Gut Instinct: Creating Scientific Theories with Online Learners

Vineet Pandey[1], Amnon Amir[2], Justine Debelius[2], Embriette R. Hyde[2], Tomasz Kosciolek[2], Rob Knight[2], Scott Klemmer[1]
[1]Design Lab [2]Department of Pediatrics
UC San Diego, La Jolla, CA
{vipandey, amamir, jdebelius, ehyde, tkosciolek, robknight, srk}@ucsd.edu

ABSTRACT
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 paper 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. These results highlight the promise and difficulty of double-bottom-line learning experiences.

Author Keywords
Online learning; citizen science; social computing systems; crowdsourcing

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

THE PROMISE OF CITIZEN SCIENCE WITH LEARNERS

People worldwide have theories about their health, environment, interpersonal interactions, and myriad other topics [26]. Some of these folk theories encapsulate generalizable insights and wisdom; many others are completely false; and some are in between [34]. How might we harvest and assess such intuitive theories to extend human knowledge, especially in domains where science is limited?

CHI 2017, May 6–11, 2017, Denver, CO, USA.

DOI: http://dx.doi.org/10.1145/3025453.3025769

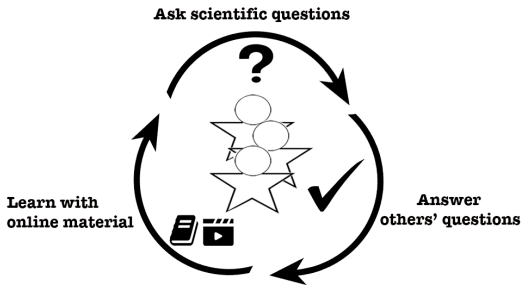

Figure 1: A dual objective: integrating citizen science and online learning

Worldwide, students collectively spend millions of hours a week testing their skills on assignments with known answers [51]. 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 [17,18]. 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 [58]. 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 paper explores the potential of coupling online citizen science with learning materials to create scientific questions (Figure 1).

Often, when citizens participate in science, it is as "embedded sensors" that are aggregated by experts. A classic example is Audubon's Christmas bird count, run since 1900 [7]. Online examples include reporting flower blooms in Project Budburst [13]; recording wildlife activity [24]; identifying galaxies from satellite imagery in GalaxyZoo [59]; and biochemistry games: finding protein structures in Foldit [19], synthesizing RNA molecules in EteRNA [44],

and aligning nucleotide sequences in Phylo [33]. 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 [20].

The main contribution of this paper 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 [37]. However, the combined condition did not show additive benefits.

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 [53]. 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 [52]. 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 [9]. 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

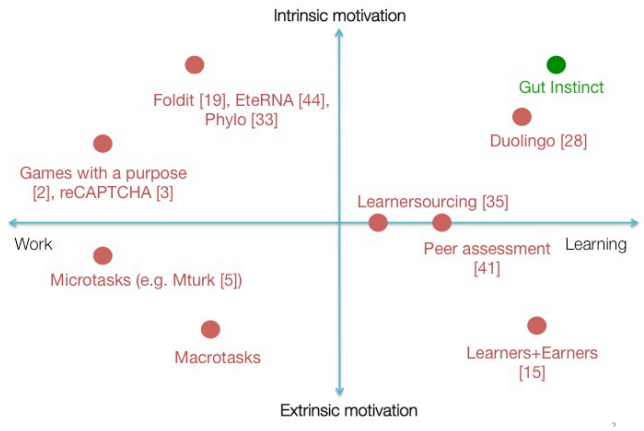

Figure 2: Crowd systems/techniques place different emphasis on work and learning. Some, like Mechanical Turk [5], emphasize work over learning. Crowd approaches also vary in their motivation. Games like Foldit [19] leverage participants' motivation to perform altruistic work while having fun. Gut Instinct helps participants learn about the gut microbiome while contributing towards the altruistic purpose of helping researchers better understand it.

how a collection of novices offers complementary contributions to experts, often in small but nonetheless useful ways.

