Crowdsourcing Research Questions in Science

Beck, Susanne; Brasseur, Tiare; Poetz, Marion; Sauermann, Henry

Document Version
Final published version

Published in:
Research Policy

DOI:
10.1016/j.respol.2022.104491

Publication date:
2022

Research Policy 51 (2022) 104491

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Crowdsourcing research questions in science

Susanne Beck [a,b,*], Tiare-Maria Brasseur [a,b], Marion Poetz [a,b], Henry Sauermann [c]

[a] Ludwig Boltzmann Gesellschaft, Open Innovation in Science Center (LBG OIS Center), Nußdorfer Str. 64, Vienna 1090, Austria
[b] Department of Strategy and Innovation, Copenhagen Business School, Kilevej 14A, Frederiksberg 2000, Denmark
[c] European School of Management and Technology, Schlossplatz 1, 10178 Berlin, Germany

ARTICLE INFO

Keywords:
Crowd science
Citizen science
Crowdsourcing
Problem solving
Problem finding
Agenda setting
Organization of science

ABSTRACT

Scientists are increasingly crossing the boundaries of the professional system by involving the general public (the crowd) directly in their research. However, this crowd involvement tends to be confined to empirical work and it is not clear whether and how crowds can also be involved in conceptual stages such as formulating the questions that research is trying to address. Drawing on five different "paradigms" of crowdsourcing and related mechanisms, we first discuss potential merits of involving crowds in the formulation of research questions (RQs). We then analyze data from two crowdsourcing projects in the medical sciences to describe key features of RQs generated by crowd members and compare the quality of crowd contributions to that of RQs generated in the conventional scientific process. We find that the majority of crowd contributions are problem restatements that can be useful to assess problem importance but provide little guidance regarding potential causes or solutions. At the same time, crowd-generated research questions frequently cross disciplinary boundaries by combining elements from different fields within and especially outside medicine. Using evaluations by professional scientists, we find that the average crowd contribution has lower novelty and potential scientific impact than professional research questions, but comparable practical impact. Crowd contributions outperform professional RQs once we apply selection mechanisms at the level of individual contributors or across contributors. Our findings advance research on crowd and citizen science, crowdsourcing and distributed knowledge production, as well as the organization of science. We also inform ongoing policy debates around the involvement of citizens in research in general, and agenda setting in particular.

1. Introduction

> *"If I had only one hour to save the world, I would spend fifty-five minutes defining the problem, and five minutes finding the solution."*
>
> (Attributed to Einstein)

Scientific research is instrumental in improving productivity, health, and social welfare in modern societies. As such, scholars have directed great attention towards understanding the institution of science and the organization of knowledge production (Dasgupta and David, 1994; Ding et al., 2010; Merton, 1973; Sauermann and Stephan, 2013). However, scientific research faces important challenges such as a steady increase in the resources required to reach the knowledge frontier and declining productivity (Jones, 2009; Pammolli et al., 2011) as well as rising skepticism towards science in the broader public (European Science Foundation, 2013; Lewandowsky et al., 2016). Partly in response to these challenges, science is undergoing a fundamental change: While research has for a long time been the domain of highly trained experts in academic or industrial sectors, an increasing number of projects now directly involve members of the broader public in knowledge production (Bonney et al., 2014; Hand, 2010; Wiggins and Crowston, 2011). Such "crowd science" or "citizen science" (CS) projects are active in a broad range of fields such as biology, medicine, ecology, physics, and even the social sciences (Franzoni et al., 2022, forthcoming; Scistarter, 2020). Policymakers and funding agencies strongly support this development in the hopes of accelerating scientific knowledge production and increasing its societal impact (European Commission, 2018; US Congress, 2016).

The rapidly growing number of crowd and citizen science projects has already resulted in a significant volume of research outputs (Irwin, 2018; Kullenberg and Kasperowski, 2016). Scholars of science have documented the diffusion of CS across disciplines, quantified the financial value of crowd contributions, and studied contributors'

* Corresponding author at: Ludwig Boltzmann Gesellschaft, Open Innovation in Science Center (LBG OIS Center), Nußdorfer Str. 64, Vienna 1090, Austria.
E-mail address: sub.si@cbs.dk (S. Beck).

https://doi.org/10.1016/j.respol.2022.104491
Received 24 May 2021; Received in revised form 24 December 2021; Accepted 24 January 2022
Available online 3 February 2022

motives (e.g., Kullenberg and Kasperowski, 2016; Lyons and Zhang, 2019; Sauermann and Franzoni, 2015). This descriptive research reveals an interesting pattern: Many projects involve crowd members in empirical stages of the research process, yet there are very few projects that involve the crowd in conceptual stages, especially in identifying problems and formulating questions that research should address (Hecker et al., 2018; Turrini et al., 2018). This observation is interesting because crowdsourcing is very effective in the related context of innovation, where it can leverage contributors' expertise to identify and solve important problems (Jeppesen and Frederiksen, 2006; Lüthje et al., 2005; Poetz and Schreier, 2012). As such, one might expect similar benefits from involving crowd members in the formulation of research questions (RQs). At the same time, the results from innovation studies may not generalize to the context of science because science is typically more removed from the general public and non-professionals may lack the scientific expertise that is often assumed necessary to formulate research questions (Merton, 1973). The crowd's ability to formulate scientific RQs should, therefore, be of interest to scholars interested in crowd science, crowdsourcing, as well as the organization of science and innovation more generally.

Insights into the crowd's ability to formulate research questions are also important for policy and scientific practice: There are increasing calls from policymakers and advocacy groups to involve citizens in defining research questions and setting research agendas (Caron--Flinterman et al., 2005; Mazzucato, 2018). These calls are not based on the notion that scientists currently lack questions to study. Rather, the hope is that involving citizens can help bridge the gap between science and society, leveraging unique perspectives from crowd members that result in different kinds of research questions (Pols, 2014; Sauermann et al., 2020). Systematic evidence on crowd-generated RQs can provide a firmer foundation to assess the merits of such practices.

Following prior work in management and economics, we define the crowd as a large group of people who self-select to carry out a task in response to an open call, and who are located outside of the boundaries of the focal organization (Afuah and Tucci, 2012; Jeppesen and Lakhani, 2010). The prior literature has studied different kinds of crowds, including crowds composed of professionals with domain-relevant expertise, individuals with deep expertise in other domains, members of the general public with user knowledge, as well as "normal" people without particular expertise related to the task at hand (e.g., Guinan et al., 2013; Piezunka and Dahlander, 2015; Tucci et al., 2018). Given the dominant profiles of contributors to CS projects (Raddick et al., 2013; Science Europe, 2018), our interest is in crowd contributors who are not professional scientific researchers themselves but may have expertise and experience in the problem domain. We note that the literature we build on conceptualizes crowds as potential contributors of valuable inputs and is distinct from literatures that use the term "crowd" with negative connotations.[1]

In the conceptual part of the paper, we discuss the nature of research questions and distinguish three dimensions of RQ quality that are commonly considered in contemporary scientific practice: novelty, scientific impact, and practical impact (Cummings et al., 2007; Thabane et al., 2009). Drawing on different conceptualizations of crowdsourcing (what we call crowdsourcing "paradigms"), we also discuss the potential merits of involving the crowd in RQ formulation. We then examine central features of crowd-generated RQs empirically by analyzing data from two crowdsourcing efforts initiated by a large research institution that invited members of the public to contribute research questions in the domains of mental illness (project 1) and traumatology (project 2). The crowd in this case consisted of a range of individuals that had relevant experience and knowledge but were not themselves engaged in scientific research, including medical practitioners such as nurses and therapists, patients, patient relatives, and other people who had personal experience with mental illness or trauma.

Our descriptive analysis shows that many research questions formulated by crowd members tend to be relatively simple problem restatements (e.g., "What can be done to speed up the body's healing process?"). These research questions do not refer to potential causes of problems or solutions and are thus "ill-structured" (Schwenk and Thomas, 1983; Von Hippel and Von Krogh, 2016). At the same time, a large share of questions is discipline-crossing, i.e., they connect concepts from different fields within medicine or even non-medical fields (e.g., "To what extent do accidental injuries make an impact on the victim's social life?").

To systematically compare crowd-generated RQs with questions developed in the conventional process, we sampled professional research questions from conference proceedings in the same fields. We then asked expert researchers to evaluate both sets of questions with respect to three key dimensions of quality while being blind to the source. In project 1, we find that the average crowd question is rated as less novel and with lower scientific impact than professional questions, but of similar practical impact. In project 2, crowd questions are rated lower with respect to all quality dimensions. However, crowd questions outperform professional ones once we focus on the best of multiple submissions by individual contributors (project 1), or on the best questions across all contributors (project 2). These patterns are consistent with prior research showing that the benefits of crowdsourcing hinge on attracting large numbers of contributors, activating and retaining prolific contributors, and finding efficient approaches to screen large numbers of contributions (Piezunka and Dahlander, 2015; Tucci et al., 2018).

Our findings contribute to several streams of literature. First, we contribute to the literature on crowd and citizen science, which has focused on the role of crowd members in empirical activities such as data collection or coding (Bonney et al., 2014; Franzoni and Sauermann, 2014; Lyons and Zhang, 2019; Nielsen, 2011). We highlight crowd involvement in conceptual stages of the research process as an important gap and provide a conceptual discussion that draws on the broader crowdsourcing literature to consider potential benefits of involving crowds in formulating research questions. Data from two real projects yield novel insights into the characteristic features of crowd-generated research questions and into their quality relative to RQs generated in the conventional scientific process.

We note that the literature on crowd and citizen science is very diverse, with different streams emphasizing different project features and goals (Haklay et al., 2021; Sauermann et al., 2020). Our paper bridges between the crowd science and citizen science streams in that we ground our discussion in different paradigms of crowdsourcing, but study a crowd that consists of citizens who are not professional scientists (see Franzoni et al., forthcoming). At the same time, we focus on the quality of research questions as the outcome of interest and we do not examine whether crowd involvement in RQ formulation can also achieve other goals such as scientific literacy, increasing awareness of certain problems, or policy change (Kimura and Kinchy, 2016; Van Brussel and Huyse, 2018). Similarly, we chose expert scientists as the relevant judges of RQ quality, recognizing that some citizen science actors challenge professional authority and performance standards, or argue that the professional system needs to change in order to better appreciate and integrate citizens' contributions (Cohen and Doubleday, 2021; Ottinger, 2010). Indeed, the performance of the crowd in our study may seem even more impressive considering that we hold research

[1] An earlier sociological literature portrays the crowd as "deficient" and "irrational" (Le Bon, 1895; Sturgis and Allum, 2004), often associating crowds with violence and de-individualization. Among others, this work examines negative consequences when individual crowd members are "carried away by the mood of the multitude" (Borch, 2012). Our conceptualization of crowds is much more neutral and builds on a different underlying societal model; indeed, it highlights the potential benefits that both professional scientists and the broader public can gain through crowdsourcing and other forms of collaboration (see also Franzoni et al., forthcoming).

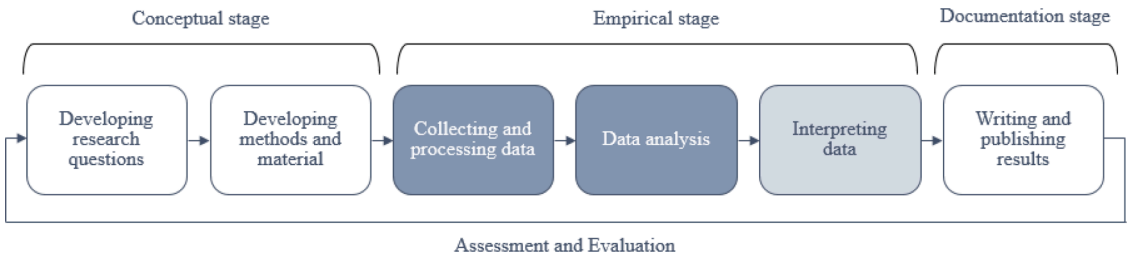

Fig. 1. Crowd involvement in different stages of the scientific research process.
Notes: We use a stylized conceptualization that abstracts from the complex and iterative nature of the scientific research process. Darker shades indicate that crowd involvement currently tends to be more common in those particular stages (see Hecker et al., 2018; Turrini et al., 2018).

questions to professional quality standards and use as benchmark professional RQs that were developed in a more extensive conventional research process that typically includes teamwork and improvements over time. In Section 5, we discuss how future research can complement our perspective.

Second, we contribute to the literature on distributed knowledge production and crowdsourcing. Scholars have made great progress by studying the crowd's ability to solve problems (Franke et al., 2013; Jeppesen and Lakhani, 2010; Majchrzak and Malhotra, 2020) or to identify problem-solution pairs (Von Hippel and Von Krogh, 2016). We add to an emerging stream of research on the crowd's role in identifying problems (e.g., Nickerson et al., 2017) by studying problem formulation in the context of science. This context is interesting because of unique challenges and opportunities arising from the large distance between the knowledge bases of crowd members and professional scientists. Although much of the prior work identifies conditions under which crowdsourcing versus traditional mechanisms are superior (Afuah and Tucci, 2012; Felin and Zenger, 2014), science is a context where contributions from crowd members and professional scientists may be complementary across different stages of knowledge production, suggesting opportunities for future research on the interactions between crowd members and professional scientists throughout this process.

Finally, we contribute to the literature on the organization of science. Much of this work has focused on features that distinguish science from other organizational realms (Dasgupta and David, 1994; Merton, 1973) or has studied changes *within* the professional scientific enterprise (Ding et al., 2010; Wuchty et al., 2007). We study an emerging mechanism that *crosses* conventional boundaries by involving non-professional scientists in research. Our results provide evidence regarding important opportunities as well as challenges arising from involving the crowd in setting research agendas, suggesting considerable value of future research on crowd involvement in science and on boundary spanning between science and the broader public.

Section 2 provides conceptual background, including a discussion of dimensions of research question quality and of potential merits of crowd involvement suggested by different paradigms of crowdsourcing and related mechanisms. We then describe the research context and measurement in Section 3. The empirical analysis in Section 4 provides descriptive results on research questions generated by the crowd and compares crowd questions to questions developed in the conventional process. We discuss our findings and identify opportunities for future research in Section 5.

2. Theoretical background

2.1. Crowd involvement in science

Although a generally accepted definition of crowd or citizen science is yet to emerge, CS can be understood as the direct participation of members of the public in scientific research projects in response to an open call for contributions (Eitzel et al., 2017; Franzoni et al., 2022; Haklay et al., 2021). Projects are typically led by professional scientists and crowd contributors participate as volunteers. Contributors are not professional scientists but often have relevant expertise in particular problem domains and may have experience with research. Although the general public has long been involved in research (Shapin, 2008), technological advances and increased policy attention have contributed to a surge in crowd science activities over the last decades (Beck et al., 2022; Science Europe, 2018). CS projects are now active in a broad range of fields. For example, the platform Zooniverse has enabled over two million volunteers to support research projects in astronomy, biology, and history by classifying images, audio files, and videos. Environmental monitoring projects such as eBird draw on geographically dispersed crowds to collect data on animal and plant populations, providing valuable input for research in biology and climatology. The project Foldit involves crowds in solving protein folding problems and has resulted in a number of top-tier publications.[2]

A common approach to classify CS projects is to conceptualize scientific research as consisting of a number of interdependent stages (see Fig. 1), and to distinguish projects according to which of these stages involve crowd members (Follett and Strezov, 2015).[3] Emerging evidence suggests that most CS projects involve crowds in empirical stages of the research, such as collecting observational data or analyzing images and other kinds of information. In contrast, there are very few projects that involve the crowd in the documentation of results or in conceptual stages, such as in identifying the problem and research questions that a project is trying to address (Bonney et al., 2014; Hecker et al., 2018; Turrini et al., 2018). The focus on empirical contributions is not surprising, given the considerable efficiencies that projects can gain from employing large numbers of contributors to collect data across time and space, or to process data in a distributed fashion (Sauermann and Franzoni, 2015; Theobald et al., 2015). However, there are increasing calls to expand the scope of crowd involvement, reflecting the hope that involving citizens in conceptual stages may allow them to contribute unique knowledge and experience to identify novel problems and to direct research efforts towards areas of societal need (Caron-Flinterman et al., 2005; Guinan et al., 2013; Mazzucato, 2018). Despite these hopes, there is very little conceptual or empirical research on the crowd's ability to contribute scientific research questions. The goal of our paper is to fill this gap. In the subsequent sections, we first conceptualize the particular task of research question formulation and discuss different

[2] www.zooniverse.org; www.ebird.org; www.fold.it

[3] Although we follow the prior literature in conceptualizing research as a linear process, we acknowledge that this strongly simplifies the complex and iterative nature of research in practice. Indeed, rich qualitative studies suggest that research questions can develop over time as the work progresses, that questions are adjusted to make them "do-able" given various multi-level constraints, or that empirical activities may be performed without clear research questions (Fujimura, 1987; Knorr-Cetina, 1999; Latour and Woolgar, 1979). Thus, while we believe that research questions are a useful construct and worthy of study, we do not claim that all research projects require well-defined research questions or that RQs are always defined and fixed at the beginning of the process.

dimensions of RQ quality. We then draw on different "paradigms" of crowdsourcing and related mechanisms to discuss potential benefits of involving crowds. Although this discussion will also point towards potential challenges, a comprehensive consideration of such challenges is left for future work (see Section 5).

