RUNNING HEAD: CROWDSOURCING SCIENCE

Scientific Utopia III: Crowdsourcing Science

Eric Luis Uhlmann[4]
Charles R. Ebersole[1]
Christopher R. Chartier[6]
Timothy M. Errington[2]
Mallory Kidwell[9]
Calvin K. Lai[8]
Randy J. McCarthy[7]
Amy Riegelman[5]
Raphael Silberzahn[3]
Brian A. Nosek[1,2]

[1] University of Virginia, [2] Center for Open Science, [3] University of Sussex, [4] INSEAD, [5] University of Minnesota, [6] Ashland University, [7] Northern Illinois University, [8] Washington University in St. Louis, [9] University of Utah

Authors' Note: This research was supported by the Center for Open Science and an R&D grant from INSEAD. Please address correspondence to Eric Uhlmann, eric.luis.uhlmann@gmail.com, or Brian Nosek, nosek@virginia.edu

Author contribution statement: This paper was outlined and the literature review conducted through a crowdsourced process to which all authors contributed. EU, CE, BN, CC, TE, CL, RM, AR, & RS drafted the body of the manuscript. CE created the figure and tables. All authors provided critical edits and revisions. The third through ninth authors are listed alphabetically.

Abstract

Most scientific research is conducted by small teams of investigators, who together formulate hypotheses, collect data, conduct analyses, and report novel findings. These teams operate independently, as vertically integrated silos. Here we argue that scientific research that is horizontally distributed can provide substantial complementary value, aiming to maximize available resources, promote inclusiveness and transparency, and increase rigor and reliability. This alternative approach enables researchers to tackle ambitious projects that would not be possible under the standard model. Crowdsourced scientific initiatives vary in terms of the degree of communication between project members, from largely independent work curated by a coordination team to crowd collaboration on shared activities. The potential benefits as well as challenges of large scale collaboration span the entire research process: ideation, study design, data collection, data analysis, reporting, and peer review. Complementing traditional small science with crowdsourced approaches holds the potential to accelerate the progress of science and improve the quality of scientific research.

Keywords: crowdsourcing, collaboration, teams, methodology, meta-science

There is no perfect research study. Scientists, in their effort to understand nature, are constrained by limited time, resources, and expertise. This may produce a dilemma between choosing a lower quality, expedient approach or conducting a better powered, more intensive investigation allowing for stronger inferences. Ideals of the scientific process can be outweighed by the pragmatic reality of scientists' available resources and pursuit of career advancement. Scientists are rewarded for being the originator of new ideas and evidence through authorship of articles. These cultural incentives foster a focus on novelty and authorship that can come at the expense of rigor and foster questionable practices (Bakker, van Dijk, & Wicherts, 2012; Greenland & Fontanarosa 2012; Nosek, Spies, & Motyl, 2012; Open Science Collaboration, 2015). One alternative is for researchers to take more time for individual studies, expend more resources on each project, and publish fewer findings. Scientists could also work more collectively, combining resources across more contributors. But such choices have implications for productivity, individual credit, and career advancement.

Here we consider the standard model of scientific investigation and describe a complementary model – crowdsourcing science. Crowdsourced approaches seek to maximize the use of available resources, diversify contributions, enable big science, and increase transparency and reliability. Adaptation of cultural norms and incentives to promote crowdsourcing as a complement to the standard model promises to make science more rigorous and inclusive, and accelerate discovery.

Two models of doing science

Standard model: Vertical integration

Much of academic research resembles a vertically integrated business in certain respects. An individual or a small research team conceives a research question, designs studies to

investigate the question, implements the studies, analyzes the data, and writes a report of what was found. The closed team conducts the entire process from conceiving the idea to reporting the outcomes. The team members responsible for these steps are active collaborators and co-authors on a manuscript reporting the research. The sought after reward is acceptance and publication in the most widely read, prominent journal possible.

This model has a number of notable characteristics. It is *localized*, with funding distributed to particular labs and institutions, and *resource intensive*, with the project work divided among a few individuals. Access to productive research pipelines is constrained, and experience and status lead to opportunities to engage in research collaborations (Merton, 1968). It produces a large quantity of *small science*, with teams of limited size conducting projects that are correspondingly limited in scope – a small team can only collect so much data, carry out so many analyses, and consider so many alternatives to their methodology. Finally, contribution is recognized and rewarded through *authorship* on the final publication.