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". For example, GalaxyZoo volunteers discovered 'green pea' galaxies overlooked by scientists who mistakenly assumed the green hue was merely an imaging artifact [54]. The converse also holds, and much more often: novices are also "uninfected" by all the knowledge that enables experts to innovate. 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 [31]. 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 [10]. Open & crowd innovation builds up on contributions by diverse online participants, and a 'bubbling up' process for strong ideas [56]. Our novel contribution is an explicit integration of learning.

Crowd workers perform better when they understand their efforts' importance. For example, Mechanical Turk workers analyzing radiology images performed better when told of the medical purpose: finding cancerous tumors [14]. Motivation can also be personal. For example, 23andMe is a genetic testing site and online service that includes a discussion board. On this forum, a user reported disliking the sounds of others eating. She's not alone; a 23andMe survey found 16,000 users with the same condition and a predictive genetic similarity among them [1].

Creative, open-ended work has rich pedagogical value. Online work, like online learning, requires appropriate scaffoldings, such as rubrics [12,41], decision trees [43,57], tutorials [6], and quick expert guidance [23]. Similar to general critique of pure discovery learning [47], simply asking participants to "figure it out" would be poor pedagogy. Hence, Gut Instinct introduces a guided discovery learning approach as Mayer advocates: expert-curated learning materials help participants start, with discovery following. Recruiting learners as citizen scientists offers a Problem-based Learning experience with context and motivation for the material students learn [50]**.** In principle, these real-world problems also provide a yardstick for measuring learning.

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 learned about the topic they edited [25]. Similarly, Kim *et al.* asked learners to create how-to video segments as part of an online curriculum [35]; 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 [2]), authentication (reCAPTCHA [3]), and learning (Duolingo [28]). Under the hood, these tasks simultaneously label images, transcribe text, and translate phrases. Such crowd work can also bootstrap machine learning [8]. This paper is distinct from prior work (Figure 2) in leveraging people's individual lived experience, knowledge, context, and folk theories, rather than treating people as interchangeable respondents.

Understanding the human microbiome requires insights into people's lifestyles

The human gut microbiome is the community of microbes (and their gene products) interacting in the human gut. However, research has only scratched the surface of understanding the microbiome and using it to improve our well-being. The American Gut Project (AGP) is the world's largest crowdfunded citizen science project [38]. AGP participants contribute their samples for bacterial marker gene sequencing and analysis [22]. Participants then receive a summary of their results with all their raw data. Anonymized data is publically available. AGP seeks to build a comprehensive map of the human microbiome, and identify its healthy and unhealthy components.

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 [27]. 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 [16]. The human microbiome is impossible to understand without information about its host [22] 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.

HYPOTHESES

This paper 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.

While learning almost by definition improves performance on similar tasks, transfer to novel tasks (like creating new & different questions) is famously uneven [10]—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.

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. 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). It is BSD open source at http://gutinstinct.ucsd.edu.

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 [36]. 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

Figure 3: A question on Topics page for diet to test understanding of the learning material

Figure 4: 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

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). Overall, these elements of the interface form the learning part of the system.

GutBoard: Discussing and answering questions

The GutBoard provides a discussion board with user-generated questions tagged by topics (Figure 5(a)). 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.

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 5(b) shows the interface.