2.2. Problems, research questions, and RQ characteristics

Developing a research question is an important step in the research process because it can critically shape later stages such as research design and research methods (Alvesson and Sandberg, 2011; Bryman, 2007). Generally speaking, a research question describes the kind of knowledge a researcher seeks to generate in order to solve a problem. Such problems can be purely curiosity-driven without immediate practical relevance (e.g., we want to know more about the history of the universe). Problems can also relate to important practical concerns (e.g., we want to reduce the rate of cancer), and of course they can be related to both (Stokes, 1997).

The two examples above are what organization theorists would call "ill-structured" problems: they lack a well-defined problem space and it is not clear what operators can be used to move from the starting state to the desired state (Felin and Zenger, 2014; Simon, 1973). The corresponding research questions (e.g., "What is the history of the universe", or "How can we cure cancer?") are similarly ill-structured and provide little guidance as to which particular elements of the problem should be investigated. Therefore, problem formulation typically involves not only finding a particular problem but also exploring potential underlying causes that, once addressed, can help solve the problem (Schwenk and Thomas, 1983; Simon, 1973; Von Hippel and Von Krogh, 2016). Research questions that entail potential causes or solutions narrow the problem space and tend to be easier to answer than ill-structured questions. For example, the question "What is the effect of regular physical exercise on the risk of cancer?" identifies a potential cause of cancer (lack of exercise) that can be systematically investigated. Formulating a *well-structured* research question by incorporating potential causes and solutions is non-trivial and is an important value-added activity: As suggested by the quote at the beginning of the paper, formulating the "right" question can be considered an important step towards solving the problem. Of course, there are many potential causes and solutions that could be included in a research question and not all of them are similarly promising, affecting the perceived "quality" of a research question (see next section).

In addition to question structure, we also consider a related characteristic, *specificity*. The examples above illustrate that well-structured RQs tend to be more specific in that they detail particular problem elements that should be investigated. However, both well- and ill-structured questions can still vary in their degree of specificity. For example, the question "What is the effect of regular physical exercise on the risk of cancer?" is well-structured but still quite broad with respect to the scope of its key constructs. An even more specific question might be "What is the effect of regular long distance running on the risk of colon cancer?".

This discussion of question structure and specificity points towards a potential tradeoff: Research questions that are ill-structured and broad tend to cover a larger problem space and, if answered, could help solve a larger problem. The challenge is that such questions are difficult to answer. Well-structured and specific questions, in turn, narrow the problem space, which makes them easier to answer but may also limit the impact of the knowledge that is generated. This point leads us to consider more explicitly different aspects of the "quality" of RQs.

2.3. The quality of research questions

We synthesize prior work in the study of science and science education to distinguish three key dimensions of research question quality:

First, research questions can differ in their degree of *novelty* (Boudreau et al., 2016; Connolly et al., 1993). A novel RQ has not been researched before and the answer to the question would extend previous findings (Cummings et al., 2007). This novelty may relate to different parts of the research question, including the primary problem of interest (e.g., curing cancer), potential causes or solutions (e.g., physical exercise), as well as their combination.

Second, research questions differ with respect to the importance of the problem that they help solve, or the extent to which they close an existing knowledge gap (Cummings et al., 2007; MacCrimmon and Wagner, 1994). Since research can investigate both problems that directly address practical concerns and problems that do not, we distinguish between the potential *scientific impact* of the RQ (e.g., if answered, how much could we learn about the history of the universe, and how much do we care about understanding the history of the universe) and the potential *practical impact* of answering it (e.g., how much could we reduce the rate of cancer, and how much do we care about curing cancer). As per the earlier discussion, potential impact may relate to question characteristics such as their structure and specificity. Impact may also depend on which particular causes or solutions are included in a RQ, and how promising these elements are with respect to solving the problem.[4]

The prior discussion raises two additional important points. First, judgments of RQ quality are most meaningful ex ante, e.g., when deciding how to allocate resources to competing research projects (Stokes, 1997). This also means that evaluations will be subjective and uncertain, and similarly qualified judges may disagree about the quality of a research question (Guinan et al., 2013; Lamont, 2009). Relatedly, assessments of RQ quality likely depend on the knowledge and preferences of the evaluator. In this paper, we focus on assessments of RQ quality made by professional scientists who are experts in a particular research field. Our rationale is that even if research questions are formulated by the crowd, they will typically be chosen and investigated by (or in collaboration with) professionals with the required training in research methods as well as the necessary physical equipment. Of course, professional scientists may have certain "biases" with respect to the content or form of RQs. We will consider quality evaluations of a broader range of stakeholders in our discussion of opportunities for future research.

2.4. Crowds and research question formulation

2.4.1. A stylized view of the conventional process of RQ formulation

Before we discuss the potential involvement of crowds, we highlight three important features of the process by which research questions are generated in the conventional professional scientific system.[5] First, questions are formulated by professional scientists who have scientific training and research experience. As such, they understand the role of research questions in the research process, including the potential benefits of well-structured questions in guiding research. Moreover, they are aware of prior and current research in the field, allowing them to draw on that work to identify research gaps and develop new questions (Fleming and Sorenson, 2004; Merton, 1973). Second, professional scientists typically work in teams (Wuchty et al., 2007). Even though there is some division of labor of team members with respect to conceptual vs. empirical work, research questions tend to be developed collaboratively

[4] Although we treat the dimensions of quality as independent, they may be related. For example, novel research questions may be judged as having higher scientific or practical impact. Similarly, there may be additional relationships between the RQ characteristics discussed in Section 2.2 and RQ quality. Although a full theoretical discussion of these relationships is beyond the scope of this paper, we will explore them empirically in Section 4.

[5] We highlight a few central features and abstract from potential differences across fields (e.g., natural vs. social sciences) but also different types of research (e.g., theoretical vs. empirical).

Table 1
Crowd paradigms and their potential relevance for generating RQs.

Crowd Paradigm (representative references)	Primary rationale for involving the "crowd"	Target crowd characteristics	Implications/ potential benefits for RQ generation
Crowd labor (Boudreau and Lakhani, 2013; Sauermann and Franzoni, 2015)	A large number of contributors can supply a high volume of labor inputs or generate a large number of ideas.	Large number of contributors, common skills.	• Crowdsourcing can yield a high number of RQ. • Assuming that RQ generation is a stochastic process, a high number of RQ candidates also increases the chance to discover high quality RQs.
Broadcast search (Afuah and Tucci, 2012; Jeppesen and Lakhani, 2010)	Broadcasting a problem or call for inputs to the crowd helps identify high-value outlier solutions or rare inputs.	Diverse crowd with characteristics that increase the probability of having high value solutions (e.g., expert knowledge in different domains).	• Assuming that people have research questions on their minds, broadcast search can identify high-value "outlier" RQs. • Assuming that RQ formulation requires rare knowledge that is difficult to target directly, broadcast search can help find individuals who possess such knowledge.
User innovation (Franke et al., 2006; Von Hippel and Katz, 2002)	Users and professional innovators have different knowledge bases. Users' experiential knowledge can be useful to identify problems that experts may not see, and to find novel solutions.	Crowd with deep experience in the problem domain.	• Crowd members may generate questions with high practical impact. • Assuming that RQ formulation also requires scientific expert knowledge, crowd-generated questions may be of low scientific impact.
Community production (Majchrzak and Malhotra, 2020; Raymond, 1999)	Collaboration among crowd members allows recombination of complementary knowledge and skills, as well as division of labor.	Diverse crowd with complementary knowledge and skills to identify and address elements of complex problems.	• Collaboration and exchange of knowledge may help in problem formulation and increase the structure of RQs. • A broad base of shared knowledge regarding existing solutions may allow communities to set aside RQs that have already been addressed and identify the most novel and open RQs.
Crowd wisdom (Mannes et al., 2014; Mollick and Nanda, 2016)	Aggregating many independent estimates or judgments mitigates individual-level biases and errors.	Large number of independent crowd members with relevant knowledge to judge a particular attribute.	• Less relevant for RQ generation. • Potentially valuable for evaluating and selecting RQs for funding. • If RQ evaluation requires experiential and scientific expert knowledge, aggregating judgments from user crowds and experts may be beneficial.

and thus reflect the knowledge and expertise of multiple researchers (Haeussler and Sauermann, 2020). Third, questions are typically not formulated at a single point in time but developed over time. For example, they may emerge gradually as scientists work on prior projects, and may be refined as current projects are discussed with peers or presented at conferences.

This discussion highlights that "professional" RQs that can be gleaned from conference presentations or articles differ from crowd-generated RQs not only with respect to different creators (professionals vs. non-professionals) but also with respect to the particular mechanisms these creators use. We will keep this point in mind when comparing the characteristics and quality of professional versus crowd-generated RQs in the empirical part of this paper. First, however, we consider different theoretical rationales for involving crowds in research question formulation.

2.4.2. Crowd "paradigms" and potential merits of involving crowds in RQ formulation

A large body of research has studied crowdsourcing and related mechanisms in the context of innovation or other types of tasks. In the following, we synthesize this work to identify five "paradigms" that highlight different rationales for involving crowds. We also discuss whether and how these paradigms might apply to the formulation of research questions. Note that our goal is not to provide a comprehensive review of the literature, and the paradigms we identify are only one possible way to cluster the growing body of work in this domain. However, we suggest that these paradigms provide complementary perspectives that are useful to consider the potential role of crowds in RQ formulation. Table 1 summarizes our discussion.

Crowd labor. An important stream of research focuses on the crowd's ability to support projects with a high volume of effort and labor inputs (Lyons and Zhang, 2019; Theobald et al., 2015). In this paradigm, tasks tend to be standardized and algorithmic in nature, often requiring only common cognitive skills (Franzoni and Sauermann, 2014). Outside of science, this mechanism is central for crowd labor platforms such as Amazon Mechanical Turk (Boudreau and Lakhani, 2013). In science, the crowd labor paradigm is useful to understand large CS projects such as Zooniverse or eBird (see Section 2.1). Studies of crowd labor document the high volume of labor inputs that projects can generate, but also highlight that contributions tend to be very uneven: Especially if projects rely on unpaid volunteers, a small share of highly motivated contributors tend to be responsible for a large share of the inputs (Sauermann and Franzoni, 2015).

The perhaps most distinct aspect of this paradigm is its focus on the volume and scale of crowd contributions. At first blush, this paradigm may seem less relevant in the context of RQ formulation given that the goal is arguably not to produce a large number of RQs, but to generate questions that are particularly novel and impactful (see Section 2.3). However, there may be an important link between quantity and quality if we conceptualize RQ formulation as a creative combinatorial process (Simonton, 2003; Singh and Fleming, 2010): The higher the number of

RQs that are generated, the higher may be the chances that this pool of questions contains a particularly novel or impactful one. In other words, the larger the number of people who get involved in RQ formulation, and the more questions each person generates, the higher may be the quality of the best of these questions.

Broadcast search. A second paradigm focuses not on the volume of contributions but on the wide range of potential contributors that can be reached by broadcasting a problem to a large and diverse crowd (Afuah and Tucci, 2012; Jeppesen and Lakhani, 2010). This has benefits if a project requires highly specialized or rare resources, such as unique skills or pre-existing solutions to a problem. While the number of potential contributors matters, this mechanism is particularly effective when organizers target the "right" crowd, namely people who are more likely to possess the required rare skills or solutions. However, identifying the right crowd is not trivial because it is not always known ex ante what the organization needs or what kinds of skills and knowledge would be most useful (Felin and Zenger, 2014). Broadcast search is featured in many discussions of crowdsourcing for general problem solving (Afuah and Tucci, 2012; Felin and Zenger, 2014; Jeppesen and Lakhani, 2010), where the ideal outcome is an "outlier" solution of particularly high value. This paradigm is also helpful to understand crowd science projects that ask crowds to solve particular (sub)problems, such as the project Foldit (Khatib et al., 2011) or crowdsourcing initiatives at NASA (Lifshitz-Assaf, 2018).

Broadcast search could be a useful lens for research question generation if only few people have access to the knowledge or experience required to formulate high quality research questions, or have pre-existing high-value RQs on their minds. Moreover, broadcast search in a crowd of non-scientists would be particularly beneficial if members of the public are more likely to possess such rare knowledge than the professional scientists who are currently generating RQs (see Section 2.4.1). One of the few documented efforts to crowdsource problems and hypotheses using the broadcast search paradigm was a prize-based contest that targeted a highly educated crowd – primarily patients, relatives, and researchers affiliated with Harvard University (Guinan et al., 2013).[6]

User innovation. The literature on user innovation suggests that users can generate innovative ideas because their personal experience gives them a deep understanding of practical problems as well as potential solutions to those problems (Poetz and Schreier, 2012; Von Hippel and Katz, 2002). In the context of medicine, for example, user innovators include patients who are deeply familiar with a disease and existing treatments, but also medical practitioners who administer treatments and see remaining challenges (Demonaco et al., 2019). The user innovation paradigm highlights the potential divide between the knowledge bases of professional innovators and users but also shows that successful innovation often requires the integration of experiential and technical knowledge (Franke et al., 2006). Organizations can support this integration by providing "toolkits" that transfer some of the technical knowledge that users may otherwise lack (Franke and Piller, 2004; Von Hippel and Katz, 2002).

The user innovation paradigm draws our attention to two particularly interesting aspects: The important role of problem identification and the notion that crowd members may have unique experiential knowledge that differs from the knowledge held by professionals (Von Hippel and Von Krogh, 2016). Applying these ideas to the task of RQ formulation, experiential knowledge may allow crowd members to identify problems that are not recognized by professionals. It may also allow them to structure problems by speculating about potential causes and solutions, perhaps even inspired by their own prior efforts to come up with solutions. By distinguishing different types of knowledge, the user innovation paradigm suggests different expectations regarding the three dimensions of research question quality: On the one hand, involving crowd members with personal experience in a problem domain could yield research questions that have particularly high practical relevance. On the other hand, crowd members who lack important professional or technical knowledge (e.g., knowledge about the state of the art in a research field), may generate questions that have low scientific impact. Expectations regarding RQ novelty are less clear: Not being constrained by expert knowledge and disciplinary boundaries may allow crowd members to generate more novel combinations of problem elements, but the lack of expert knowledge may also prevent them from discarding questions that have already been addressed (Singh and Fleming, 2010).

Community production. The crowd labor, broadcast search, and user innovation paradigms focus on the contributions of individual crowd members who are not interacting. A fourth paradigm - community production - highlights interactions and collaborative contributions. Research in this stream shows that joint efforts of crowd members with diverse knowledge and skills can lead to superior solutions, especially when problems are complex (Raymond, 1999). This research also discusses organizational mechanisms related to division of labor and coordination that facilitate distributed work (Franke and Shah, 2003; Von Krogh et al., 2003).