The standard model is akin to the philosopher model of scholarly contribution. An independent thinker conceives and generates a stand-alone piece of scholarship. After peer review by a small number of select colleagues, that scholarship is entered into the marketplace of ideas for others to examine, discuss, critique, and extend. Independence in developing and enacting the idea allows the scholar to dig deeply into a question or idea without interference, and credit allocation is straightforward. The scholar is evaluated based on the reception of her work in the idea marketplace. Outstanding ideas may become permanently linked to the scholar's identity, securing a lasting reputation and impact.

So what is wrong with the standard approach to science? For many research questions and contributions, nothing. Independently generated contributions are an efficient means of

getting initial evidence for many ideas into the marketplace. Indeed, the decentralized nature of science is presumed to feed productive generation and culling of ideas by the independent actions of scholars with different priors, assumptions, expertise, and interests. Oftentimes small teams work together repeatedly and develop co-specializations enabling deep dives into a methodology or phenomenon. A community of scientists then shares their work, exchanges feedback, and serially builds on each other's findings, mimicking crowd collaboration in some respects.

At the same time, for some research questions and contributions, the standard model may limit progress. There are trade-offs that individual researchers and small teams must consider when directing their research efforts. They could vary design elements and stimuli instead of holding them constant, collect larger samples for fewer studies instead of smaller samples for more studies, and, they could replicate their findings across multiple conditions or contexts rather than demonstrate a phenomenon and then move on. Researchers inevitably weigh these tradeoffs against the potential rewards. And, because the present culture prizes innovation and discovery (Bakker et al., 2012), some behaviors that would foster research credibility and cumulative progress are performed ineffectively or infrequently. Underperformed behaviors include collecting large, cross-cultural samples to evaluate generalizability and estimate effect sizes precisely (Henrich, Heine, & Norenzayan, 2010), replicating findings systematically in independent laboratories (Klein et al., 2014; Makel, Plucker, & Hegarty, 2012; Mueller-Langer, Fecher, Harhoff, & Wagner, in press; Simons, 2014), obtaining many different perspectives on how to analyze the same data (Silberzahn et al., in 2018), and employing a wide variety of study designs and stimuli (Judd, Westfall, & Kenny, 2012; Wells & Windschitl, 1999).

Alternative model: Horizontal distribution

The alternate model – crowdsourcing – eschews vertical integration and embraces horizontal distribution of ownership, resources, and expertise (Howe, 2006). In a distributed collaboration, numerous researchers each carry out specific components of a larger project, usually under the direction of a core coordination team (such that crowd projects are rarely perfectly horizontally distributed). Notably, modern science is already stretching the standard model in more collaborative directions (Supplement 1). Solo authorship is now the exception in most fields. This is partly due to the diversification of expertise required to conduct research with modern tools (Börner et al., 2010). Across disciplines, team size almost doubled from 1.9 in the 1960s to 3.5 in 2005 (Wuchty, Jones, & Uzzi, 2007a/b), and working in teams is associated with greater individual career success (Kniffin & Hanks, 2018). Team-authored papers are more cited than solo-authored papers, and this gap in scholarly impact has increased over time (Wuchty et al., 2007a/b).

Rather than two qualitatively distinct categories of research, the vertically integrated and horizontally distributed approaches are better conceived as a continuum, with variation in the depth of contribution by any given individual and the number of individuals contributing to the project. New opportunities and challenges emerge when moving further across the continuum from singular, independent scholars to a distributed, interdependent community. Crowdsourcing carefully selected research questions, in parallel to the necessarily far greater number of small team projects, holds a number of potential benefits for science— among these enabling the conduct of large-scale research projects, democratizing who contributes to science, and assessing the robustness of findings.

Enabling big science. An inclusive, diversified contribution model enables ambitious projects that would be unattainable by individuals working in isolation. Combining resources enables crowdsourced teams to enact research designs that vastly exceed what could be accomplished locally. Instead of holding sampling, stimulus, or procedural variables constant and hoping they do not matter, crowdsourced teams can allow them to vary and test whether they do. Instead of carrying out a low-powered, imprecise test, crowdsourced teams can conduct high-powered, precise studies and draw confident conclusions. Crowdsourcing complex activities seeks to mobilize the crowd's competencies, knowledge, and skills, and may leverage underused resources such as a better way to analyze the data, access to hard-to-recruit populations, knowledge of unpublished research or articles published in other languages, and translation of research materials into local languages and dialects. Crowdsourcing flips research planning from "what is the best we can do with the resources we have to investigate our question," to "what is the best way to investigate our question, so that we can decide what resources to recruit."