Nudges to think creatively and to stay on task

Gut Instinct employs several best practices for increasing high-quality contributions [32,49]. 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

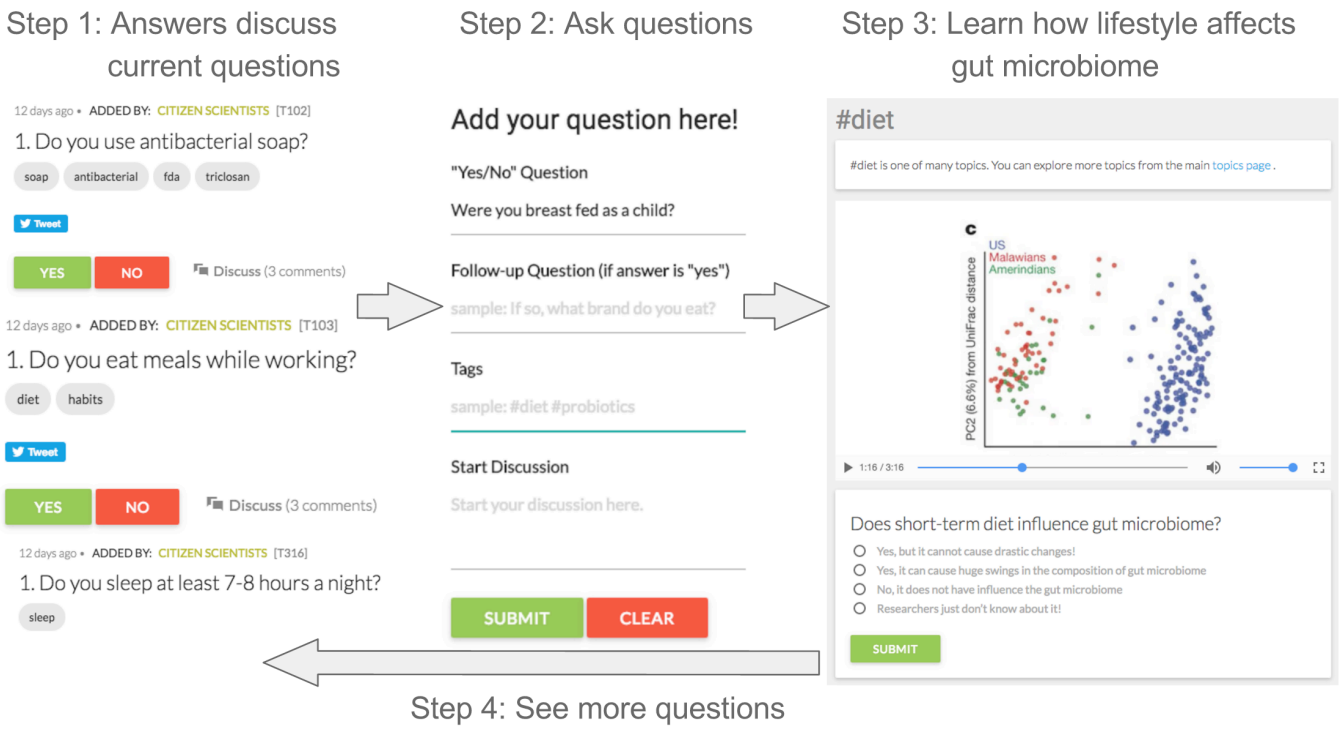

Figure 5: Gut Instinct is a web system to learn about the gut microbiome and create causal theories about gut microbiome (a) A discussion board where learners add their questions and discuss them with other learners (b) "Add question" box for people to add their own questions, (c) A tutorial video showing how gut microbiome varies across countries with different food habits [55]

to gut microbiome researchers (Figure 4). 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.

EXPERIMENT: WORK, LEARNING, & COMBINED

A between-subjects experiment compared the work and learning performance of participants across three different conditions: *Contribute, Learn and Combined* (Figure 6). In the Learn condition, participants were provided learning material and some practice problems, both curated from the Coursera microbiome class [36]). 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.

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

Table 1: Demography info for 44 participants. Some participants did not complete portions of survey

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.