The formulation of research questions is arguably less complex than the tasks studied in early work on community production (e.g., software development). As pointed out by more recent work, however, problem identification and structuring can similarly benefit from knowledge sharing and iterative collaboration (Foss et al., 2016; Majchrzak and Malhotra, 2020). By drawing on a broader set of shared knowledge, communities may also be able to identify already existing solutions, allowing them to focus their problem formulation on the most pressing open needs (see Singh and Fleming, 2010). Indeed, such knowledge exchange and collective problem formulation can be observed in online medical communities such as Patientslikeme.com or Cysticfibrosis.com, where members discuss their experiences, share existing solutions, and sometimes develop hypotheses regarding causes of conditions or potential new treatments.

Crowd wisdom. Research on the "wisdom of crowds" is a well-established stream of work that focuses on the advantages that crowds have in making predictions or estimating values (Galton, 1907; Surowiecki, 2005). The key mechanism is that if judgments are at least somewhat independent, the biases and errors at the level of individuals may cancel out in larger crowds. More recent work has explored how the wisdom of crowds depends on the decision problem, the distribution of expertise, how crowd members interact, and how individual judgments are aggregated (Mannes et al., 2012, 2014; Simmons et al., 2011). Whereas the four paradigms discussed up to this point consider how crowds *create* objects (e.g., ideas, innovations, new products), crowd wisdom is particularly useful in thinking about how crowds *select*. One particularly relevant example is crowdfunding of innovative projects, which aggregates the judgments of many people with respect to project attributes such as the likelihood of technical success or fit with consumer preferences (Butticè et al., 2017). While experts may have advantages with respect to judging some of these dimensions, consumers may have advantages with respect to others (Mollick and Nanda, 2016). As such, the composition of the crowd matters, and approaches that integrate judgments from both experts and laypeople may be particularly effective (Mannes et al., 2014).

The crowd wisdom paradigm appears less relevant when considering crowd involvement in the *generation* of research questions. However, it may prove useful when thinking about how to involve crowds in selecting research questions or proposals. For example, the scientific crowdfunding platform Experiment.com aggregates crowd judgments regarding the scientific merit and practical relevance of research

[6] This project was implemented via InnoCentive and participants could compete for 30,000 USD in prizes. The project was advertised in the Harvard Catalyst community, InnoCentive community, as well as the journal *Nature*. Although the general public could contribute, 41% of the contributions came from Harvard faculty, students, and staff (Guinan et al., 2013).

projects. Similarly, the Council of the Region of Southern Denmark recently asked the general public to vote on funding proposals, hoping to generate a more "democratic" assessment of the societal relevance of projects (Franzoni et al., 2021). The effectiveness, boundary conditions, and potential biases of these approaches to use crowds in RQ selection still have to be investigated.[7]

Before we turn to the empirical analysis, we note two more general points. First, the different crowd paradigms we identified highlight different mechanisms, consider somewhat different types of crowds, suggest different approaches to involve crowds, and highlight different aspects of performance. At the same time, there are overlaps between the paradigms, and the different mechanisms may complement each other in practice. For example, the crowd science project Foldit seeks to identify outlier solutions to a particular protein folding problem and relies on contributors who tend to have unusual skills in solving 3D puzzles (i.e., key features of broadcast search). But even individuals with such rare skills do not have solutions ready and need to spend considerable time developing solutions (crowd labor), which involves not only much trial and error but also collaboration among contributors (community production).

Second, the relevance of the different paradigms partly depends on the nature of the task, e.g., the degree to which the task requires crude effort vs. creativity vs. scarce skills and knowledge. Given the importance and complexity of research question formulation, we suggest that each paradigm highlights important aspects of crowd involvement that may be relevant when studying crowdsourcing RQs – except perhaps for crowd wisdom, which we included to anticipate the role of crowds in judging RQs. Thus, the paradigms will provide a useful lens to probe mechanisms and interpret results in the subsequent empirical analysis. However, we will not be able to "test" paradigms or speak to the general merits of one paradigm over another because the explanatory power of a given paradigm also depends on how a crowdsourcing project is set up, which, in turn, depends on the assumptions that organizers have about the applicability of different paradigms. The projects we study in the next sections were not designed by the organizers with any particular paradigm in mind. Rather, the idea was to enable multiple crowd-related benefits to materialize, with the overarching goal to create high-quality research questions. As such, the goal of our empirical analysis is to assess the crowd's ability to generate research questions, as evidenced in the particular projects we study. We will interpret the empirical results in light of the five paradigms in the subsequent discussion (see Section 5).

3. Methods

3.1. Crowdsourcing projects and RQ extraction

We analyze data from two projects that crowdsourced scientific research questions. These projects were designed and implemented by the Ludwig Boltzmann Gesellschaft (LBG), a European research and funding institution focusing on the medical sciences. The institution's goal was to identify new and promising topics that could serve as input for its research efforts; it explicitly sought to incorporate knowledge that does not originate from within the professional scientific discourse and to link research more closely to societal challenges.[8] Although similar in their aims, the two projects differed somewhat regarding the targeted scientific field, the campaigning strategy, the composition of the crowd, and how crowd contributions were elicited (see Appendix A for an overview).

Project 1 aimed to identify research questions in the field of mental illnesses. The project sought to recruit individuals with high experience in this area, including patients, caregivers (e.g., relatives of patients), and medical practitioners such as nurses and physical therapists. As such, online and offline campaigns targeted relevant stakeholder groups such as patient organizations as well as associations for medical practitioners, asking their members to participate.

Project 2 was conducted in the field of traumatology (i.e., accidental physical injuries of bones, tissues and ligaments). Two channels were used to recruit contributors. First, LBG initiated an online and offline campaign to attract participants with experience in the field of accidental injuries, including patients, caregivers and medical practitioners. Second, the general-purpose crowdsourcing platform Clickworker.com was used to invite individuals with personal or professional experience with accidental injuries to participate for a small monetary reward. Participants recruited via both channels were directed to the same custom-designed website to contribute their RQs.

Both projects invited contributors to suggest research questions in the focal domain (mental illness and traumatology, respectively). Project 1 allowed contributors to submit any form and quantity of text via a website that was designed for the purpose of that project. To extract research questions from these submissions, we followed a three-step process. First, two trained social scientists examined all "raw" contributions and marked text passages that included research questions or question-like statements such as "I would find it interesting to study whether…". Second, the social scientists extracted RQs that were either identical to the original text (for questions) or as close as possible to it (for question-like statements).[9] Third, disagreements between the two social scientists about the extraction were fully resolved. Out of the 422 raw contributions, 140 contained no research questions or question-like statements. From the remaining 282 contributions (by 155 unique crowd participants) a total of 753 research questions could be extracted. The crowd contributions varied greatly in length (from 54 to a maximum of 20,004 characters) and in the number of questions that were extracted. While 29.0% of the contributors generated only one question, 65.2% contributed between two and ten questions. Nine (5.8%) crowd members contributed more than ten questions, including four that submitted 28, 30, 46, and 86 questions, respectively.

Project 2 followed a different process in that crowd members were invited to directly submit exactly one research question in an entry box on a customized website, leading to a total of 180 submissions from 180 individuals.[10] We excluded 29 contributions that were incomplete or invalid (e.g., entries such as "N/A").

To allow for meaningful expert evaluations, we excluded questions that were completely unrelated to the respective fields (7 in project 1

[7] Our discussion of the paradigms has focused on the respective rationales and benefits of involving crowds, but there are also challenges that need to be considered. Some of these have been discussed in prior research, including identifying potential contributors, motivating (especially unpaid) contributions, enabling collaboration and knowledge sharing, and screening high volumes of contributions for quality. We will speak to some of these issues in the empirical part, as well as in our discussion of opportunities for future research (see Section 5).

[8] https://ois.lbg.ac.at/en/projects/crowdsourcing-research-questions-in-science

[9] To illustrate, an example for a question-like statement is "Of interest would be studies that clarify how high the costs are that arise because prescribed and purchased psychotropic drugs are not taken by the patients concerned, especially, for example, because the drugs are thrown away or are not taken properly." and the extracted research question "How high are the costs that arise because prescribed and purchased psychotropic drugs are not taken by the patients concerned, especially because the drugs are thrown away or are not taken properly?".

[10] In addition to the process described here, a random sample of crowd members in project 2 was asked to submit RQs using structured forms designed to help the formulation of RQs. We do not use these cases in the present paper to obtain unbiased insights about the nature and quality of crowdsourced RQs without assistance.

Table 2a
Research question characteristics, Project 1 (Mental Illness).

Category	Level	Example	Share/ Mean/SD
Structure	Problem restatement	How can you help people with hoarding disorder?	59.1%
	Problem and cause	How does the biography of the parent and grandparent generation influence the development of mental illnesses?	18.6%
	Problem and solution	Can mental health strategies be optimized through personalization (e.g., considering a patient's past experiences, understanding the patient's personality etc.)?	12.3%
	Relationship	What is the interaction between autoimmune diseases and autism/Asperger's syndrome?	9.3%
Specificity	Specific question	Do patients who have received milieu therapy have greater self-confidence and self-esteem than patients without this treatment?	Mean: 3.91 SD: 0.96
	Broad question	How can the humanities (e.g. philosophy, ethics, law, sociology, psychology, pedagogy/education) be included in research into the lifeworld of mentally ill persons?	Min: 1 Max: 5
Discipline-crossing nature	Not discipline-crossing	What is the role of traumatic experiences in the development of mental illnesses?	55.2%
	Medical field *(e.g., immunology)*	What is the role of the regulation of the immune system in the diagnosis and therapy of psychosomatic diseases?	10.1%
	Non-medical field *(e.g., media/ technology)*	What is the interplay between new media (like the internet) and mental health?	34.3%
Question length	Short question	How are poverty and mental illnesses related?	Mean: 136.29 SD: 75.58 Min: 31 Max: 1172
	Long question	Which basic conditions in the intramural area have a lasting favorable or less favorable effect on patients (e.g., distance from the place of residence or social environment, visit "ban", stronger obligation to participate in therapy programs, single room accommodation, more individual therapies, qualitatively better nutrition, new/less frequently used therapy forms as accompanying measures)?	

Note: The illustrative examples include both high and low quality RQs.

and 4 in project 2).[11] Our final sample includes 746 questions from 155 crowd contributors in project 1 and 147 questions from 147 unique contributors in project 2. The contributors were diverse with respect to their age (average 45 years in project 1 and 37 years in project 2) as well as gender (62.6% female in project 1 and 46.3% in project 2). In project 1, large shares of contributors were medical practitioners (50.32%) as well as patients and relatives of patients (13.6%). In project 2, the crowd consisted mainly of patients and relatives of patients (76.2%) and medical practitioners (19.1%).[12] Note that medical practitioners such as nurses and physical therapists fall within our definition of "crowd" because they are not engaged in scientific research. In other words, they are professional "users" of scientific knowledge but, unlike professional scientists, they are not "producers" of such knowledge. Just like patients and patient relatives, medical practitioners are particularly promising members of the crowd because their experience in the problem domain may provide them with unique knowledge that can help identify valuable research questions. Appendix B summarizes additional characteristics of the crowd members and their submissions.

3.2. Professional research questions

To compare the quality of crowd-generated questions to that of questions generated as part of the conventional scientific research process, we sampled research questions from international conference proceedings. Towards this end, we asked a professional scientist in mental health to list and rank the most relevant international research conferences in that field, and another professional scientist to do the same for traumatology. We then checked each conference for topical fit, the public availability of conference proceedings, and whether it took place in the same year as the corresponding crowdsourcing project. This process identified eleven conference proceedings for each project, from which we randomly selected a set of working papers. We then extracted from these papers research questions using a very similar procedure as in the case of text submitted by crowd members (see Section 3.1).[13] This process resulted in 103 professional research questions on mental illness (project 1) and 100 RQs in traumatology (project 2). Applying the same exclusion criteria as to the crowdsourced questions (i.e., dropping questions that were evaluated as completely unrelated to the field) resulted in a final set of 99 professional RQs in mental illness and 83 questions in traumatology.

We selected questions from conference proceedings rather than published articles to get questions that are "earlier stage" and more comparable to the crowdsourced questions. As noted in Section 2.4.1, however, even these questions are the result of longer and more complex processes (e.g., involving team work, multiple revisions by authors, peer feedback).[14] Moreover, the questions that appear in conference proceedings are likely selected based on perceived quality by both the authors submitting papers and the reviewers who accept them. Thus, differences between crowd-generated and professional RQs reflect not only differences resulting from involving different types of creators but also differences resulting from different RQ generation processes. Given that RQs such as those extracted from conference proceedings represent the "status quo" available in the professional system, it is particularly interesting and relevant to see how crowd-generated questions compare.

3.3. Measures

Two social scientists evaluated the structure, specificity, and the discipline-crossing nature of all research questions.[15] Professional scientists in the fields of mental illness and traumatology evaluated the

[11] We asked evaluators (see next section) to what extent a question is related to the focal scientific field (1="not at all related"; 2="somewhat related"; 3="very related")).

[12] Demographic information stems from crowd members' registration on the custom-designed project website. This information was optional in project 1, leading to missing data regarding age (5.8%), gender (2.6%), and type of experience with mental illnesses (36.1%).

[13] If the conference article stated a clear research question, we used this question. In the case of question-like statements, we made minor adjustments to form a regular question structure. Just like for crowd contributions, we dropped an article from consideration if it did not include a research question or question-like statement.

[14] Highlighting the different RQ generating processes, most of the conference papers were co-authored (75.0% in project 1 and 96.0% in project 2).

[15] One social scientist is a co-author of this paper and was blind to the source of the research questions. The other was independent and blind to the purpose of this study and to the source of the research questions.

Table 2b
Research question characteristics, Project 2 (Traumatology).

Category	Level	Example	Share/ Mean/ SD
Structure	Problem restatement	What should you do if you have cut off a finger?	58.5%
	Problem and cause	Does gender affect the reporting of an accidental injury to a medical professional?	23.8%
	Problem and solution	What influence do tactile stimuli, such as effleurage, brushing, lentil baths, etc. have on pain processing and edema reduction?	14.3%
	Relationship	Is there a correlation between mental disorders and accidental injury rates?	1.3%
Specificity	Specific question	Do accidents occur more frequently amongst different ages, gender, races, nationalities?	Mean: 2.69 SD: 1.33
	Broad question	What methods of injury prevention are effective?	Min: 1 Max: 5
Discipline-crossing nature	Not discipline-crossing	Can chronic pain as a consequence of injury to the wrist be reduced/ avoided through therapeutic interventions in occupational therapy?	29.9%
	Medical field *(e.g., pharma)*	To what extent can DMSO (dimethyl sulfoxide) accelerate/ usefully complement the healing process?	3.4%
	Non-medical field *(e.g., sociology)*	To what extent do accidental injuries make an impact on the victim's social life?	50.3%
Question length	Short question	How can internal injuries be recognized more quickly?	Mean: 93.36
	Long question	What influence does a regular psychotherapy consultation taking place as early as possible in the context of the treatment of moderate accidental injuries have on the duration and the experienced difficulty of the healing process?	SD: 42.63 Min: 27 Max: 298

Note: The illustrative examples include both high and low quality RQs.

quality of research questions with respect to novelty, scientific impact, and practical impact. Appendix C summarizes descriptive statistics of all measures.

3.3.1. Measures of RQ characteristics

Structure. Reflecting our theoretical discussion of question structure (see Section 2.2), the evaluators distinguished RQs that were simple problem restatements, RQs that included a problem with potential causes or solutions, and RQs about undirected relationships between different constructs (see Tables 2a and 2b for examples). The raters had an initial agreement of 72.7% in project 1 and 75.1% in project 2; they reached consensus after discussing disagreements. The dummy variable *well_structured* is coded as 0 for questions that are simple problem restatements and 1 for research questions that include problems and causes/solutions, or undirected relationships between two constructs.