Democratizing science. Although person factors (Clemente, 1973; Hirsch, 2007; Williamson & Cable, 2003) and merit play a role in success in science, scientific careers also exhibit a Matthew effect (Merton, 1968). Early advantages in doctoral institution rank, professional connections, and grant funding accumulate benefits over time (Bol, De Vaan, & van de Rijt, in press; Clauset, Arbesman, & Larremore, 2015). Grant funding is overallocated to elite universities, and returns on investment would be greater if the funds were distributed more evenly (Wahls, 2018). At the other end of the academic hierarchy, early career researchers from less well-known institutions, underrepresented demographic groups, and countries that lack economic resources may never have a fair chance to compete (Petersen, Jung, Yang, & Stanley, 2011; Wahls, 2018). Academic fields are generally talent rich, such that globally distributed

projects can recruit individuals with advanced training and much to offer, yet too few resources to enact the vertical model competitively on their own. Few people enjoy the resource benefits of research intensive institutions including laboratory space, professional staff to support grant writing and management, graduate students, light teaching loads, and a community of colleagues for developing ideas and sharing infrastructure. Crowdsourcing aims to provide a new avenue through which those outside of major research institutions can contribute to high-profile projects, increasing inclusiveness, merit, and returns on investment (Chargaff, 1978; Feyerabend, 1982).

Assessing the robustness of findings. A crowdsourced approach is uniquely advantaged in determining the reliability and generalizability of findings. The ecosystem of standard science leads to the publication of massive numbers of small-sample studies (Pan, Petersen, Pammolli, & Fortunato, 2016), each with observations typically drawn from a single population (e.g., undergraduates from the researchers' home institution in the case of behavioral experiments; Sears, 1986). Combined with the filter of an academic review process that primarily permits statistically significant results to appear in the published record (Fanelli, 2010), the end result is research literatures filled with inaccurately estimated effect sizes due to publication bias (Ioannidis, 2005, 2008). The standard approach to science is also susceptible to issues such as study designs generated from a single theoretical perspective (Monin, Pizarro, & Beer, 2007), unconsidered cultural differences (Henrich et al., 2010), and researcher degrees of freedom in data analysis (Gelman & Loken, 2014; Simmons, Nelson, & Simonsohn, 2011). Large scale collaboration helped transform epidemiology into a more reliable field (Ioannidis, Tarone, & McLaughlin, 2011; Panagiotou, Willer, Hirschhorn, & Ioannidis, 2013), and this process is currently underway in psychology and other scientific disciplines. Multi-lab collaborations facilitate directly replicating findings (same materials and methods, new observations; Ebersole

et al., 2016; Klein et al., 2014) and conceptually replicating them (new approach to testing the same idea; Landy et al., 2018). Crowdsourcing research is a part of a changing landscape of science that seeks to improve research reliability and advance the credibility of academic research (LeBel, McCarthy, Earp, Elson, & Vanpaemel, in press; Nosek et al., 2012).

At the same time there are opportunity costs and diminishing returns involved in organizing many laboratories to carry out a single scientific investigation. Organizing a collective for a globally distributed project can create bureaucracy and transaction costs. For the same effort, a larger number of interesting ideas with initial supporting evidence could have been introduced into the literature by smaller teams working separately. Crowdsourcing allows for systematically examining cross-population variability, but it is important to begin by making sure the effect emerges reliably in at least one location. It will often be beneficial to rely on small-teams research for these reasons, especially when it comes to new areas of inquiry. Crowd projects with dozens or even hundreds of authors also create credit ambiguity and lack extrinsic incentives for participation, topics we address in depth later when we discuss structural reforms to encourage greater crowdsourcing. We believe the two models should coexist, with individual investigators and small teams generating initial evidence for new ideas, and crowdsourced initiatives employed to select particularly critical questions for intense examination. A diverse array of scientific projects, everywhere along the continuum from lone researchers to huge collectives, may produce the greatest return of useful knowledge from the resources invested. The remainder of this article discusses circumstances in which crowdsourcing offers particular opportunities and challenges as a complement to the standard model.