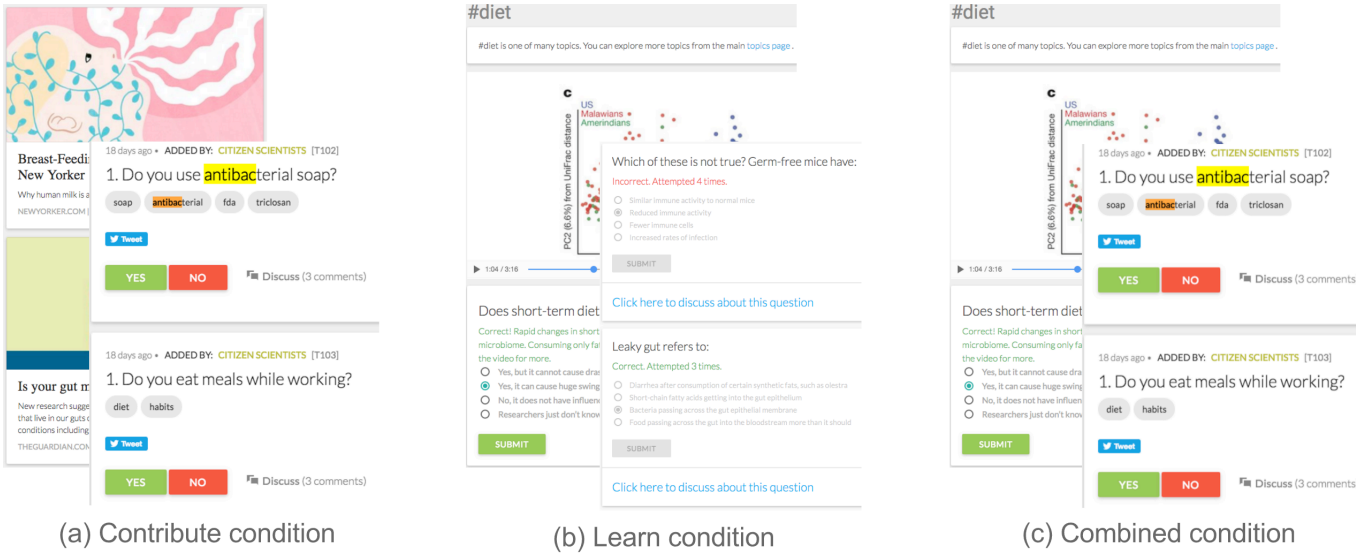

Figure 6: Three conditions for experiment. (a) Contribute condition where participants read some general articles about microbiome and added questions and answered others' questions (b) Learn condition where participants saw curated topic videos (e.g. about diet) and answered practice problems from a Coursera class [36] (c) Combined condition where participants saw curated

