

[158] VALENTINE, M. A., RETELNY, D., TO, A., RAHMATI, N., DOSHI, T., AND BERNSTEIN, M. S. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (2017), ACM, pp. 3523–3537.

[159] VAN DER MEIJ, H., AND CARROLL, J. M. Principles and Heuristics for Designing Minimalist Instruction. *Technical Communication 42*, 2 (1995), 243–261.

[160] VON AHN, L., LIU, R., AND BLUM, M. Peekaboom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)* (2006), pp. 55–64.

[161] VON AHN, L., MAURER, B., MCMILLEN, C., ABRAHAM, D., AND BLUM, M. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science 321*, 12 September 2008 (2008), 1465–1468.

[162] VON HIPPEL, E. *Democratizing innovation: The evolving phenomenon of user innovation*. MIT, 2005.

[163] VON HIPPEL, E. Democratizing innovation: The evolving phenomenon of user innovation. *Journal fur Betriebswirtschaft 55*, 1 (2005), 63–78.

[164] WANG, N.-C., HICKS, D., AND LUTHER, K. Exploring Trade-Offs Between Learning and Productivity in Crowdsourced History. *Proceedings of the ACM on Human-Computer Interaction 2*, CSCW (2018), 178.

[165] WEISSE, A. B. Self-experimentation and its role in medical research. *From the Texas Heart Institute of St. Luke's Episcopal Hospital, Texas Children's Hospital 39*, 1 (2012).

[166] WICKS, P., VAUGHAN, T. E., MASSAGLI, M. P., AND HEYWOOD, J. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology 29*, 5 (2011), 411–414.

[167] WIKIPEDIA. Community Insights/2018 Report.

[168] WIKIPEDIA. Bristol stool scale, 2018.

[169] WILLETT, W., HEER, J., HELLERSTEIN, J., AND AGRAWALA, M. CommentSpace: structured support for collaborative visual analysis. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (2011), ACM, pp. 3131–3140.

[170] WING, J. M. Computational thinking. *Communications of the ACM 49*, 3 (2006), 33–35.

[171] WOBBROCK, J. Designing, Running, and Analyzing Experiments, 2018.

[172] WU, L., WANG, D., AND EVANS, J. A. Large teams develop and small teams disrupt science and technology. *Nature 566*, 7744 (2019), 378.

[173] YANG, J., HAUFF, C., BOZZON, A., AND HOUBEN, G.-J. Asking the Right Question in Collaborative Q&A Systems. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media* (New York, NY, USA, 2014), HT '14, ACM, pp. 179–189.

[174] YATSUNENKO, T., REY, F. E., MANARY, M. J., TREHAN, I., DOMINGUEZ-BELLO, M. G., CONTRERAS, M., MAGRIS, M., HIDALGO, G., BALDASSANO, R. N., ANOKHIN, A. P., HEATH, A. C., WARNER, B., REEDER, J., KUCZYNSKI, J., CAPORASO, J. G., LOZUPONE, C. A., LAUBER, C., CLEMENTE, J. C., KNIGHTS, D., KNIGHT, R., AND GORDON, J. I. Human gut microbiome viewed across age and geography. *Nature 486*, 7402 (2012), 222–227.

[175] YU, Y., STAMBERGER, J., MANOHARAN, A., AND PAEPCKE, A. EcoPod: a mobile tool for community based biodiversity collection building. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06)* (2006), pp. 244–253.

[176] ZHANG, X., ZHANG, D., JIA, H., FENG, Q., WANG, D., LIANG, D., WU, X., LI, J., TANG, L., LI, Y., LAN, Z., CHEN, B., LI, Y., ZHONG, H., XIE, H., JIE, Z., CHEN, W., TANG, S., XU, X., WANG, X., CAI, X., LIU, S., XIA, Y., LI, J., QIAO, X., AL-AAMA, J. Y., CHEN, H., WANG, L., WU, Q.-J., ZHANG, F., ZHENG, W., LI,

Y., ZHANG, M., LUO, G., XUE, W., XIAO, L., LI, J., CHEN, W., XU, X., YIN, Y., YANG, H., WANG, J., KRISTIANSEN, K., LIU, L., LI, T., HUANG, Q., LI, Y., AND WANG, J. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med 21*, 8 (aug 2015), 895–905.

[177] ZOONIVERSE. Galaxy Zoo, 2007.