Forms of Scientific Crowdsourcing

Rather than supplanting the standard approach, organizing many individuals and laboratories into shared projects seeks to offset some of the weaknesses of vertically-integrated science. Crowd initiatives vary on multiple dimensions that can create advantages and disadvantages depending on the research application (Lakhani et al., 2007; Muffatto, 2006; Salganik, 2017; Srinarayan, Sugumaran, & Rajagopalan, 2002; Surowiecki, 2005). For example, crowdsourced projects vary in terms of the degree of communication between project members, from largely independent work curated by a coordination team to crowd collaboration on shared activities. Also, crowd science initiatives vary in their inclusivity, from open calls for collaborators to carefully chosen groups of topic experts.

Figure 1 crosses the horizontal dimension of communication (anchored at the left end by *curated contributions* and at the right by *crowd collaboration*) with the vertical dimension of selectivity to create a 2 x 2 matrix. Examples of relevant crowdsourced projects are placed in this matrix as illustrations. These projects are described greater detail in the next section and in Tables 1 and 2 and Supplements 1 and 2. Citizen science initiatives that include anyone willing to collect data involve a high degree of independence between actors, and thus fall into the bottom-left quadrant (Gura, 2013). Posing a research question to specialists (e.g., moral judgment researchers) and asking them to independently design studies to test the same idea falls into the top-left quadrant (Landy et al., 2018). Iterative contests in which topic experts work together to improve experimental interventions (Lai et al., 2014), and the collective development of open-source software (Muffatto, 2006) are in the top-right, and more inclusive forms of crowd writing (Christensen & van Bever, 2014) the bottom-right. Open peer review, in which anyone can publicly comment on a scientific manuscript or article, falls into the lower-right quadrant,

and crowd review by experts carefully chosen by a journal editor in the top-right quadrant. Traditional small-teams research, with unrestricted communication and select membership, falls outside the extreme upper-right corner of the matrix at the far end of both axes.

Multi-stage projects may operate in different locations in this space during the research lifecycle. For example, to explore consensus building about disparate findings from the same dataset, Silberzahn et al. (2018) segued from isolated individual work to round-robin feedback and then open group debate. Indeed, much crowdsourced science moves gradually from left to right on the communication dimension over the life course of the project, culminating in collective email exchanges and editing of the manuscript draft. Likewise, crowd projects tend to rely more on selective expertise over time (i.e., move up the vertical axis), as project coordinators and specialized sub-teams of statistical experts check the collective work for errors and play leading roles in producing the final report.

On the vertical dimension, greater inclusivity facilitates scaling up for massive initiatives. In contrast, selectivity in project membership prioritizes specific areas of expertise for contribution. It is not yet clear under what conditions involving large crowds of contributors (i.e., moving downward on the vertical axis) compromises overall project quality, relative to applying mild or strong selectivity standards for contribution (Budescu & Chen, 2015; Mannes, Soll, & Larrick, 2014). Research done by lone scientists and small teams is already known to be error prone (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; Garcia-Berthou & Alcaraz, 2004; Salter et al., 2014; Westra et al., 2011), and the quality-quantity tradeoff that can accompany scaling up is potentially offset by the numerous eyes available to catch mistakes (e.g., Silberzahn et al., 2018). The available evidence suggests data collected by citizen scientists are comparable in error rates and general quality to those assembled by professionals (Kosmala, Wiggins,

Swanson, & Simmons, 2016; Thelen & Thiet, 2008), online coders and political scientists reach near-perfect agreement on the policy positions in political manifestos (Benoit, Conway, Lauderdale, Laver, & Mikhaylov, 2016), Wikipedia entries are as accurate as the Encyclopedia Britannica (Giles, 2005), highly published and less prolific researchers are similarly likely to successfully replicate a given behavioral effect (Bench et al., 2017; see also Klein et al., 2018), and crowds of investigators do not exhibit measurably different "flair" at designing studies that obtain significant findings (Landy et al., 2018).