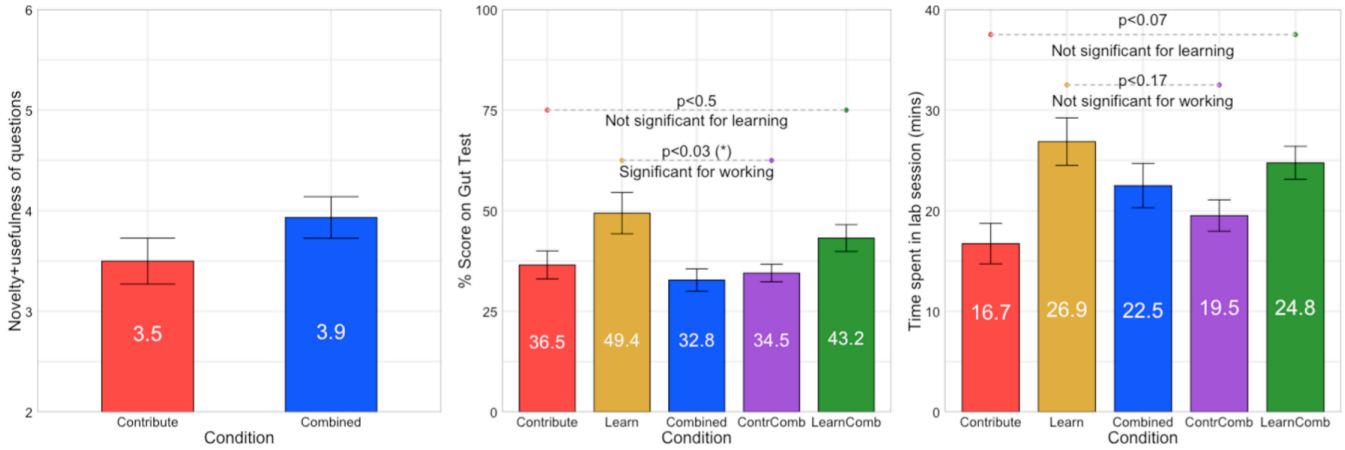

Figure 7: a) Participants in Contribute and Combined conditions created questions of similar quality

b) Participants in Learn condition performed the best on a summative test. Learning did not show a significant effect on score but working did

c) There were no significant differences in time spent in lab session across the conditions

Participants

44 participants were recruited from a Southern California university (Table 1). Participants were novices in terms of their knowledge of the human microbiome. Random assignment balanced g ender 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 info, 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).

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 7a). Figure 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 9).

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

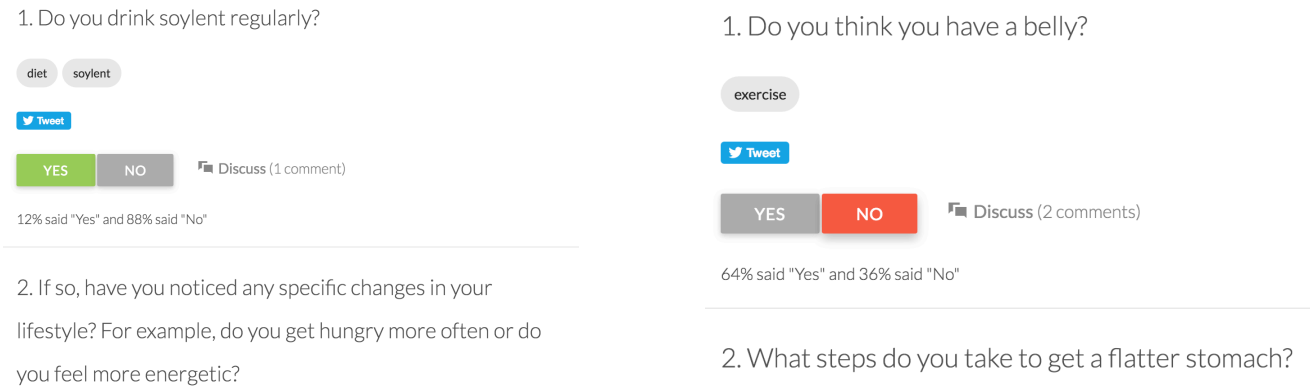

Figure 8: An example of a good and bad question added by participants. Soylent question was scored 5/6 (2 on novelty and 3 on usefulness) while the belly question was rated 2/6 (1 on novelty and 1 on usefulness)

Figure 9: Most participants reported that the learning experience was helpful and the system was enjoyable. 65% of participants asked questions; the mode was 1

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 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 7b). The effect size for working was small (Cohen's effect size $d=.11$).

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 7c). 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 2 summarizes results from the experiment.

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, 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 [43] 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 [45] 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 [37]. 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.

Measures (mean values)	Combined	Contribute	Learn	*p*
No difference in quality or quantity of questions across Combined or Contribute conditions				
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 points)	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	-

Table 2: Summary of results from experiment

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 [39] 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 [31]. 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.

SCIENCE WITH LEARNERS: PROMISE & CHALLENGES

This paper 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 [41]. Furthermore, the newly-aware sometimes articulate useful insights that familiarity has blinded experts to [30]. 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 paper'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 [22]. 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 [40] 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 [50] and just in time learning [11] to the scale of Internet. For example, browsing StackOverflow before fixing programming questions leads to better work, and lateral learning [46]. Similarly, global-scale distributed contributions like peer review have enabled massive online courses to offer creative, open-ended assignments through peer review [41]. Such active learning approaches seek a dual objective of content learning and metacognitive growth [21].