Specificity. Even for a given question structure, some questions entail general constructs or relationships (e.g., "What is the relationship between capitalism and mental illnesses?"), while others entail more specific constructs or relationships (e.g., "To what extent does the use of neuro feedback help psychosis patients to recognize or prevent emerging panic attacks/episodes?"). Raters assessed each question on a 5-point scale ranging from 1 (very broad) to 5 (very specific). After a discussion between the coders, their agreement was satisfactory considering the creative nature of the content (Boudreau et al., 2016; Gwet, 2014).[16]

Discipline-crossing. Some questions only included concepts from the targeted scientific field (i.e., mental illness or traumatology). Others integrated concepts from different medical fields (e.g., "Are patients with mental illnesses such as depression at greater risk of injury?") or even non-medical fields (e.g., "What is the interplay between new media (like the internet) and mental health?"). To cross-check the coding of questions that used specialist medical terminology, the coders consulted experts form the respective scientific fields. The two raters had an initial agreement of 68.4% in project 1 and 82.8% in project 2. After discussing disagreements, they reached consensus in project 1 and 83.6% agreement in project 2.[17] For descriptive purposes, we distinguish three levels of this measure (1 = not discipline-crossing, 2 = discipline-crossing with medical field, 3 = with non-medical field). In regression analyzes, we use a dummy variable *discipline-crossing* that takes on the value of 0 for questions that only included concepts from the focal field and 1 for questions that also included concepts from other medical or non-medical fields.

Question length. Although we had not discussed question length in the conceptual part, we capture this attribute for descriptive purposes, using the count of characters. Given the skewed nature of this count, we perform regression analyzes using the logarithmized value (*ln_length*).

3.3.2. Measures of RQ quality

In project 1, a senior professional scientist rated each RQ along the three quality dimensions in an online evaluation tool using 5-point scales. The dimensions were defined as follows. *Novelty*: "How novel is this question compared to the current state of research on mental illness?" (from 1="not at all" to 5 = "very novel"). *Scientific impact*: "To what extent would answering this question advance research on mental illness?" (from 1 = "not at all" to 5 = "great extent"). *Practical impact*: "To what extent could answering this question affect the lives of patients, relatives, caregivers or medical practitioners in the field of mental illness?" (from 1 = "not at all" to 5 = "great extent"). Each rating scale also provided a "can't assess" option for the case that the research question was unclear to the expert and a particular quality dimension could not be assessed (coded as missing). We validated the expert evaluation by investigating a small sample of RQs independently using online tools.[18]

In addition to using the ordinal measures, we also computed a set of dichotomous measures that indicate whether a particular question was "top rated", defined as receiving a score of 4 or 5 on the respective quality dimension. The resulting measures are *novelty_top, scientific_impact_top*, and *practical_impact_top*. Finally, we code the dummy variable *all_top* that takes on the value of 1 if a question received top scores

[16] In project 1, Gwet's coefficient (ordinal weights) increased from 0.63 (initial agreement) to 0.93 (after discussing disagreements); in project 2, Gwet's coefficient increased from 0.55 to 0.69 (Gwet, 2014).

[17] If no agreement could be reached, the variable *discipline-crossing* was recoded as missing.

[18] We randomly selected 6 questions. Two social scientists who were blind to the expert ratings searched scientific databases and Google Scholar to identify papers and conference contributions that covered the same content as the research questions. They then assessed the novelty of the RQs based on how many articles on the subject were available at the time of the crowdsourcing project. They assessed scientific impact based on the citations of the identified articles and checking for existing special issues or specialized conferences. They assessed practical impact based on the articles' Altmetric scores and by performing Google searches to determine the "popularity" of the content of the questions. Based on these various inputs, each quality dimension was rated on the same 5-point scale used by the experts. This systematic search yielded an exact agreement between the social scientists and the expert on 72.2% of the ratings.

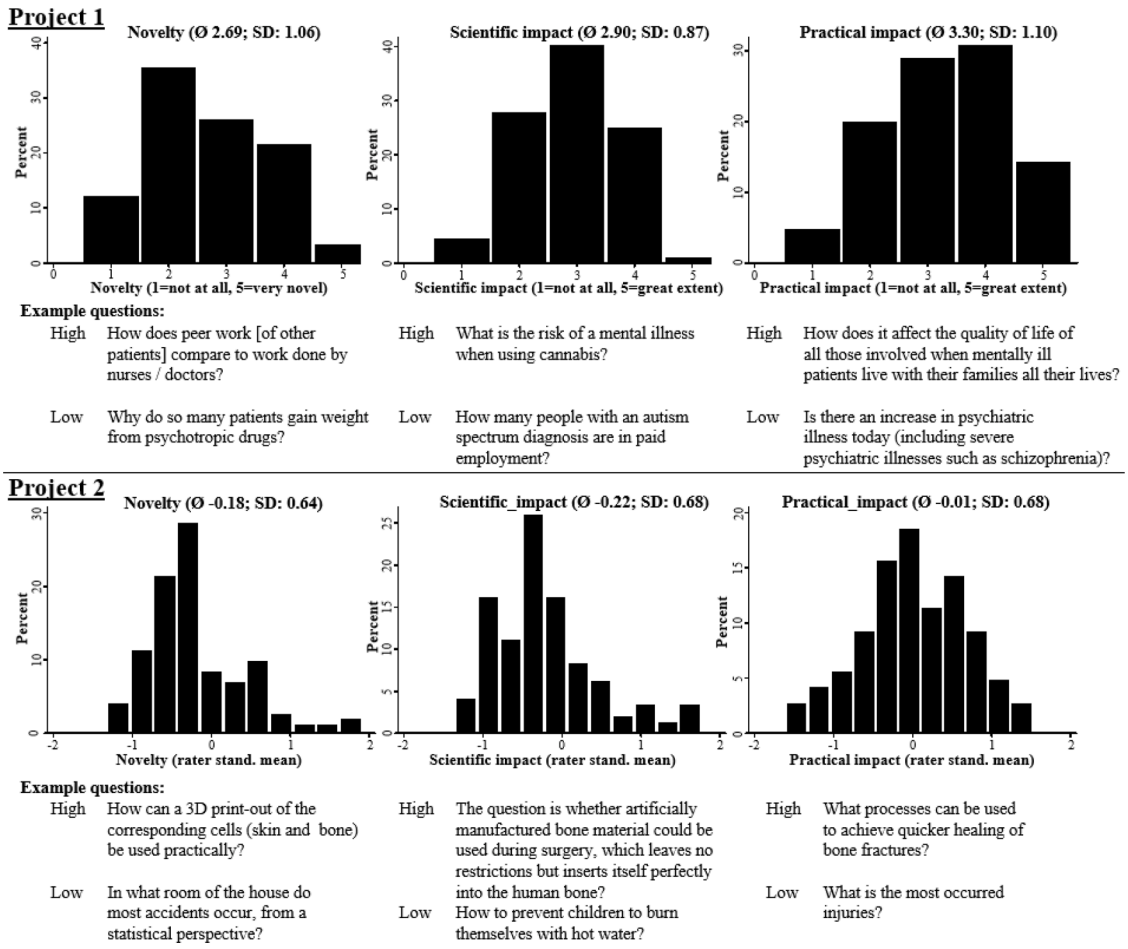

Fig. 2. Distribution of quality ratings for project 1 and project 2.
Notes: All three quality dimensions were evaluated on 5-point rating scales ranging from 1 = not at all to 5 = very/great extent. Questions in project 1 were rated by one scholarly expert ($N = 746$). RQs in project 2 were rated by multiple experts and figures display the average of the rater-standardized ratings ($N = 147$).

on each of the three dimensions.

The evaluation in project 2 was similar except for two aspects. First, research questions were evaluated by four scholarly experts (vs. just one). After asking the raters to evaluate all questions individually using an online tool (similar to project 1), we conducted a one-day workshop to discuss disagreements among raters (Krippendorff, 2004).[19] To address potential rater-specific evaluation differences (e.g., some raters being "tougher" than others), we standardized the ratings for each dimension by evaluator. We then computed the final quality measures for each question as the means of these standardized scores (*novelty, scientific_impact*, and *practical_impact)*. Second, in addition to the "can't assess" option, we also asked evaluators explicitly to rate the clarity of each research question ("How clear is it to you what this question seeks to examine?"; 1 = "not at all" to 5 = "fully clear") (Durand and Van-Huss, 1992). To ensure greater reliability of the quality ratings, we recoded quality ratings as missing if an evaluator indicated that a question was "not at all" clear.[20]

Similar to project 1, we computed a set of dichotomous measures that indicate whether a particular question was "top rated" (*novelty_top, scientific_impact_top*, and *practical_impact_top*). Given that the means of the standardized scores are continuous, we define "top" scores as those in the top 40% of the range of the scores considering both crowd-generated and professional RQs (analogous to project 1, where we used 40% of the range by selecting questions with scores of 4 and 5). We again computed a dummy variable *all_top* that equals 1 if a question received top scores on all three dimensions.

3.3.3. Other measures

The variable *questioncount* captures the number of questions submitted by a particular contributor and ranges from 1 to 86 in project 1. By construction, all questions submitted by a particular contributor have the same count. Project 2 allowed only one submission per contributor and this variable takes the value of 1. Given the skewed distribution of *questioncount*, we use the natural log in regression analyzes (*ln_questioncount)*.

[19] Experts could change their ratings during the workshop. After the workshop, Gwet's agreement coefficients (ordinal weights) ranged from .44–.62. Although satisfactory, these metrics are not very high, consistent with prior studies using expert evaluations of RQ quality (Guinan et al., 2013) and likely reflecting the inherent subjectivity of such judgments (see Section 2.3).

[20] Excluding an evaluator's quality ratings when a question was not at all clear is comparable to the exclusion of crowd submissions that did not allow for extracting a valid research question in project 1. A particular quality dimension is coded as missing for the question as a whole if only one evaluator gave a valid rating.

Table 3
Relationships between question characteristics and quality dimensions.

	Project 1			Project 2		
	(1) Novelty	(2) Scientific impact	(3) Practical impact	(4) Novelty	(5) Scientific _impact	(6) Practical _impact
Well-structured	0.482**	0.524**	0.032	0.239^{+}	0.012	−0.258^{+}
	(0.169)	(0.143)	(0.154)	(0.142)	(0.160)	(0.154)
Specificity	−0.275**	−0.290**	−0.168*	0.017	−0.080	−0.126*
	(0.094)	(0.080)	(0.079)	(0.049)	(0.052)	(0.051)
Discipline-crossing	0.281*	0.330*	0.272^{+}	−0.038	−0.272^{+}	−0.035
	(0.127)	(0.141)	(0.143)	(0.124)	(0.151)	(0.132)
Ln_length	0.777**	0.706**	0.554**	0.453**	0.452**	0.591**
	(0.164)	(0.161)	(0.151)	(0.143)	(0.167)	(0.166)
Constant				−2.297**	−1.801*	−2.104**
				(0.596)	(0.712)	(0.735)
Observations	715	731	732	119	120	119
Pseudo R^2	0.0273	0.0302	0.0110			
R^2				0.183	0.084	0.152

Notes: Models 1–3: ordered logistic regressions (Ologit). Models 4–6: OLS regressions; robust standard errors in parentheses (clustered by contributor for Project 1).
** $p < 0.01$, * $p < 0.05$, + $p < 0.10$.
In project 1, a single evaluator rated all RQs on each quality dimension on 5-point scales.
In project 2, multiple evaluators rated all RQs on each quality dimension on 5-point scales. These ratings were standardized by rater and then averaged.

4. Results

4.1. Characteristics of crowd-generated research questions

In a first set of analyzes, we explore the structure, specificity, discipline-crossing nature, and length of crowd-generated RQs. Given the different approaches to elicit questions in projects 1 and 2, we show results separately in Tables 2a and 2b.[21]

With respect to structure, we observe that research questions most often take the form of a problem restatement (59.1% in project 1 and 58.5% in project 2), i.e., contributors stated a problem without suggesting potential causes or solutions. Even though such "ill-structured" questions provide little guidance regarding how a problem might be studied or solved, they can be valuable if they point to important problems that may have been ignored in the past.[22] Among the well-structured questions, the most common type includes problems and potential causes (18.6% in project 1 and 23.8% in project 2), followed by questions that include problems and potential solutions. The smallest group of questions asked about a non-directional relationship between two constructs. The examples shown in Tables 2a and 2b illustrate that question structure does not simply refer to the particular language or terminology that respondents used. Rather, well-structured questions include additional informational content: they go beyond identifying a potentially important problem to give additional insights into contributors' experience with that problem and implicit theories about the problem.

The average rating of specificity in project 1 was 3.91 on a 5-point scale, indicating that crowd-generated RQs tended to be quite specific. Questions were less specific in project 2 (average = 2.69). One potential explanation for the difference between projects is that contributors had unlimited space to describe their experiences and questions in project 1 but were given a more restrictive entry box in project 2.

We observe that 44.4% of the questions in project 1 are discipline-crossing, with most of these combining medical and non-medical fields (34.3%) and a smaller share combining different medical fields (10.1%). In project 2, the share of discipline-crossing questions is even larger (53.7%), again with the majority of these questions crossing the disciplinary boundary between medicine and other fields. Thus, many crowd members do not restrict themselves to disciplinary boundaries and often connect medical with non-medical fields. This observation is interesting because it supports the notion that the lack of scientific training may allow crowd members to think outside disciplinary silos, but also that their practical experience with particular problems allows them to see those problems in a broader context (see Section 2.4.2, especially on the user innovation paradigm).

Finally, the average length of research questions is 136 characters in project 1 and 93 characters in project 2. The difference is likely to reflect again the more open-ended approach to elicit RQs in project 1.

Appendix D shows correlations between question characteristics for both projects. Among others, well-structured questions tend to be longer, more specific, and more likely to be discipline-crossing. Specificity is also positively correlated with question length.

4.2. Research question quality

Fig. 2 shows the distributions of our measures of novelty, scientific impact, and practical impact for both projects. We also include examples of questions that received very high and very low ratings for each dimension.

In project 1, the average rating is highest for practical impact (mean = 3.30), followed by scientific impact (mean = 2.90) and novelty (mean=2.69). Comparatively few questions receive ratings of 4 or 5 on novelty (mean *novelty_top* = 0.26) and scientific impact (mean *scientific_impact_top*=0.26), but almost half of the questions contributed by the crowd receive top scores on practical impact (mean *practical_impact_top*=0.46).

For project 2, we do not report average scores given the use of standardization (see Section 3.3.2).[23] Yet again we see that relatively few crowdsourced questions receive ratings in the upper 40% of the scale on novelty (mean *novelty_top*=0.13) and scientific impact (mean *scientific_impact_top*=0.13), while the share of questions that are top-rated with respect to practical impact is considerably higher (mean

[21] In case of missing data, we use pairwise deletion throughout this paper (Newman, 2014).

[22] The frequency with which particular problems are raised may also give scientists and policymakers insights into problem importance and be helpful in setting research priorities at the level of broader research portfolios. Of course, problem importance depends not only on frequency but also severity (Murray and Lopez, 2013). Moreover, the frequency with which problems are mentioned will be influenced by the composition of the crowd and may be biased if crowds are not representative of the broader population (Sauermann et al., 2020).

[23] Unstandardized ratings are as follows. Novelty: mean = 1.94, SD = 0.61; scientific impact: mean = 2.20, SD = 0.80; practical impact: mean = 3.01, SD = 0.80. These figures suggest that the practical impact of crowd-generated questions is higher than novelty and scientific impact also in project 2.

Table 4
Project 1: Crowdsourced vs. professional research questions.