These null findings are surprising— there must be some point at which a crowd project becomes overly inclusive and insufficiently expert members compromise overall quality. One possibility is that coordinators of the crowd projects thus far have chosen the degree of inclusiveness and communication best suited to their research question (i.e., the correct location in Figure 1), leading to judicious scaling without losses in quality. Logically, only individuals with specialized training (e.g., with physiological equipment) would be recruited to collect data for certain projects (e.g., pooling fMRI data across laboratories; upper left quadrant of Figure 1). Even with an open call, potential contributors may volunteer for projects where they feel they can add value (e.g., an avid bird watcher volunteers to help track migrations), leading to self-screening based on relevant skill sets. Testing the conditions under which crowdsourcing increases and decreases project quality will inform future investments in crowdsourced research.

In contrast, there is little direct evidence regarding the consequences of information exchange between project members in crowdsourced scientific initiatives. Nevertheless, potential costs and benefits of crowd communication are suggested by the literatures on group influence and decision making. One of the virtues of crowds of independent agents, especially demographically and intellectually diverse ones, is their tendency to balance out individual

biases and errors in the aggregate (Galton, 1907; Larrick, Mannes, & Soll, 2012; Surowiecki, 2005). Crowdsourcing scientific investigations with little to no communication between project members (i.e., the far left regions of Figure 1) may help avoid the potentially biasing effect of individuals' overcommitment to intellectual claims (Berman & Reich, 2010; Luborsky et al., 1999; Manzoli, Flacco, D'Addario, Capasso, DeVito, Marzuillo, et al., 2014; Mynatta, Dohertya, & Tweneya, 1977), and path dependencies in which knowledge of others' approaches has an inordinate influence (Derex & Boyd, 2016). The effectiveness of crowds is more difficult to evaluate in situations that lack normatively correct answers or objective measures of accuracy. Yet even then, the diversity in approaches and results on the part of independent scientists, for example in analytic choices and study designs, is at least made transparent to the reader (Landy et al., 2018; Silberzahn et al., 2018).

That the "wisdom of the crowd" effect is spoiled when peer influence between members of the crowd is possible (Lorenz, Rauhut, Schweitzer, & Helbing, 2011), suggests that the more one moves toward crowd collaborations (i.e., right on the horizontal axis), the more conformity and deference to authority become risks. The one crowdsourced project that has tracked individual beliefs under conditions of gradually increasing communication found little evidence of convergence over time, beyond what would be expected based on sensitivity to new evidence (see Figure 4 of Silberzahn et al., 2018). The circumstances under which conformity effects occur in crowd science remains an open empirical question, and future projects should consider manipulating factors such as task interdependence and anonymity of communications.

Importantly, allowing information exchange and creating interdependencies between project members also comes with potential benefits. One of the hypothesized benefits of crowd collaboration is the ability of members of the community to learn from each other (Wenger,

1988). For example, teams in the Lai et al. (2014) intervention contest observed the effectiveness of others' interventions between rounds, and used those insights to improve their own interventions. Likewise, the round-robin feedback between different analytic teams in the crowdsourcing data analysis initiative (Silberzahn et al., 2018) helped several analysts to identify clear errors and adopt improved specifications. These are only anecdotal examples, and further research is needed to examine when peer learning occurs systematically in iterative, multi-stage crowd collaborations, and how it might best be facilitated. As reviewed next, evidence of the viability of crowdsourcing across all stages of the research process has accumulated rapidly in recent years.

Crowdsourcing science in action

Science can benefit from crowdsourcing activities that span the entire research process (see Table 1). These include coming up with research ideas, assembling the research team, designing the study, collecting and analyzing the data, replicating the results, writing the paper, obtaining reviewer feedback, and deciding next steps for the program of research. Table 2 and Supplement 2 summarize some recent crowdsourced scientific initiatives, organized by which of these respective stages they focused crowd efforts on.

Ideation

Crowds of scientists can be organized to collaborate virtually on complex problem solving challenges, each proposing ideas for solving components of the problem and commenting on each others' suggestions (open communication, the far right regions of Figure 1). This approach has been used to great effect in the Polymath projects, resulting in a number of important mathematical proofs (Ball, 2014; Polymath, 2012; 2014; Tao, Croot, & Helfgott, 2012). Similar to a product design contest (Poetz & Shreier, 2012), crowds of researchers can

also be used to generate original research hypotheses and select which ideas are most likely to be of broad interest and impact (Jia et al., 2018; Schweinsberg, Feldman, et al., 2018). This approach may be particularly useful when it comes to datasets which for legal or ethical reasons cannot be publicly posted or further distributed— for instance, the personnel records of a private firm, who might agree to share them with one research team or institution but not for general distribution. Even in such cases, the core coordination team who serve as custodians of the data can post an overview of the variables and sample online, and publicly solicit ideas for testing (Jia et al., 2018). The crowdsourced generation and selection of research ideas is one way to open up datasets and collaboration opportunities that would otherwise remain closed to most scientists.