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 [4,42]. 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 [29]. 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 [48]. 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.

CONCLUSION

This paper 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 paper 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.

ACKNOWLEDGMENTS

We thank all participants who used Gut Instinct and provided feedback. We thank members of Design Lab, especially Steven Dow and Derek Lomas, and Michael Bernstein for their useful comments on this work. We thank Brian Soe and Aliff Macapinlac for help developing the Gut Instinct website and running pilot studies. A Google Research Award and gift from SAP helped support this work.

REFERENCES

1. 23andMe. 2015. Something to Chew On. Retrieved December 31, 2016 from https://blog.23andme.com/23andmeresearch/something-to-chew-on

2. Luis von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*, 55–64. https://doi.org/10.1145/1124772.1124782

3. Luis Von Ahn, Benjamin Maurer, Colin Mcmillen, David Abraham, and Manuel Blum. 2008. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321, 12 September 2008: 1465–1468. https://doi.org/10.1126/science.1160379

4. Vincent Aleven, Bruce Mclaren, Ido Roll, and Kenneth Koedinger. 2006. Toward meta-cognitive tutoring: A model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education* 16, 2: 101–128. https://doi.org/10.1.1.121.9138

5. Amazon. 2016. Mechanical Turk. Retrieved December 31, 2016 from https://www.mturk.com

6. Erik Andersen, Eleanor O Rourke, Yun-en Liu, Richard Snider, Jeff Lowdermilk, David Truong, Seth Cooper, and Zoran Popovi. 2012. The Impact of Tutorials on Games of Varying Complexity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*, 59–68. https://doi.org/10.1145/2207676.2207687