	Crowd (random pick) vs. prof RQs			Crowd (best pick) vs. prof RQs			Crowd (random pick) vs. prof RQs			Crowd (best pick) vs. prof RQs		
	(1) Novelty	(2) Scientific impact	(3) Practical impact	(4) Novelty	(5) Scientific impact	(6) Practical impact	(7) Novelty	(8) Scientific impact	(9) Practical impact	(10) Novelty	(11) Scientific impact	(12) Practical impact
Crowd	−0.758**	−0.756**	0.285	0.619**	0.622**	1.623**	−0.906**	−0.592+	0.256	−0.984**	−0.727*	0.241
	(0.247)	(0.231)	(0.218)	(0.238)	(0.239)	(0.232)	(0.334)	(0.321)	(0.295)	(0.349)	(0.326)	(0.335)
Ln_questioncount							0.137	−0.158	0.028	1.622**	1.386**	1.564**
							(0.195)	(0.210)	(0.183)	(0.208)	(0.204)	(0.228)
N	219	252	254	223	252	254	219	252	254	223	252	254
Pseudo R^2	0.0138	0.0167	0.00210	0.00908	0.0114	0.0618	0.0149	0.0181	0.00214	0.124	0.100	0.147

Notes: Ordered logistic regression (Ologit); robust standard errors in brackets. Professional research questions as benchmark.
** $p < 0.01$, * $p < 0.05$, + $p < 0.1$.
The "Random pick" sample includes one randomly drawn question from each crowd contributor. The "Best pick" sample includes the best question from each crowd contributor on a particular dimension.

practical_impact_top=0.35).

Taken together, the crowd generates research questions that are evaluated by experts as, on average, moderately novel and with moderate scientific impact. Questions tend to be rated higher on practical impact than on novelty and scientific impact, consistent with the idea that crowds with experiential knowledge may have a particular strength with respect to recognizing important problems and/or solutions that may have practical impact (see Section 2.4.2). Furthermore, there is a wide distribution of quality, including a few questions rated very highly on novelty and scientific impact, and a considerable share of questions with top ratings on practical impact.

The correlation matrix (Appendix D) shows that the three dimensions of RQ quality are positively correlated with each other in both projects. Moreover, quality is related to some of the characteristics explored in the prior section (structure, specificity, discipline-crossing, and length). We explore the latter relationships further by estimating a series of regressions using the quality measures as dependent variables and question characteristics as predictors (see Table 3). All regressions use robust standard errors, in project 1 clustered by contributor.

In project 1 (models 1–3), well-structured questions are rated as significantly more novel and of higher scientific impact. This is consistent with the notion that well-structured research questions provide more opportunities to create novel combinations (e.g., of problems and potential causes or solutions), and are more useful in guiding research (see Section 2.2). Specificity has a positive correlation with structure, but including both jointly in the regression yields a negative relationship between specificity and RQ quality. As per our conceptual discussion, a potential explanation is that specificity tends to narrow the problem space, reducing potential impact. Discipline-crossing questions are rated higher on novelty as well as scientific impact, which is consistent with arguments made by proponents of interdisciplinary research (e.g., National Academies, 2004). Question length has a positive relationship with RQ quality. Regressions using data from project 2 (models 4–6) suggest similar qualitative results but coefficients tend to be somewhat smaller and have lower statistical significance, likely reflecting the smaller sample size.[24]

The results in Table 3 should be considered as purely correlational. Future research is needed to examine how exactly question characteristics influence expert judgments, and how quality judgments may differ between experts and other types of judges. For the subsequent analyzes, we take research question characteristics and expert evaluations of quality as given.

4.3. Comparing crowd-generated and professional research questions

We now examine how the quality of crowd-generated research questions compares to that of RQs generated in the conventional scientific process.

4.3.1. Project 1

Recall that project 1 (mental illness) allowed contributors to submit multiple questions; the number of questions submitted ranged from 1 to 86, with an average of five questions per contributor. To avoid biases due to the dominance of questions from particularly prolific contributors, we use only one question from each contributor, selecting that question using two different approaches. First, we randomly draw one question from those submitted by a particular contributor (in the case of a single submission, that question is selected), resulting in a set of *random-pick* questions.[25] Second, we select the *best* question from each contributor, using the expert's rating on a particular dimension (*best-pick*).[26] As expected, the quality of best-pick questions is significantly higher than that of random-pick questions (average novelty 3.22 vs. 2.53; scientific impact 3.34 vs. 2.85; practical impact 3.74 vs. 3.07).

To compare the quality of the two sets of questions to that of professional questions, we regress the dimensions of RQ quality on a dummy *crowd* that takes the value of 1 for crowd-generated questions and 0 for questions generated in the conventional professional scientific process (see Table 4). Models 1–3 use the *random-pick* crowd questions and show that these questions have lower novelty and scientific impact than professional research questions; we see no significant difference in practical impact. The picture changes dramatically when we focus on the *best-pick* crowd questions (models 4–6): crowd questions now have significantly higher novelty, scientific impact, and especially practical impact than professional questions. One interpretation of this finding is that the average – or randomly selected – crowd question has lower

[24] A notable difference compared to project 1 is that the coefficients of *discipline-crossing* are no longer positive and even slightly negative. A supplementary analysis reveals that this result is driven by questions that combine traumatology with non-medical fields, not by questions that combine traumatology with other medical fields (which were more common in project 1, see Table 2a and 2b). Given that our measures of novelty and impact specifically use the focal medical field as a reference point, a potential explanation is that integrating fields with a moderate distance (i.e., other medical) is perceived by professional scientists to be on average more beneficial than integrating fields with greater distance (i.e., non-medical).

[25] As a robustness check, we ran all regressions for project 1 also with the entire sample of RQs. The results are very similar to those obtained using the "random pick" questions.

[26] If multiple questions are tied with the highest score, we again pick randomly. For example, consider a contributor who submitted 5 questions, where 2 questions are rated novelty = 5 and 3 questions have lower novelty scores. For the "random pick" set, we randomly draw one of the 5 questions. For the "best pick" set, we randomly choose one of the two questions rated 5. The "best-pick" questions are chosen separately for each quality dimension.

quality, but allowing individual crowd members multiple "trials" increases the chance that they contribute a high-quality question (see Simonton, 2003). An alternative explanation could be that prolific contributors who submit many questions are more intrinsically motivated or have more experience and thus produce questions of higher average quality (Amabile, 1996). To distinguish these two possibilities, we re-estimate regression models 1–6 and additionally include the number of research questions submitted by a particular contributor (*ln_questioncount*).[27] Models 7–9 use the sample of *random-pick* questions and show no significant coefficient of ln_*questioncount*. In other words, randomly drawn questions from contributors who submitted many questions are not better than questions from contributors who only submitted a single question. Using the *best-pick* questions, however, the question count variable has a significant positive coefficient: The more questions someone submitted, the better is the "best" of these questions (models 10–12). Once ln_*questioncount* is included, the coefficients of *crowd* change noticeably and return to levels seen in the regression using the *random-pick* questions (e.g., model 12 vs. model 3). As such, the selection from among multiple submissions seems to explain the outperformance of *best-pick* questions relative to professional questions. Although we find no evidence that prolific contributors are per se "better" (the quality of their randomly drawn RQs is not higher), the fact that they submitted a larger number of research question points to higher levels of motivation or interest in the crowdsourcing project.

Given the wide distribution of RQ quality, we move beyond mean comparisons to also examine questions of particularly high quality. For this purpose, we use the dummy variables indicating whether a question was rated as 4 or 5 with respect to a particular dimension (*novelty_top, scientific_impact_top*, and *practical_impact_top*) and whether a question was rated very highly on all dimensions (*all_top*). Table 5, models 1–4 estimate linear probability models using the random-pick sample and show that crowd-generated research questions are just as likely to be top rated as professional questions.[28] Thus, although random-pick crowd questions underperformed with respect to mean novelty and scientific impact (see Table 4), this disadvantage disappears once we focus on top questions. Models 5–8 use the best-pick sample and again show a strong positive coefficient of *crowd*; crowd questions outperform when picking the best of multiple submissions from a given contributor. The variable ln_*questioncount* has no significant coefficient in the sample of random-pick questions (models 9–12) but has a significant positive coefficient in the best-pick sample (models 13–16). Moreover, including this variable reduces the coefficient of *crowd* to insignificance in the best-pick sample, again suggesting that selection is the primary explanation for the outperformance of *best-pick* questions.[29]

4.3.2. Project 2

In project 2 (traumatology), contributors were asked to submit only a single research question. In principle, those submissions could simply be the first questions that came to contributors' mind (similar to the random-pick sample in project 1). However, contributors may also have thought about multiple questions and then submitted what they thought

Table 5
Project 1: Top-rated crowdsourced vs. professional research questions.

	Crowd (random pick) vs. prof RQs				Crowd (best pick) vs. prof RQs				Crowd (random pick) vs. prof RQs				Crowd (best pick) vs. prof RQs			
	(1) Novelty_top	(2) Scientific_impact_top	(3) Practical_impact_top	(4) All_top	(5) Novelty_top	(6) Scientific_impact_top	(7) Practical_impact_top	(8) All_top	(9) Novelty_top	(10) Scientific_impact_top	(11) Practical_impact_top	(12) All_top	(13) Novelty_top	(14) Scientific_impact_top	(15) Practical_impact_top	(16) All_top
Crowd	−0.039	−0.062	0.075	0.036	0.236**	0.229**	0.359**	0.244**	−0.050	−0.035	0.089	0.049	−0.020	−0.022	0.115	0.051
	(0.060)	(0.059)	(0.060)	(0.049)	(0.065)	(0.062)	(0.060)	(0.055)	(0.074)	(0.074)	(0.076)	(0.063)	(0.072)	(0.073)	(0.075)	(0.064)
Ln_question count									0.011	−0.026	−0.014	−0.013	0.243**	0.240**	0.234**	0.184**
									(0.041)	(0.041)	(0.045)	(0.037)	(0.032)	(0.033)	(0.034)	(0.036)
Constant	0.232**	0.320**	0.293**	0.118**	0.232**	0.320**	0.293**	0.118**	0.232**	0.320**	0.293**	0.118**	0.232**	0.320**	0.293**	0.118**
	(0.051)	(0.048)	(0.046)	(0.039)	(0.051)	(0.048)	(0.046)	(0.039)	(0.051)	(0.048)	(0.046)	(0.039)	(0.051)	(0.048)	(0.046)	(0.039)
N	219	252	254	218	223	252	254	220	219	252	254	218	223	252	254	220
R^2	0.002	0.004	0.006	0.002	0.050	0.050	0.122	0.062	0.002	0.006	0.006	0.003	0.191	0.168	0.233	0.158

Notes: Linear probability models, robust standard errors in parentheses. ** $p < 0.01$. Professional research questions as benchmark.
The "Random pick" sample includes one randomly drawn question from each crowd contributor. The "Best pick" sample includes the best question from each crowd contributor on a particular dimension. Top-rated questions are those that received ratings of >=4 on a particular dimension. "All_top" questions received ratings of >=4 on all three quality dimensions.

[27] For professional research questions, *questioncount* is set to 1.

[28] We feature linear probability model regressions for ease of interpretation (Angrist and Pischke, 2008). Robustness checks using logistic regressions show very similar results (available upon request).

[29] Crowd members who submitted multiple questions generally did so in a single session, but the data received from the organizers do not allow us to see the exact sequence in which RQs were submitted. As such, we cannot investigate time trends with respect to question quality or topic flow. Our reading of the submissions suggests that most of the prolific contributors covered a small set of topics (e.g., the perception of mentally ill people in society) but explored rather broadly within those topics to look at the issue from different angles. Future research could use methods such as think-aloud protocols to study how exactly individuals generate RQs, and to explore potential individual heterogeneity in cognitive processes.

Table 6
Project 2: Average and top-rated crowdsourced vs. professional research questions.

	All crowd-generated RQs vs. professional RQs							Best 20% crowd-generated RQs (per dimension) vs. professional RQs						
	(1) Novelty	(2) Scientific_impact	(3) Practical_impact	(4) Novelty_top	(5) Scientific_impact_top	(6) Practical_impact_top	(7) All_top	(8) Novelty	(9) Scientific_impact	(10) Practical_impact	(11) Novelty_top	(12) Scientific_impact_top	(13) Practical_impact_top	(14) All_top
Crowd	−1.079**	−0.979**	−0.554**	−0.383**	−0.539**	−0.389**	−0.301**	−0.044	0.076	0.360**	−0.241*	−0.046	0.294**	−0.023
	(0.124)	(0.105)	(0.089)	(0.065)	(0.064)	(0.068)	(0.061)	(0.140)	(0.127)	(0.085)	(0.097)	(0.108)	(0.056)	(0.159)
Constant	0.899**	0.759**	0.543**	0.419**	0.667**	0.706**	0.323**	0.899**	0.759**	0.543**	0.419**	0.667**	0.706**	0.323**
	(0.111)	(0.088)	(0.068)	(0.063)	(0.057)	(0.056)	(0.060)	(0.112)	(0.089)	(0.068)	(0.063)	(0.057)	(0.056)	(0.060)
N	201	210	207	201	210	207	200	90	98	97	90	98	97	72
R^2	0.324	0.306	0.143	0.240	0.302	0.135	0.190	0.001	0.003	0.102	0.055	0.002	0.111	0.000

Notes: OLS Regressions, robust standard errors in parentheses. ** $p < 0.01$, * $p < 0.05$. Professional research questions as benchmark.
Top-rated questions are those that received ratings in the upper 40% of the ratings scale. "All_top" questions received a top score on all three quality dimensions.

was the best question (similar to the best-pick sample in project 1). The project organizers did not instruct participants to pursue one or the other approach. Table 6 compares the crowd submissions to professional research questions. Models 1–3 show that crowd questions have significantly lower average novelty, scientific impact, and practical impact than professional questions. This pattern of lower quality persists when we examine the likelihood of generating "top" questions (models 4–7).

A potential interpretation of the lower quality of crowd-generated RQs in project 2 is that contributors submitted the first question that came to mind rather than considering multiple options and submitting the best one.[30] Given that each contributor only submitted one question, we cannot examine whether picking the "best" of multiple ideas from an individual provides an advantage to crowd contributors (we could do so for project 1; see Table 4, models 4–6). However, we can explore a similar idea by selecting the best questions across all contributors. Towards this end, we pick the best 20% of crowd-generated RQs on each dimension (a share similar to the share of "best pick" RQs in project 1) and compare them to professional RQs. Models 8–10 in Table 6 show that the best crowd-generated questions have similar novelty and scientific impact as professional questions. They outperform significantly with respect to practical impact. Finally, we compare the best crowd-generated questions to professional questions with respect to the likelihood of receiving a "top" rating. Models 11–14 show that even the best crowd questions are less likely to be rated as "top" with respect to novelty, but they are significantly more likely to be rated as "top" than professional questions with respect to practical impact.

Summarizing results from projects 1 and 2, the average crowd contribution tends to be rated as lower quality than professional questions obtained from conference proceedings with respect to novelty and scientific impact. In project 2, crowd questions also have lower ratings for practical impact. However, crowd questions outperform professional RQs when a selection process is applied either by picking the best of multiple submissions by an individual crowd contributor (project 1) or when picking the best questions from all crowd submissions (project 2). Fig. 3 visualizes these results.[31]

5. Discussion

An increasing number of research projects actively involve citizens who are not professional scientists, but this involvement is often limited to empirical work such as data collection and analysis. In this paper, we explored whether and how the "crowd", understood as a large number of individuals self-selecting in response to an open call for participation based on their knowledge, skills or experiences, can be involved in formulating research questions. Both our conceptual discussion and the

[30] Further supporting this interpretation is that most contributors in project 2 were recruited through a general-purpose crowdsourcing platform and worked for fixed pay, without explicit incentive to submit high quality RQs.

[31] Sample size limitations and incomplete data on crowd members' background prevent us from systematically examining whether contributors' characteristics are correlated with the quality of RQs. In unreported supplementary analyses, however, we used the available data to compare RQs generated by medical practitioners to those generated by other crowd members (e.g., patients and relatives). In project 1, the average quality of research questions is not significantly different, but medical practitioners are more likely to generate RQs that are top ranked with respect to novelty. In project 2, medical practitioners generated RQs that are more novel on average but we find no differences with respect to other dimensions of quality, or the likelihood of generating top rated questions. Thus, preliminary evidence suggests that medical practitioners generate somewhat more novel questions, which may reflect that they have greater knowledge of the state of the literature but also that they have richer (vicarious) problem-related knowledge from having worked with a larger number of patients. Both the results and our interpretation should be considered extremely preliminary, but they suggest interesting avenues for future research on heterogeneity within the crowd.