Assembling resources

Genome-wide association studies (GWAS) distribute the task of investigating the entire genome across many collaborators and institutions with specialized roles, leading to important discoveries related to genes and pathways of common diseases (Visscher et al., 2012). Consider the innumerable lost opportunities for similarly combining resources across laboratories in other scientific fields. For instance, a researcher at one institution may have a great idea, but lacks access to the right equipment or sample of subjects to test it. Elsewhere, another team find they have an excess of research resources (e.g., they compensate participants for a 30-minute session for completing a 15-minute study). Some researchers have resources that could productively be used by other researchers who need those resources to meet their research goals. One way to attempt to minimize the collective waste and maximize researchers' collective ability to meet their research goals is to match 'haves' with 'needs' using online platforms such as Science Exchange (scienceexchange.com) and StudySwap (http://osf.io/view/StudySwap/). Such exchanges, which could be expanded into full-scale online academic labor markets similar to

oDesk or Elance (Horton, 2010), seek to push academic communities into the upper right quadrant of Figure 1 by opening novel lines of communication and creating opportunities to connect resources and expertise.

Study design

Another limitation to standard science is narrow sampling of the constructs-of-interest (Baribault et al., in press; Judd et al., 2012; Monin & Oppenheimer, 2014; Wells & Windschitl, 1999). A small team is at risk of generating a limited set of stimuli, operationalizations of variables, and study designs. Another team might have carried out a very different test of the same idea, based on different prior training and theoretical assumptions. Even seemingly small differences in methods might produce substantial differences in research results. An alternative crowd approach is to assign the same research question to different experts, who then independently design studies aimed at answering it (upper left corner of Figure 1, low communication combined with high expertise). Landy et al. (2018) did precisely this, finding that variability in effect sizes due to researcher design choices was consistently high. Indeed, study designs from different researchers produced significant effects in opposite directions for four of five research questions related to negotiation, moral judgment, and implicit cognition. Crowdsourcing conceptual replications more effectively reveals the true consistency in support for a scientific claim.

Data collection

Online platforms for crowdsourced labor such as Amazon's Mechanical Turk (MTurk) have become widely used as a source of inexpensive research participants and coders (Stewart, Chandler, & Paolacci, 2017) (Supplement 3). Rather than merely serving as research subjects, members of the general public can also be recruited to collect data and observations. This

strategy moves the project into the bottommost left corner of Figure 1 of inclusive projects with low communication, with anyone willing to help being included as a project member. The tradition of citizen science dates back to Denison Olmsted's use of observations from a crowd of both amateur and professional astronomers to track the great meteor storm of 1833 (Littmann & Suomela, 2014; Olmsted, 1934). Citizen science today is a movement to democratize science (Chargaff, 1978; Feyerabend, 1982), engage the public, create learning opportunities, and gather data and solve problems at minimal cost with the aid of a host of volunteers (Cavalier & Kennedy, 2016; Gura, 2013). Amateur, non-professional scientists participate actively in scientific investigations in biology, astronomy, ecology, conservation, and other fields, working under the direction of professionals at research institutions. A related approach is to gamify scientific problems and recruit citizen scientists to aid in cracking them, as in the video game Quantum Moves in which players move digital renditions of atoms (Sørensen et al., 2016), the online EyeWire game in which players help reconstruct eye cells (Kim et al., 2014), and the protein folding game FoldIt (Cooper et al., 2010). Notably, for some types of citizen science projects contributors may have substantial skills and knowledge, or even formal training such as an advanced degree, and in such cases are far from novices. One of the strengths of crowdsourcing is the ability to tap into the expertise of individuals outside of mainstream academia who are able and willing to contribute to science.