7. Audubon. 2016. Audubon Science. Using data to realize the best conservation outcomes. Retrieved December

31, 2016 from http://www.audubon.org/conservation/science/christmas-bird-count

8. Michael S. Bernstein. 2012. Crowd-powered Systems. Retrieved December 31, 2016 from http://hci.stanford.edu/msb/files/job_search/research-statement.pdf
9. Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*, 313–322. https://doi.org/10.1145/1866029.1866078
10. Margaret A. Boden. 2004. *"The Story so far". The Creative Mind: Myths and Mechanisms.* Routledge.
11. Michele K. Bolton. 1999. The Role Of Coaching in Student Teams: A "Just-in-Time" Approach To Learning. *Journal of Management Education* 23: 233–250.
12. David Boud. 1995. *Enhancing learning through self-assessment*. Kogan Page, London.
13. Project BudBurst Boulder Colorado. 2016. Project BudBurst: An online database of plant phenological observations. Retrieved December 31, 2016 from http://budburst.org/
14. Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior and Organization* 90: 123–133. https://doi.org/10.1016/j.jebo.2013.03.003
15. Guanliang Chen, Dan Davis, Markus Krause, Efthimia Aivaloglou, Claudia Hauff, and Geert-Jan Houben. 2016. Can Learners be Earners? Investigating a Design to Enable MOOC Learners to Apply their Skills and Earn Money in an Online Market Place. *IEEE Transactions on Learning Technologies* PP, 99: 1. https://doi.org/10.1109/TLT.2016.2614302
16. I. Cho and M.J. Blaser. 2012. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics* 13, 4: 260–270. https://doi.org/10.1038/nrg3182
17. The Human Microbiome Project Consortium. 2012. A framework for human microbiome research. *Nature* 486, 7402: 215–221. https://doi.org/10.1038/nature11209.A
18. The Human Microbiome Project Consortium. 2013. Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* 486, 7402: 207–214. https://doi.org/10.1038/nature11234.Structure
19. Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307: 756–760. https://doi.org/10.1038/nature09304
20. Michael J. Coren and Fast Company. 2011. Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks. Retrieved December 31, 2016 from https://www.scientificamerican.com/article/foldit-gamers-solve-riddle/
21. Catherine H. Crouch and Eric Mazur. 2001. Peer Instruction: Ten years of experience and results. *American Journal of Physics* 69, 9: 970. https://doi.org/10.1119/1.1374249
22. Justine W Debelius, Yoshiki Vázquez-Baeza, Daniel McDonald, Zhenjiang Xu, Elaine Wolfe, and Rob Knight. 2016. Turning Participatory Microbiome Research into Usable Data: Lessons from the American Gut Project. *Journal of Microbiology & Biology Education* 17, 1: 46–50. https://doi.org/10.1128/jmbe.v17i1.1034
23. Steven P. Dow, Anand Kulkarni, Scott R. Klemmer, and Bjorn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*, 1013–1022. https://doi.org/10.1145/2145204.2145355
24. Siamak Faridani, Bryce Lee, Selma Glasscock, John Rappole, Dezhen Song, and Ken Goldberg. 2009. A networked telerobotic observatory for collaborative remote observation of avian activity and range change. *IFAC Proceedings Volumes (IFAC-PapersOnline)*: 56–61. https://doi.org/10.3182/20091006-3-US-4006.0015
25. Rosta Farzan and Robert E Kraut. 2013. 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)*, 783–792. https://doi.org/10.1145/2470654.2470765
26. Susan A. Gelman and Cristine H Legare. 2011. Concepts and folk theories. *Annu Rev Anthropol*: 379–398. https://doi.org/10.1146/annurev-anthro-081309-145822
27. SR Gill, Mihai Pop, RT DeBoy, and PB Eckburg. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* 312, 5778: 1355–1359. https://doi.org/10.1126/science.1124234.Metagenomic
28. Severin Benedict Hans Hacker. 2014. Duolingo: Learning a Language while Translating the Web. Ph.D Dissertation.
29. Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Most People are not WEIRD. *Nature* 466, July 2010. https://doi.org/10.1017/S0140525X0999152X

30. Pamela J. Hinds. 1999. The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology: Applied* 5, 2: 205–221. https://doi.org/10.1037/1076-898X.5.2.205

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

32. Charlene Jennett and Anna L. Cox. 2014. Eight Guidelines for Designing Virtual Citizen Science Projects. *Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP '14)*: 16–17.

33. Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Luis Sarmenta, Mathieu Blanchette, and Jérôme Waldispühl. 2012. Phylo: A citizen science approach for improving multiple sequence alignment. *PLoS ONE* 7, 3. https://doi.org/10.1371/journal.pone.0031362

34. Willett Kempton. 1986. Two theories of home heat control. *Cognitive Science* 10, 1: 75–90. https://doi.org/10.1016/S0364-0213(86)80009-X

35. Juho Kim. 2015. Learnersourcing : Improving video learning with collective learner activity. Ph.D Dissertation. Retrieved from https://juhokim.com/files/JuhoKim-Thesis.pdf

36. R. Knight, J. Metcalf, and K. Amato. 2016. Gut Check: Exploring Your Microbiome. Coursera. Retrieved December 31, 2016 from https://www.coursera.org/learn/microbiome

37. KnightLab. 2016. American Gut Project. Login. Retrieved December 31, 2016 from http://microbio.me/americangut/

38. KnightLab. 2016. American Gut - What's in your gut? Retrieved December 31, 2016 from http://americangut.org/

39. Yasmine Kotturi, Chinmay E. Kulkarni, Michael S Bernstein, and Scott Klemmer. 2015. Structure and messaging techniques for online peer learning systems that increase stickiness. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale (L@S '15)*, 31–38. https://doi.org/10.1145/2724660.2724676

40. Michel Krieger, Emily Margarete Stark, and Scott R Klemmer. 2009. Coordinating tasks on the commons: designing for personal goals, expertise and serendipity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*, 1485–1494. https://doi.org/10.1145/1518701.1518927

41. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. 2013. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction* 20, 6: 1–31. https://doi.org/10.1145/2505057

42. Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael Terry, and Krzysztof Z Gajos. 2016. Curiosity Killed the Cat, but Makes Crowdwork Better. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '16)*. https://doi.org/10.1145/2858036.2858144

43. Doris Lee, Joanne Lo, Moonhyok Kim, and Eric Paulos. 2016. Crowdclass: Designing classification-based citizen science learning modules. In *Proceedings of the Fourth AAAI Conference on Human Computation and Crowdsourcing (HCOMP '16)*.

44. Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille, and Rhiju Das. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* 111, 6: 2122–2127. https://doi.org/10.1073/pnas.1313039111

45. Yi-Chieh Lee, Wen-Chieh Lin, Fu-Yin Cherng, Hao-Chuan Wang, Ching-Ying Sung, and Jung-Tai King. 2015. Using Time-Anchored Peer Comments to Enhance Social Interaction in Online Educational Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*, 689–698. https://doi.org/10.1145/2702123.2702349

46. Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. 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)*, 2857–2866. https://doi.org/10.1145/1978942.1979366

47. Richard E Mayer. 2004. Should There Be a Three-Strikes Rule Against Pure Discovery Learning? The case for guided methods of instruction. *American Psychologist* 59, 1: 14–19. https://doi.org/10.1037/0003-066X.59.1.14

48. Katharina Reinecke, Ann Arbor, and Krzysztof Z Gajos. 2015. LabintheWild : Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*, 1364–1378. https://doi.org/10.1145/2675133.2675246

49. R. Resnick, P. and Kraut. 2011. *Building Successful Online Communities: Evidence-based social design*. MIT Press, Cambridge, MA.

50. John R Savery and Thomas M Duffy. 1995. Problem based learning: An instructional model and its constructivist framework. *Educational Technology* 35, 5: 31–38. https://doi.org/47405-1006

51. Dhawal Shah. 2015. By The Numbers: MOOCS in 2015. *Class Central.* Retrieved from https://www.class-central.com/report/moocs-2015-stats/

52. Rion Snow, Brendan O Connor, Daniel Jurafsky, Andrew Y Ng, Dolores Labs, and Capp St. 2008. 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)*, 254–263. https://doi.org/10.1.1.142.8286

53. James Surowiecki. 2005. *The Wisdom of Crowds*. Anchor.

54. Ramine Tinati, Max Van Kleek, Elena Simperl, Markus Luczak-Roesch, Robert Simpson, and Nigel Shadbolt. 2015. Designing for Citizen Data Analysis: A Cross-Sectional Case Study of a Multi-Domain Citizen Science Platform. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*, April: 4069–4078. https://doi.org/10.1145/2702123.2702420

55. T Yatsunenko, F E Rey, M J Manary, I Trehan, M G Dominguez-Bello, M Contreras, M Magris, G Hidalgo, R N Baldassano, A P Anokhin, A C Heath, B Warner, J Reeder, J Kuczynski, J G Caporaso, C A Lozupone, C Lauber, J C Clemente, D Knights, R Knight, and J I Gordon. 2012. Human gut microbiome viewed across age and geography. *Nature* 486, 7402: 222–227. https://doi.org/10.1038/nature11053

56. Lixiu Yu, Jeffrey V Nickerson, and Yasuaki Sakamoto. 2012. Collective creativity: Where we are and where we might go. *Collective Intelligence Conference*: 1–8.

57. YuanYuan Yu, J.A. Stamberger, A. Manoharan, and A. Paepcke. 2006. 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)*, 244–253. https://doi.org/10.1145/1141753.1141807

58. Xuan Zhang, Dongya Zhang, and Huijue et al. Jia. 2015. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* 21, 8: 895–905. https://doi.org/10.1038/nm.3914

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