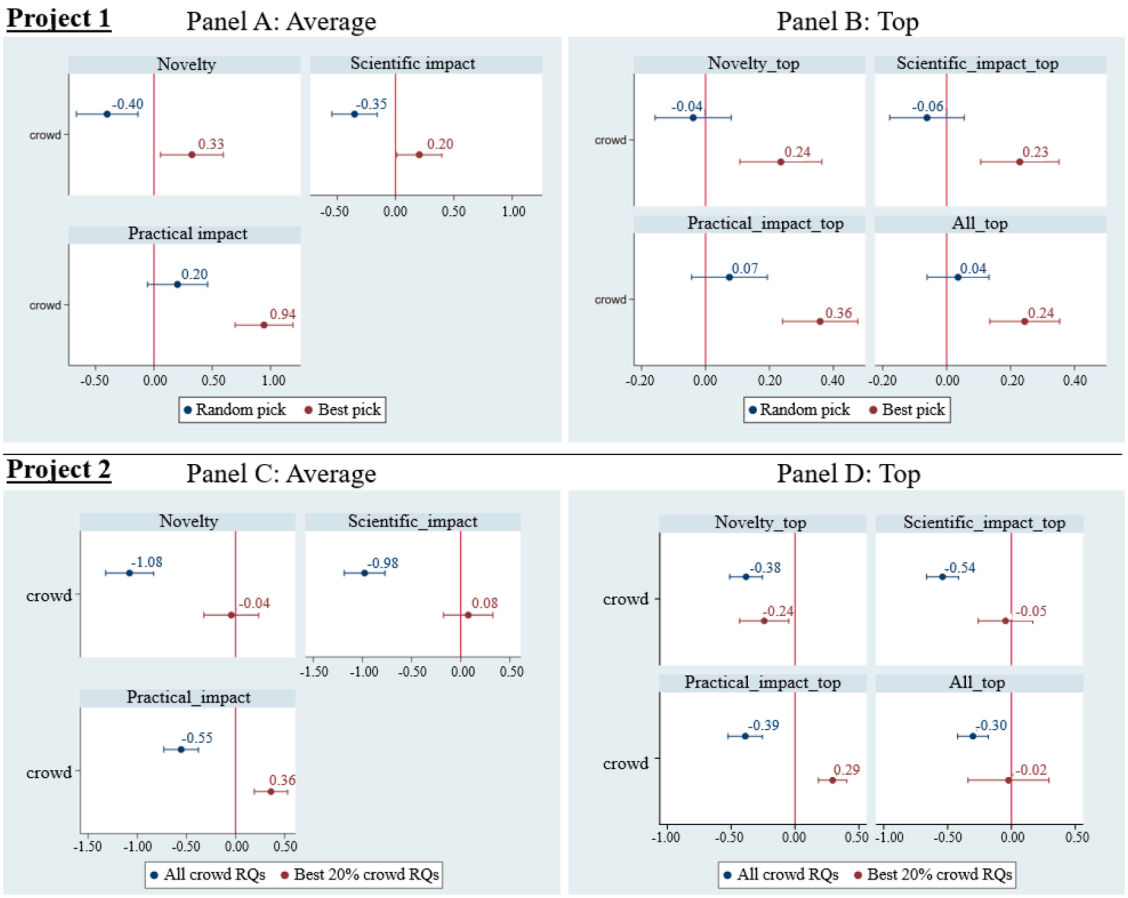

Fig. 3. Crowd versus professional questions: Parameter estimates with 95% confidence intervals.
Notes: Figure shows estimated coefficients of the dummy variable for crowd-generated RQs (relative to professional RQs). Panel A visualizes results equivalent to Table 4 (models 1–6), using quality dimensions as dependent variables but re-estimated using OLS for ease of interpretation. Panel B visualizes results from Table 5 (models 1–8), using the dummies whether a question was rated "top" on a particular dimension as dependent variables. Panel C visualizes results from Table 6 (models 1–3 and 8–10), using quality dimensions as dependent variables. Panel D visualizes results from Table 6 (models 4–7 and models 11–14), using the dummies whether a question was rated "top" on a particular dimension as dependent variables. In project 1, the "random pick" sample of crowd-generated questions includes one randomly drawn question from each contributor. The "best pick" sample includes the best question on a particular dimension from each contributor. In project 2, the "Best 20% crowd RQs" sample includes the best 20% of all crowd-generated RQs on the respective quality dimension.

empirical results suggest that some crowd members can generate high-quality research questions, even when judged by experts and relative to questions generated in the conventional professional system. Indeed, the institution running the crowdsourcing projects found the outcomes sufficiently promising to fund two research groups to investigate crowd-generated RQs on mental illness, and a third group to investigate crowd-generated RQs in traumatology.[32]

However, the data also show more nuanced patterns that point to potential limitations of crowdsourcing and the need for thoughtful project design. Among others, many submissions were ill-structured questions that draw attention to potentially important problems and may help set research priorities at a more general level but provide little guidance regarding which particular causes or solutions should be studied. Moreover, the average novelty and scientific impact of crowd-generated RQs was lower than that of RQs generated though the conventional – more complex – scientific process. Consistent with many other applications of crowdsourcing, the strength of the crowd emerged primarily when we moved beyond averages and applied selection mechanisms that focused attention on the smaller share (but still considerable number) of high-quality contributions. Attention to the "best" crowd-generated questions, however, does not necessarily create a bias in favor of the crowd because the professional RQs that made their way into published conference proceedings are likely also a selected subset of those questions initially conceived by professional scientists.

Our conceptual discussion identified five different "paradigms" that highlight different rationales for involving crowd members (see Table 1). We now build on that discussion to interpret our empirical observations and to note some promising areas for future research as seen through the lens of each paradigm.

Several of our observations can be interpreted through the lens of *crowd labor*. First, participants invested a considerable amount of effort and generated a large volume of contributions. This was the case especially in project 1, which did not constrain contributors regarding the volume of text submitted. Consistent with prior evidence from crowd labor platforms such as Zooniverse, this project also showed a skewed distribution of contributions across individuals, with a small number of prolific contributors making a large share of total contributions (Sauermann and Franzoni, 2015). Although the best research questions were more likely to come from such prolific contributors, this did not reflect generally higher quality submitted by these individuals. Rather, it appears that these individuals had more "trials", some of which resulted in

[32] The institution applied an internal evaluation process that was independent from the evaluations used in this study. Among others, that process involved clustering and combining research questions, and the institution considered a broad range of factors in deciding which projects to fund. Further details are available at https://ois.lbg.ac.at/en/projects/crowdsourcing-research-questions-in-science.

high quality contributions. This pattern is consistent with theories of scientific creativity as a stochastic process, where the quality of the best ideas is a function of the number of ideas that are generated (Simonton, 2003). An important question for future research derived from this perspective is how the project design can be improved to shift the distribution of contributions to increase overall quality. Borrowing from scientific projects that use crowd labor for empirical activities, for example, organizers could consider more extensive training, gamification, or technical tools that facilitate work and guide contributors towards formulating questions that meet certain pre-established criteria (e.g., being well-structured).

Broadcast search focusses our attention on a smaller number of outlier contributions and suggests that such contributions may come from individuals who have exceptional relevant knowledge. Although several RQs had very high quality (see Fig. 2 for examples), we observed no questions that one would think as orders-of-magnitude better than the average question. Looking at quality ratings alone, this may simply reflect that the rating scales were bounded at the top, creating ceiling effects. More fundamentally, this result raises the intriguing question of whether the quality distribution of *problems* identified by the crowd may be naturally less skewed than the quality distribution of crowd-generated *solutions* to important problems. After all, many people will be able to recognize cancer as a key problem to solve but only very few people will have breakthrough ideas on how to solve that problem. Unfortunately, the data limited our ability to study a second aspect highlighted by the broadcast search paradigm, namely the characteristics of those individuals who make high-value contributions. Future research should study how RQ quality is related to individual characteristics, and what types of crowds project organizers should target to increase the chance of finding high value contributions.

The *user innovation* paradigm is useful to understand several of our findings. First, many crowd-generated RQs crossed disciplinary silos, often linking medical to non-medical fields (see Section 4.1). This could reflect that experience as medical practitioner or patient gives crowd members access to unique knowledge that provides a holistic understanding of the nature of problems and suggests a broad range of causes and solutions. Similarly, we found that crowd-generated RQs outperformed professional questions with respect to practical impact, consistent with the idea that experiential knowledge allows users to identify important problems and solutions that may be less salient to experts (Poetz and Schreier, 2012). At the same time, the novelty and scientific impact of crowd-generated RQs tended to be lower, perhaps reflecting a lack of scientific expertise that is particularly relevant for those dimensions of quality (see Section 2.4.2). Relatedly, we observed higher RQ quality across all three dimensions in project 1 versus project 2. Although the projects differed in several respects (see Appendix A), one potential explanation is that project 1 included a larger share of individuals with deep expertise in the problem domain ("experts by experience"), including patients and medical practitioners. In addition, mental illnesses are often more chronic and complex than accidental injuries (e.g., with respect to socio-economic implications), potentially creating more extensive experiences among patients and caregivers (e.g., Holmes and Deb, 2003; Rüsch et al., 2005). Although these interpretations are tentative, they point to the value of considering different types of knowledge, including expert knowledge held primarily by scientific researchers but also "experiential" knowledge held by individuals who are confronted with practical problems on a regular basis. Future research is needed to tie different dimensions of RQ quality more directly to explicit measures of such prior knowledge. Building on prior research on the value of "toolkits" (Von Hippel and Katz, 2002), future research could also examine how crowd members can be supplied with the scientific knowledge that they may currently lack in order to formulate RQs that have higher potential practical *and* scientific impact.

The *community production* paradigm is less relevant in our context because neither of the projects allowed contributors to interact or share knowledge with each other. On the one hand, the project design focusing on individual contributions may have made it more challenging for crowd members to come up with good questions. Recall that the professional questions we included for comparison were predominantly created by teams of scientists, which may be part of the reason for their higher quality on some dimensions (Singh and Fleming, 2010; Wuchty et al., 2007). On the other hand, collaboration often requires a greater time commitment from contributors and creates coordination challenges (Cummings and Kiesler, 2014; Dahlander and O'Mahony, 2010), raising the possibility that a collaborative approach would have attracted fewer contributors and generated fewer research questions. Moreover, there are potential downsides of collaboration such as group-think and social influence that may be particularly problematic in creative tasks (Knudsen and Srikanth, 2014; Paulus, 2000) as well as in tasks involving estimation (see the benefits of independent judgments highlighted by the crowd wisdom paradigm). In the particular context of medicine, collaborative approaches may also create additional participation challenges related to social stigma associated with particular conditions, or concerns about anonymity and confidentiality.

To understand the potential benefits and costs of collaboration in RQ formulation, future work could study crowdsourcing mechanisms that allow crowd members to interact with each other, but also how professional scientists perform when asked to generate RQs as individuals rather than teams. Even more interestingly, the user innovation and the community production paradigms jointly would suggest that collaboration between professional scientists and citizens might be an effective mechanism to integrate expert and experiential knowledge in RQ formulation. One approach to do so is co-creation of research projects by citizens and professional scientists working side-by-side. Of course, such collaborations are not trivial given the need to develop a shared language and address perceived and real differences in norms, status, and incentives (Hidalgo et al., 2021; Suess-Reyes et al., 2021). An alternative might be sequential approaches, where crowds draw on their experiential knowledge to create a pool of potential research questions, and experts subsequently use their scientific knowledge to select the most promising RQs and develop them further (Guinan et al., 2013).

Finally, we noted in Section 2 that the *crowd wisdom* paradigm has limited applicability to our empirical context because it primarily focuses on accuracy in estimation tasks. However, follow-up research could study whether crowds could be useful for *evaluating* crowd-generated RQs, and whether the reduction of errors and biases due to aggregation is one of the main benefits. Indeed, recent work suggests that wisdom of crowd effects apply not only to estimates of facts but also to estimates of preferences (Müller-Trede et al., 2018), which may be particularly important when assessing the potential social impact of research (Franzoni et al., 2021). Future research on evaluation is particularly relevant given the high number of questions that can be generated through crowdsourcing, requiring efficient mechanisms to screen and evaluate contributions (Piezunka and Dahlander, 2015).

We note again that the degree to which the different paradigms may explain our findings also depends on the particular design of the projects we studied. As such, the foregoing discussion does not imply that any particular paradigm is generally "better" or more useful to study crowdsourcing of RQs than another. We suspect that the most effective crowdsourcing approaches combine and customize mechanisms highlighted by several paradigms, and future research on such an integration may be particularly valuable from an applied perspective.

Before we conclude, we acknowledge important limitations that provide additional opportunities for future research. First, we studied projects in the medical sciences and it is not clear how much our results generalize to other fields. The medical sciences may be particularly amenable to crowdsourcing RQs to the extent that the general public can relate more easily to scientific research than in other fields, and that many members of the general public have relevant experiential knowledge as patients, relatives, or health practitioners. It would be important to examine how crowdsourcing of RQs performs in other fields, especially more "basic" fields such as genomics or quantum physics. Indeed,

it would be exciting to study whether the benefits of crowdsourcing, and the relevance of different crowdsourcing paradigms, depend on the "distance" between professional scientists and the general public in terms of relevant scientific expert and experiential knowledge.

Notwithstanding potential limitations with respect to generalizability, our context of the medical sciences allows us to add to a vibrant literature on the interactions between professionals and citizens in this important field. In particular, both scholars and policymakers have called for involving patients and caregivers along the entire process of clinical research (e.g., Patrick-Lake and Hernandez, 2017), and have highlighted medical practitioners as contributors who may have aggregated valuable experiential knowledge through their frequent interactions with patients (e.g., Collins et al., 2017; Forsythe et al., 2019). Our paper documents an ambitious effort to involve patients as well as medical practitioners in early stages of the research process, using a large-scale crowdsourcing mechanism that is quite different from more common personal and smaller-scale efforts to involve citizens (Kaiser, 2012). At the same time, we complement prior research that has documented a pioneering crowdsourcing effort in the medical sciences (Guinan et al., 2013) by providing a conceptual discussion of relevant crowd paradigms, quantitative evidence on project outcomes, and an explicit comparison between crowd contributions and professional RQs.

A second potential limitation is that we measured RQ quality using evaluations by expert scientists. This approach is consistent with prior research on crowdsourcing performance in science and other domains (Guinan et al., 2013; Jeppesen and Lakhani, 2010; Poetz and Schreier, 2012) and is also appropriate given the empirical context, where the research institution sought to identify research questions that could be investigated by professional scientists. However, expert evaluators may have their own sets of biases and may fail to fully appreciate the quality of crowd-generated research questions, especially if those questions deviate from the style or language scientific experts are used to. In that sense, the high quality ratings that (selected) RQs received from experts in the current study are all the more remarkable and support those CS advocates who believe that the crowd can make important contributions *and* that professional scientists will appreciate the resulting advantages and are willing to work with the crowd. At the same time, future research could investigate whether and how quality evaluations of other stakeholders such as citizens and policymakers differ from those of experts. Such work could consider differences and even disagreements regarding standards and values between experts and the broader public (Cohen and Doubleday, 2021; Ottinger, 2010) but could also explore unique strengths of crowd evaluations as suggested by the "wisdom of crowds" paradigm. A related interesting question is how the willingness to "listen" to the crowd differs across professional scientists and how crowdsourcing processes can be set up to reduce tensions and facilitate effective collaboration between professionals and citizens.