Data analysis

Researchers working with a complex dataset are confronted with a multitude of choices regarding potential statistical approaches, covariates, operationalizations of conceptual variables, and the like. In a quantitative review, Carp (2012a, 2012b) found that 241 published fMRI papers used 223 distinct analytic strategies. Researchers may consciously or unconsciously choose

statistical specifications that yield desired results, in particular statistically significant results in support of a favored theory (Bakker et al., 2012; Ioannidis, 2005; Ioannidis & Trikalinos 2007; Simonsohn, Nelson, & Simmons, 2014; Simmons et al., 2011). One way to maximize transparency is to turn the analysis of data over to a crowd of experts. The same dataset is distributed to numerous scientists who are asked to test the same theoretical hypothesis, at first without knowledge of the specifications used by their colleagues (upper left quadrant of Figure 1, high expertise combined with low communication). This offers an opportunity to assess how even seemingly minor differences in choices may affect research outcomes, and reduces pressure to observe any particular outcome – at least for purposes of publishability. Silberzahn et al. (2018) found that 29 different teams of analysts used 29 distinct specifications and returned effect size estimates for the same research question ("Do dark skin toned soccer players receive more red cards?") that ranged from slightly negative to large positive effects. Crowdsourcing the analysis of the data reveals the extent to which research conclusions are contingent on the defensible, yet subjective decisions made by different analysts.

The growth of large-scale data has created opportunities to leverage this diversity to identify the most robust means of analyzing such complex and massive datasets. Crowdsourced challenges have been used by researchers for benchmarking new computational methods, as with for instance the DREAM (Dialogue for Reverse Engineering Assessments and Methods) Challenge focused on predicting survival of breast cancer patients (Saez-Rodriguez et al., 2016; Stolovitzky, Monroe, & Califano, 2007). Organizers provide a test data set and a particular question to be addressed to many independent analysts (an upper left quadrant approach), then apply the analytic strategies to a hold-out dataset to evaluate their robustness.

Another innovative method is to hold constructs, models, and covariates constant, and leverage a network of researchers to carry out this same analysis on different existing datasets (a *coordinated analysis*; Hofer & Piccinin, 2009). This approach was pioneered by the Integrative Analysis of Longitudinal Studies on Aging (IALSA) network (Lindwall et al., 2012). Testing a research question of common interest (e.g., does education protect against cognitive decline?; Piccinin et al. 2013) on existing datasets that include the same constructs (e.g., measures of cognitive function such as memory, reasoning, and fluency) and yet measures them in disparate ways in different populations (e.g., Sweden, Austria, Netherlands, and the United Kingdom) far more systematically assesses the generalizability of the results than relying on a single data source. Since members of this network of experts communicate extensively to agree upon their shared analytic approach and measures to use from each longitudinal dataset, a coordinated analysis falls into the upper right quadrant of Figure 1.

Note that all these approaches are qualitatively different from fields in which many researchers independently leverage a central data source (e.g., the General Social Survey; GSS). In fields like political science, resources like the GSS are used to investigate separate research questions, such that aggregation and meta-scientific comparisons are less informative. Crowdsourcing is especially useful, we suggest, for fields that are reliant on local resources that can remain siloed. That said, the data corpus generated by crowdsourced projects often serve as public resources after the publication of the article (e.g., Open Science Collaboration, 2015; Tierney et al., 2016).

Replicating findings prior to publication

Individual laboratories are typically constrained in the amount and type of data they can collect. Replicating unpublished findings in independent laboratories before they are submitted

for publication (Schooler, 2014; Tierney, Schweinsberg, & Uhlmann, in press) addresses power and generalizability directly. Authors can specify *a priori* in which replication samples and laboratories they expect their findings to emerge, for example selecting only topic experts as their replicators and thus moving up the vertical axis of Figure 1. This approach, which thus far returns modest reproducibility rate even under the seemingly best of conditions (Schweinsberg et al., 2016) has recently been integrated into graduate and undergraduate methods classes (Schweinsberg, Vignanola, et al., 2018), thus traveling downward along the vertical axis towards greater inclusiveness. Such crowdsourced pedagogical initiatives are one means of turning replication into a commonplace aspect of how science is conducted and students are educated (Everett & Earp, 2015; Frank & Saxe, 2012; Grahe et al., 2012).