Third, we focused on the potential benefits of crowdsourcing and paid less attention to challenges and costs. Both of the projects we studied required considerable resources to set up project infrastructure, recruit participants, and evaluate submissions. As such, future work should study how projects can be improved to increase both effectiveness and efficiency. Relatedly, our theorizing and empirical analysis focused on the "productivity" benefits of crowdsourcing, as measured by the quality of crowd-generated RQs.[33] Policymakers and CS advocates argue that involving the general public in RQ formulation and agenda setting may also have a number of other positive outcomes such as citizen learning, increased trust in science, and greater adoption of scientific advances in the public (European Science Foundation, 2013; Sauermann et al., 2020). Future research should study a broader range of project outcomes, whether there are trade-offs between different project goals, and how such trade-offs can be mitigated.

Finally, we share with much of the literature on crowdsourcing and citizen science the premise that engaging crowds and citizens in research can yield important scientific as well as broader benefits, notwithstanding potential costs and challenges. Scholars in other fields – especially Science and Technology Studies – provide more critical perspectives that emerge especially when crowd science is seen in the broader socio-economic context (for a review, see Kimura and Kinchy, 2016). Among others, there are concerns that unpaid contributions from crowd members support and facilitate "neoliberalist" tendencies, including funding cuts to academic science. Similarly, projects under the leadership of professional scientists may be "neocolonialist" in that experts rather than citizens decide which types of knowledge and methods are accepted as legitimate (see also our earlier discussion regarding quality evaluations by different stakeholders). Other observers are concerned that engaging citizens in research may support the "scientization" of public debates: credentialed science is privileged as the only legitimate basis for public debate, to the detriment of political discourse and attention to distributional and socio-political issues. Finally, some organizations may involve citizens only for the purpose of improving their public image rather than trying to generate scientific or non-scientific outcomes that are of broader value (Blacker et al., 2021). Although these issues are beyond the scope of the present paper, we see intriguing opportunities for integration in future research. In particular, many of these concerns have emerged from research on projects that use citizens for data collection and it would be interesting to study how these concerns apply when crowds and citizens get engaged in other stages of the research process, especially agenda setting and the formulation of research questions. Relatedly, one might explore whether these concerns apply to the same extent to the different "crowd paradigms" discussed in Section 2.4.2, e.g., crowdsourcing to procure large volumes of labor versus crowdsourcing to identify creative outlier solutions to problems or to tap into the experiential knowledge of users. And of course, future research should study what tools and policies can be designed to harness the various benefits of crowd involvement while ensuring that citizens are engaged in a transparent and responsible way (Cohen and Doubleday, 2021).

To conclude, our conceptual discussion identified several mechanisms that suggest benefits from involving crowds in research question formulation. Our empirical results show that crowdsourcing RQs can indeed "work", while also providing deeper insights into the nature of crowd-generated RQs as well as their strengths and weaknesses relative to questions developed though the conventional scientific process. We already discussed in Section 1 contributions of this study to the literatures on crowd science, crowdsourcing, and the organization of science more generally. Although it would be premature to make practical recommendations, we hope that policymakers and science practitioners will also find this study helpful in thinking about the potential merits of involving members of the broader public in research generally, and in formulating questions for research in particular.

CRediT authorship contribution statement

Susanne Beck: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Tiare-Maria Brasseur:** Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Marion Poetz:** Conceptualization, Methodology, Validation, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Henry Sauermann:** Conceptualization, Methodology, Validation,

[33] Going beyond the quality of the research questions per se, another potential "productivity" benefit of involving citizens in RQ formulation is that the resulting projects may be of greater interest and relevance to citizens, potentially increasing their willingness to participate in other stages of the research such as data collection or clinical trials. Such benefits have been documented, among others, in the context of community-based participatory research (Bhardwaj et al., 2019; Sofolahan-Oladeinde et al., 2015).

Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Two of the authors are employed as researchers at LBG, working in a research unit that had been given the opportunity to collect and analyze data on the crowdsourcing projects. LBG had no influence on the design or implementation of the present study.

Acknowledgments

We are grateful for extremely helpful feedback at several conferences and seminars, including the AoM Annual Meeting, CBS Department of Strategy and Innovation Seminar, DRUID Annual Conference, Duke Workshop on Field Experiments in Strategy, Innovation and Entrepreneurship, ESMT Berlin Faculty Seminar, Johns Hopkins Management & Organizations Seminar, Kellogg School of Management Innovation Seminar, Max Planck RISE Workshop, Open Innovation in Science Research Conference, R&D Management Conference, SEI Faculty Consortium, and the Workshop on the Organisation, Economics, and Policy of Scientific Research. We would also like to thank the editor and two anonymous reviewers for their very constructive comments and suggestions. We thank the Ludwig Boltzmann Gesellschaft for providing data, expertise and background information on their two crowdsourcing projects. This study was funded by the Austrian National Foundation for Research, Technology and Development, grant for Open Innovation in Science. All errors remain ours.

Appendix A. : Project 1 and 2 – similarities and differences

	Project 1 (Mental illness)	Project 2 (Traumatology)
Organizer	A large research institution with a focus on the medical sciences, located in Europe	
Project goal	Generation of research questions as basis for initiating future research projects. "What questions about mental illness/traumatology should research address"?	
Project promise/ commitment to participants	All RQs will be openly shared (anonymously), so that scientists from any kind of institution and field worldwide can access and potentially further pursue them. At least one question (or group of questions) is selected through an evaluation process initiated by the institution to create and fund a new research group within the institution.	
Registration	Registration (incl. demographic data) was voluntary. Contributors remained anonymous but could opt-in to be informed about project results and subsequent initiatives of the institution.	Registration (incl. demographic data) was mandatory. Contributors remained anonymous but could opt-in to be informed about project results and subsequent initiatives of the institution.
Submission platform	Custom-designed project website, available in English and German. Participants could choose their preferred language as part of the registration.	
Elicitation of contributions	Large text field (unlimited space). Focus on RQs but crowds were able to write about other things such as their experiences, problems, reasons why they want scientists to research this question.	Text-field with a single line. Focus on eliciting a single research question per contributor.
Recruiting strategy	Offline and online outreach, e.g., via patient organizations and organizations of medical practitioners.	Offline and online outreach, e.g., via patient organizations and organizations of medical practitioners. In addition, the general-purpose platform Clickworker was used to crowdsource RQs for a small monetary reward (3 €)
Crowdsourcing period	April - July 2015	May-Aug 2018, Oct 2018 (Clickworker)

Appendix B. Descriptive statistics of crowd contributions and contributors from project 1 and project 2

Tables B1 and B2

Table B1
Crowd contributions.

	Project 1 "Mental Illness"	Project 2 "Traumatology"
Crowd "raw" contributions	422	180
Contributions from which research question(s) could be extracted/clear contributions	282	151
Crowd generated RQs	753	151
RQs not related to the research field (excluded)	7	4
Total sample of crowd research questions (excl. duplicates and invalid)	**746**	**147**
of which: submitted as a question	401	147
of which: extracted from question-like statement	345	–

Table B2
Characteristics and experience of crowd contributors.

	Project 1: "Mental Illness" ($N = 155$)			Project 2 "Trauma" ($N = 147$)		
Variable	mean	min	max	mean	min	max
Total crowd contributors						
Crowd characteristics						
Age	45.42	21	76	36.70	18	65
Female	.63	0	1	.46	0	1
Recruited through general-purpose crowdsourcing platform	–	–	–	.83	0	1
Crowd experience						
No experience	–	–	–	.05	0	1
Experience as patient or patient relative	.14	0	1	.76	0	1
Experience as medical practitioners (no patient/relative experience)	.38	0	1	.01	0	1
Experience as medical practitioners and patient/relative	.12	0	1	.18	0	1

Notes: In project 1, the provision of personal data was optional, leading to a higher rate of missing data.

Appendix C. Descriptive statistics of research question evaluations from project 1 and project 2

	Project 1: "Mental illness" ($N = 746$)				Project 2: "Traumatology" ($N = 147$)			
	mean	sd	min	max	mean	sd	min	max
Ratings (not standardized)								
Well-structured	.40		0	1	.42		0	1
Questions with structure:								
Problem & cause	.19		0	1	.24		0	1
Problem & solution	.12		0	1	.14		0	1
Relationship	.09		0	1	.01		0	1
Can't assess (=missing)	.01		0	1	.02		0	1
Discipline-crossing	.44		0	1	.54		0	1
Discipline-crossing with:								
Medical fields	.10		0	1	.03		0	1
Non-medical fields	.34		0	1	.50		0	1
Specificity	3.91	.96	1	5	2.69	1.33	1	5
RQ length (characters)	136.29	75.58	31	1172	93.36	42.63	27	298
Questions per contributor (questioncount)	4.81	8.60	1	86	1			
Novelty	2.69	1.06	1	5	1.94	.61	1	3.75
Scientific impact	2.90	.87	1	5	2.20	.80	1	4.67
Practical impact	3.30	1.10	1	5	3.01	.80	1.25	4.75
Novelty _top	.26	.44	0	1	–	–	–	–
Scientific_impact_top	.26	.44	0	1	–	–	–	–
Practical_impact_top	.46	.50	0	1	–	–	–	–
All_top	.16	.37	0	1	–	–	–	–
Rater-standardized ratings								
Novelty	–	–	–	–	−0.18	.64	−1.32	1.92
Scientific_impact	–	–	–	–	−0.22	.68	−1.36	1.76
Practical_impact	–	–	–	–	−0.01	.68	−1.61	1.52
Novelty _top	–	–	–	–	.13	.34	0	1
Scientific_impact_top	–	–	–	–	.13	.33	0	1
Practical_impact_top	–	–	–	–	.35	.48	0	1
All_top	–	–	–	–	.03	.19	0	1

Appendix D. : Pairwise correlations (all crowd-generated research questions)

A: Project 1

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1) Well-structured	1.00											
(2) Specificity	0.18*	1.00										
(3) Discipline-crossing	0.08*	−0.00	1.00									
(4) RQ length	0.15*	0.08*	0.02	1.00								
(5) ln_questioncount	0.02	−0.07	−0.07*	−0.01	1.00							
(6) Novelty	0.14*	−0.08*	0.09*	0.13*	0.12*	1.00						
(7) Scientific impact	0.15*	−0.09*	0.10*	0.12*	0.06	0.69*	1.00					
(8) Practical impact	0.02	−0.06	0.09*	0.09*	0.10*	0.68*	0.58*	1.00				
(9) Novelty_top	0.10*	−0.05	0.06	0.08*	0.09*	0.81*	0.52*	0.58*	1.00			
(10) Scientific_impact_top	0.10*	−0.08*	0.09*	0.12*	0.05	0.60*	0.79*	0.46*	0.56*	1.00		
(11) Practical_impact_top	0.07*	−0.03	0.07	0.06	0.09*	0.64*	0.54*	0.85*	0.52*	0.44*	1.00	
(12) All_top	0.07	−0.04	0.07	0.09*	0.02	0.61*	0.58*	0.53*	0.74*	0.72*	0.47*	1.00

Note: * $p<0.05$; $N = 746$

B: Project 2

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1) Well-structured	1.00										
(2) Specificity	0.45*	1.00									
(3) Discipline-crossing	0.43*	0.04	1.00								
(4) RQ length	0.43*	0.43*	0.21*	1.00							
(5) Novelty	0.35*	0.30*	0.11	0.36*	1.00						
(6) Scientific_impact	0.05	0.01	−0.13	0.19*	0.61*	1.00					
(7) Practical_impact	−0.07	−0.09	−0.02	0.20*	0.48*	0.74*	1.00				
(8) Novelty_top	0.15	0.13	0.15	0.10	0.56*	0.38*	0.25*	1.00			
(9) Scientific_impact_top	−0.10	−0.09	−0.19*	0.06	0.31*	0.76*	0.49*	0.27*	1.00		
(10) Practical_impact_top	−0.10	−0.09	−0.13	0.12	0.39*	0.60*	0.77*	0.20*	0.43*	1.00	
(11) All_top	0.08	0.05	0.12	0.09	0.41*	0.40*	0.26*	0.77*	0.38*	0.22*	1.00

Note: $^{*}p < 0.05$; $N = 147$

References

Afuah, A., Tucci, C.L., 2012. Crowdsourcing as a solution to distant search. Acad. Manag. Rev. 37, 355–375.

Alvesson, M., Sandberg, J., 2011. Generating research questions through problematization. Acad. Manag. Rev. 36, 247–271.

Amabile, T., 1996. Creativity in Context. Westview Press, Boulder, Colo.

Angrist, J.D., Pischke, J.-S., 2008. Mostly Harmless econometrics: An empiricist's Companion. Princeton University Press.

Beck, S., Bergenholtz, C., Bogers, M., Brasseur, T.M., Conradsen, M.L., Di Marco, D., Distel, A.P., Dobusch, L., Dörler, D., Effert, A., Fecher, B., Filiou, D., Frederiksen, L., Gillier, T., Grimpe, C., Gruber, M., Haeussler, C., Heigl, F., Hoisl, K., Hyslop, K., Kokshagina, O., LaFlamme, M., Lawson, C., Lifshitz-Assaf, H., Lukas, W., Nordberg, M., Norn, M.T., Poetz, M.K., Ponti, M., Pruschak, G., Pujol Priego, L., Radziwon, A., Rafner, J., Romanova, G., Ruser, A., Sauermann, H., Shah, S.K., Sherson, J.F., Suess-Reyes, J., Tucci, C.L., Tuertscher, P., Vedel, J.B., Velden, T., Verganti, R., Wareham, J., Wiggins, A., Xu, S.M., 2022. The open innovation in science research field: a collaborative conceptualisation approach. Ind. Innov. 29 (2), 1–50.

Bhardwaj, P., Kumar, J., Yadav, R.K., 2019. Patients driving the clinical trial designs–democracy in clinical research. Rev. Recent Clin. Trials 14, 237–246.

Blacker, S., Kimura, A.H., Kinchy, A., 2021. When citizen science is public relations. Soc. Stud. Sci. 51 (5), 780–796.

Bonney, R., Shirk, J.L., Phillips, T.B., Wiggins, A., Ballard, H.L., Miller-Rushing, A.J., Parrish, J.K., 2014. Next steps for citizen science. Science 343, 1436–1437.

Borch, C., 2012. The Politics of Crowds: An Alternative History of Sociology. Cambridge University Press.

Boudreau, K.J., Guinan, E.C., Lakhani, K.R., Riedl, C., 2016. Looking across and looking beyond the knowledge frontier: intellectual distance, novelty, and resource allocation in science. Manag. Sci. 62, 2765–2783.

Boudreau, K.J., Lakhani, K.R., 2013. Using the crowd as an innovation partner. Harv. Bus. Rev. 91, 60–69, 140.

Bryman, A., 2007. The research question in social research: what is its role? Int. J. Soc. Res. Methodol. 10, 5–20.

Butticè, V., Franzoni, C., Rossi-Lamastra, C., Rovelli, P., Afuah, A., Tucci, C.L., Viscusi, G., 2017. The road to crowdfunding success: a review of extant literature. Creating and Capturing Value Through Crowdsourcing. Oxford University Press.

Caron-Flinterman, J.F., Broerse, J.E., Bunders, J.F., 2005. The experiential knowledge of patients: a new resource for biomedical research? Soc. Sci. Med. 60, 2575–2584.

Cohen, K., Doubleday, R., 2021. Future Directions For Citizen Science and Public Policy. Centre for Science and Policy, Cambridge.

Collins, S.P., Levy, P.D., Holl, J.L., Butler, J., Khan, Y., Israel, T.L., Fonarow, G.C., Alikhaani, J., Sarno, E., Cook, A., 2017. Incorporating patient and caregiver experiences into cardiovascular clinical trial design. JAMA Cardiol. 2, 1263–1269.

Connolly, T., Routhieaux, R.L., Schneider, S.K., 1993. On the effectiveness of group brainstorming: test of one underlying cognitive mechanism. Small Group Res. 24, 490–503.

Cummings, J.N., Kiesler, S., 2014. Organization theory and the changing nature of science. J. Organ. Des. 3, 1–16.

Cummings, S.R., Browner, W.S., Hulley, S.B., Hulley, S.B., Cummings, S.R., Browner, W. S., Grady, D.G., Newman, T.B., 2007. Conceiving the research question. Designing Clinical Research. Lippincott Williams & Wilkins.

Dahlander, L., O'Mahony, S., 2010. Progressing to the center: coordinating project work. Organ. Sci. 22, 961–979.

Dasgupta, P., David, P.A., 1994. Toward a new economics of science. Res. Policy 23, 487–521.

Demonaco, H., Oliveira, P., Torrance, A., Von Hippel, C., Von Hippel, E., 2019. When Patients Become Innovators. MIT Sloan Management Review, pp. 81–88.

Ding, W., Levin, S., Stephan, P., Winkler, A., 2010. The impact of information technology on academic scientists' productivity and collaboration patterns. Manag. Sci. 56, 1439.

Durand, D.E., VanHuss, S.H., 1992. Creativity software and DSS: cautionary findings. Inf. Manag. 23, 1–6.

Eitzel, M., Cappadonna, J., Santos-Lang, C., Duerr, R., West, S.E., Virapongse, A., Kyba, C., Bowser, A., Cooper, C., Sforzi, A., 2017. Citizen science terminology matters: exploring key terms. Citiz. Sci. Theory Pract. 2, 1–20.

European Commission, 2018. Open Science Policy Platform Recommendations. European Commission.

European Science Foundation, 2013. Science in Society: Caring for our Futures in Turbulent Times. European Science Foundation.

Felin, T., Zenger, T.R., 2014. Closed or open innovation? Problem solving and the governance choice. Res. Policy 43, 914–925.

Fleming, L., Sorenson, O., 2004. Science as a map in technological search. Strateg. Manag. J. 25, 909–928.

Follett, R., Strezov, V., 2015. An analysis of citizen science based research: usage and publication patterns. PLoS One 10, e0143687.

Forsythe, L.P., Carman, K.L., Szydlowski, V., Fayish, L., Davidson, L., Hickam, D.H., Hall, C., Bhat, G., Neu, D., Stewart, L., 2019. Patient engagement in research: early findings from the patient-centered outcomes research institute. Health Aff. 38, 359–367.

Foss, N.j., Frederiksen, L., Rullani, F., 2016. Problem-formulation and problem-solving in self-organized communities: how modes of communication shape project behaviors in the free open-source software community. Strateg. Manag. J. 37, 2589–2610.

Franke, N., Piller, F., 2004. Value creation by toolkits for user innovation and design: the case of the watch market. J. Prod. Innov. Manag. 21, 401–415.

Franke, N., Poetz, M.K., Schreier, M., 2013. Integrating problem solvers from analogous markets in new product ideation. Manag. Sci. 60, 1063–1081.

Franke, N., Shah, S., 2003. How communities support innovative activities: an exploration of assistance and sharing among end-users. Res. Policy 32, 157–178.

Franke, N., Von Hippel, E., Schreier, M., 2006. Finding commercially attractive user innovations: a test of lead-user theory. J. Prod. Innov. Manag. 23, 301–315.

Franzoni, C., Poetz, M., Sauermann, H., 2022. Crowds, citizens, and science: a multi-dimensional framework and agenda for future research. Ind. Innov. 29 (2), 1–34.

Franzoni, C., Sauermann, H., 2014. Crowd Science: the organization of scientific research in open collaborative projects. Res. Policy 43, 1–20.

Franzoni, C., Sauermann, H., Di Marco, D., 2021. When citizens judge science: evaluations of social impact and support for research, Working Paper.

Fujimura, J.H., 1987. Constructing 'Do-Able' problems in cancer research: articulating alignment. Soc. Stud. Sci. 17, 257–293.

Galton, F., 1907. Vox populi. Nature 75, 450–451.

Guinan, E., Boudreau, K.J., Lakhani, K.R., 2013. Experiments in open innovation at Harvard medical school. MIT Sloan Manag. Rev. 54, 45–52.

Gwet, K.L., 2014. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. Advanced Analytics, LLC.

Haeussler, C., Sauermann, H., 2020. Division of labor in collaborative knowledge production: the role of team size and interdisciplinarity. Res. Policy 49, 103987.

Haklay, M.M., Dörler, D., Heigl, F., Manzoni, M., Hecker, S., Vohland, K., Vohland, K., Land-Zandstra, A., Ceccaroni, L., Lemmens, R., Perello, J., Ponti, M., Samson, R., Wagenknecht, K., 2021. What is citizen science? The challenges of definition. The Science of Citizen Science. Springer, pp. 13–33.

Hand, E., 2010. People power. Nature 466, 685–687.

Hecker, S., Garbe, L., Bonn, A., Hecker, S., Haklay, M., Bowser, A., Makuch, Z., Vogel, J., Bonn, A., 2018. The European citizen science landscape – a snapshot. Citizen science: Innovation in Open science, Society and Policy. UCL Press, London, pp. 190–200.

Hidalgo, S.E., Perelló, J., Becker, F., Bonhoure, I., Legris, M., Cigarini, A., Katrin, V., Land-Zandstra, A., Ceccarini, L., Lemmens, R., Perelló, J., Ponti, M., Samson, R., Wagenknecht, K., 2021. Participation and co-creation in citizen science. The Science of Citizen Science. Springer, pp. 199–218.

Holmes, A.M., Deb, P., 2003. The effect of chronic illness on the psychological health of family members. J. Ment. Health Policy Econ. 6, 13–22.

Irwin, A., 2018. No PhDs needed: how citizen science is transforming research. Nature 562, 480–482.

Jeppesen, L.B., Frederiksen, L., 2006. Why do users contribute to firm-hosted user communities? The case of computer-controlled music instruments. Organ. Sci. 17, 45–63.
Jeppesen, L.B., Lakhani, K.R., 2010. Marginality and problem-solving effectiveness in broadcast search. Organ. Sci. 21, 1016–1033.
Jones, B., 2009. The burden of knowledge and the "death of the renaissance man": is innovation getting harder? Rev. Econ. Stud. 76, 283–317.
Kaiser, J., 2012. NIH funding shifts with disease lobbying, study suggests. Science 338, 181.
Khatib, F., DiMaio, F., Cooper, S., Kazmierczyk, M., Gilski, M., Krzywda, S., Zabranska, H., Pichova, I., Thompson, J., Popović, Z., Jaskolski, M., Baker, D., Foldit Contenders Group, Foldit Void Crushers Group, 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nat. Struct. Mol. Biol. 18, 1175–1177.
Kimura, A.H., Kinchy, A., 2016. Citizen science: probing the virtues and contexts of participatory research. Engag. Sci. Technol. Soc. 2, 331–361.
Knorr-Cetina, K., 1999. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, Cambridge, MA.
Knudsen, T., Srikanth, K., 2014. Coordinated exploration: organizing joint search by multiple specialists to overcome mutual confusion and joint myopia. Adm. Sci. Q. 59, 409–441.
Krippendorff, K., 2004. Content Analysis: An introduction to Its Methodology. Sage Thousand Oaks, California.
Kullenberg, C., Kasperowski, D., 2016. What is citizen science? – A scientometric meta-analysis. PLoS One 11, e0147152.
Lamont, M., 2009. How Professors Think. Harvard University Press.
Latour, B., Woolgar, S., 1979. Laboratory Life: The Social Construction of Scientific Facts. Sage, Beverly Hills.
Le Bon, G., 1895. The crowd: A study of the Pupular Mind. Fisher Unwin, London.
Lewandowsky, S., Mann, M.E., Brown, N.J., Friedman, H., 2016. Science and the public: debate, denial, and skepticism. J. Soc. Polit. Psychol. 4, 537–553.
Lifshitz-Assaf, H., 2018. Dismantling knowledge boundaries at NASA: the critical role of professional identity in open innovation. Adm. Sci. Q. 63, 746–782.
Lüthje, C., Herstatt, C., Von Hippel, E., 2005. User-innovators and "local" information: the case of mountain biking. Res. Policy 34, 951–965.
Lyons, E., Zhang, L., 2019. Trade-offs in motivating volunteer effort: experimental evidence on voluntary contributions to science. PLoS One 14, e0224946.
MacCrimmon, K.R., Wagner, C., 1994. Stimulating ideas through creative software. Manag. Sci. 40, 1514–1532.
Majchrzak, A., Malhotra, A., 2020. Unleashing the Crowd: Collaborative Solutions to Wicked Business and Societal Problems. Springer Nature.
Mannes, A.E., Larrick, R.P., Soll, J.B., Krueger, J., 2012. The social psychology of the wisdom of crowds. Frontiers of Social Psychology. Social Judgment and Decision Making Psychology Press, pp. 227–242.
Mannes, A.E., Soll, J.B., Larrick, R.P., 2014. The wisdom of select crowds. J. Pers. Soc. Psychol. 107, 276.
Mazzucato, M., 2018. Mission-oriented innovation policies: challenges and opportunities. Ind. Corp. Chang. 27, 803–815.
Merton, R.K., 1973. The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago.
Mollick, E., Nanda, R., 2016. Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. Manag. Sci. 62, 1533–1553.
Müller-Trede, J., Choshen-Hillel, S., Barneron, M., Yaniv, I., 2018. The wisdom of crowds in matters of taste. Manag. Sci. 64, 1779–1803.
Murray, C.J., Lopez, A.D., 2013. Measuring the global burden of disease. N. Engl. J. Med. 369, 448–457.
National Academies, 2004. Facilitating Interdisciplinary Research. National Academies Press, Washington, DC.
Newman, D.A., 2014. Missing data: five practical guidelines. Organ. Res. Methods 17, 372–411.
Nickerson, J.A., Wuebker, R., Zenger, T., 2017. Problems, theories, and governing the crowd. Strateg. Organ. 15, 275–288.
Nielsen, M., 2011. Reinventing Discovery: The New Era of Networked Science. Princeton University Press.
Ottinger, G., 2010. Buckets of resistance: standards and the effectiveness of citizen science. Sci. Technol. Hum. Values 35, 244–270.
Pammolli, F., Magazzini, L., Riccaboni, M., 2011. The productivity crisis in pharmaceutical R&D. Nat. Rev. Drug Discov. 10, 428.
Patrick-Lake, B., Hernandez, A.F., 2017. When should patients be involved in cardiovascular clinical trial design?: Always, early, and often. JAMA Cardiol. 2, 1269–1270.
Paulus, P., 2000. Groups, teams, and creativity: the creative potential of idea-generating groups. Appl. Psychol. 49, 237–262.
Piezunka, H., Dahlander, L., 2015. Distant search, narrow attention: how crowding alters organizations' filtering of suggestions in crowdsourcing. Acad. Manag. J. 58, 856–880.
Poetz, M.K., Schreier, M., 2012. The value of crowdsourcing: can users really compete with professionals in generating new product ideas? J. Prod. Innov. Manag. 29, 245–256.
Pols, J., 2014. Knowing patients: turning patient knowledge into science. Sci. Technol. Hum. Values 39, 73–97.
Raddick, M.J., Bracey, G., Gay, P.L., Lintott, C., Cardamone, C., Murray, P., Schawinski, K., Szalay, A., Vandenberg, J., 2013. Galaxy zoo: motivations of citizen scientists. Astron. Educ. Rev. 12, 010106.
Raymond, E., 1999. The cathedral and the bazaar. Knowledge. Technology & Policy 12, 23–49.
Rüsch, N., Angermeyer, M.C., Corrigan, P.W., 2005. Mental illness stigma: concepts, consequences, and initiatives to reduce stigma. Eur. Psychiatry 20, 529–539.
Sauermann, H., Franzoni, C., 2015. Crowd science user contribution patterns and their implications. Proc. Natl. Acad. Sci. 112, 679–684.
Sauermann, H., Stephan, P., 2013. Conflicting logics? A multidimensional view of industrial and academic science. Organ. Sci. 24, 889–909.
Sauermann, H., Vohland, K., Antoniou, V., Balázs, B., Göbel, C., Karatzas, K., Mooney, P., Perelló, J., Ponti, M., Samson, R., 2020. Citizen science and sustainability transitions. Res. Policy 49, 103978.
Schwenk, C., Thomas, H., 1983. Formulating the mess: the role of decision aids in problem formulation. Omega 11, 239–252 (Westport).
Science Europe, 2018. Science Europe Briefing Paper on Citizen Science. Science Europe.
Scistarter, 2020. Project finder. https://scistarter.org/.
Shapin, S., 2008. The Scientific Life: A Moral History of a Late Modern Vocation. University of Chicago Press.
Simmons, J.P., Nelson, L.D., Galak, J., Frederick, S., 2011. Intuitive biases in choice versus estimation: implications for the wisdom of crowds. J. Consum. Res. 38, 1–15.
Simon, H.A., 1973. The structure of ill structured problems. Artif. Intell. 4, 181–201.
Simonton, D.K., 2003. Scientific creativity as constrained stochastic behavior: the integration of product, person, and process perspectives. Psychol. Bull. 129, 475–494.
Singh, J., Fleming, L., 2010. Lone inventors as sources of breakthroughs: myth or reality? Manag. Sci. 56, 41–56.
Sofolahan-Oladeinde, Y., Mullins, C.D., Baquet, C.R., 2015. Using community-based participatory research in patient-centered outcomes research to address health disparities in under-represented communities. J. Comp. Eff. Res. 4, 515–523.
Stokes, D., 1997. Pasteur's Quadrant: Basic Science and Technological Innovation. Brookings Institution Press, Washington, DC.
Sturgis, P., Allum, N., 2004. Science in society: re-evaluating the deficit model of public attitudes. Public Underst. Sci. 13, 55–74.
Suess-Reyes, J., Beck, S., Poetz, M., Sauermann, H., 2021. Co-creation in (citizen) science: unbundling the concept and identifying key challenges. In: Proceedings of the 4S Annual Meeting. Online.
Surowiecki, J., 2005. The wisdom of crowds. Anchor books, New York.
Thabane, L., Thomas, T., Ye, C., Paul, J., 2009. Posing the research question: not so simple. Can. J. Anesth. 56, 71–79. Journal canadien d'anesthésie.
Theobald, E.J., Ettinger, A.K., Burgess, H.K., DeBey, L.B., Schmidt, N.R., Froehlich, H.E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M., 2015. Global change and local solutions: tapping the unrealized potential of citizen science for biodiversity research. Biol. Conserv. 181, 236–244.
Tucci, C.L., Afuah, A., Viscusi, G., 2018. Creating and Capturing Value Through Crowdsourcing. Oxford University Press, New York, USA.
Turrini, T., Dörler, D., Richter, A., Heigl, F., Bonn, A., 2018. The threefold potential of environmental citizen science-Generating knowledge, creating learning opportunities and enabling civic participation. Biol. Conserv. 225, 176–186.
US Congress, 2016. Crowdsourcing and Citizen Science Act of 2016. 1–12. Accessible here: https://www.govinfo.gov/content/pkg/BILLS-114hr6414ih/pdf/BILLS-114hr6414ih.pdf.
Van Brussel, S., Huyse, H., 2018. Citizen science on speed? Realising the triple objective of scientific rigour, policy influence and deep citizen engagement in a large-scale citizen science project on ambient air quality in antwerp. J. Environ. Plan. Manag. 62, 1–18.
Von Hippel, E., Katz, R., 2002. Shifting innovation to users via toolkits. Manag. Sci. 48, 821–833.
Von Hippel, E., Von Krogh, G., 2016. Crossroads—identifying viable "need–solution pairs": problem solving without problem formulation. Organ. Sci. 27, 207–221.
Von Krogh, G., Spaeth, S., Lakhani, K.R., 2003. Community, joining, and specialization in open source software innovation: a case study. Res. Policy 32, 1217–1241.
Wiggins, A., Crowston, K., 2011. From conservation to crowdsourcing: a typology of citizen science. In: Proceedings of the 44th Hawaii International Conference on Systems Sciences (HICSS). IEEE, Hawaii, pp. 1–10.
Wuchty, S., Jones, B., Uzzi, B., 2007. The increasing dominance of teams in the production of knowledge. Science 316, 1036–1039.