Writing research reports

The conceptualization, drafting, and revision of research articles represents another opportunity to leverage distributed knowledge. The article "The Capitalist's Dilemma," conceptualized and written by two professors and 150 of their MBA students, is one example (Christensen & van Bever, 2014). As with other forms of collaborative writing online, such as Wikipedia, channeling the contributions of many collaborators into a quality finished paper requires a few group leaders who complete a disproportionate amount of the work, and also organize and edit the written material of others (Kittur & Kraut, 2008; Kittur, Lee, & Kraut, 2009). Our personal experience with many-authored papers is that a large number of contributors commenting publicly on the draft greatly facilitates working out a solid framework and set of arguments, identifying relevant articles and literatures to cite (especially unpublished work), ferreting out quantitative and grammatical errors, and tempering claims appropriately. More radically, efforts such as CrowdForge suggests that non-experts (e.g., elite Mechanical Turk

workers), are surprisingly capable at drafting quality summaries of scientific findings for lay readers (Kittur, Smus, & Kraut, 2011). Such quality raw material could be carefully vetted and included in reviews of scientific research for practitioners and lay audiences. This suggest cautious optimism in moving down the vertical axis of Figure 1 to allow for written work from unconventional contributors, with the degree of inclusiveness varying by the technical expertise and topic knowledge required for a given paper.

Peer review

In the current system of academic peer review, an unpublished paper is submitted to a journal and evaluated by the editor and usually 2-5 external referees, each of whom provide detailed feedback, often over multiple rounds of revisions and serially across multiple journals. Even when successful, it can be a slow and arduous process taking months or years. For example, Nosek and Bar-Anan (2012) reported a case study of a researcher's corpus of publications and found that the average time from manuscript submission to ultimate publication was 677 days. There is little doubt that detailed feedback from colleagues can be immensely helpful, yet it remains unknown whether research reports are consistently improved by the review process (Anonymous, 2005). Empirical studies indicate that the interrater reliability of independent assessors is low, with median reliability coefficients of .30 for journal articles and .33 for grant reviews (Bornmann & Daniel, 2010; Cicchetti, 1991; Marsh, Jayasinghe, & Bond, 2008), and bias in favor of authors with strong networks (Wenneras & Wold, 1997). There are also the diminishing returns on time investments to consider – completing iterative rounds of review and revisions consumes time that might have been better allocated to pursuing a novel scientific discovery. The reviewers, typically anonymous, receive minimal professional benefit from their work, and the broader community may never hear worthy criticisms left unaddressed

in the published version of the paper. Ultimately, publication in a prestigious outlet is a poor signal of an article's scholarly impact, with journal impact factors driven by outlier articles and only a weak predictor of the citations accrued by the typical article in the journal (Baum, 2011; Holden, Rosenberg, Barker, & Onghena, 2006; Seglen, 1994).

An alternative is to open scientific communication and crowdsource the peer review process (Nosek & Bar-Anan, 2012). This moves rightward on the horizontal axis by opening communication, and downward on the vertical axis to the extent the review process is inclusive of many commentators. Both might be accomplished simultaneously using a centralized platform for review and discussion of research reports, with a content feed similar to social media sites (e.g., Facebook, Twitter) and users able to comment on and evaluate content as with the websites run by Reddit, Yelp, Amazon, and others (Buttliere, 2014). Posted files could include not only manuscripts, but also datasets, code, materials, and re-analyses, replications, and critiques by other scientists. Peer review would be open, credited, and citable, and for prominent articles that attract attention evaluation would be carried out by a potentially more reliable crowd of scientists rather than a small group of select colleagues. Further, reviewers would have access to the underlying data, facilitating the early identification of errors (Sakaluk, Williams, & Biernat, 2014). Measures of contribution would be diverse, with scholarly reputation enhanced not just via citations to authored manuscripts, but also intellectual impact via proposals of novel ideas, posting of data and code that others find useful, providing insightful feedback on others' work, and curation of content related to specialized topic areas (e.g., replicability of the effects of mood on helping behaviors; Lebel et al. in press). Original authors would have the opportunity to update their article in light of new evidence or arguments, with older versions archived, as in the Living Reviews group of journals in Physics.

In contrast to such a radical bottom-right quadrant approach (open communication, highly inclusive), upper-right quadrant versions of peer review would invite a crowd of topic experts carefully selected by a journal editor. However, in this more conservative scenario journal reviews would still be public, citable, and greater in number than is currently the norm. Open and citable reviews allow readers who weight traditional credentials highly to do so, while individuals lower in formal expertise but whose comments are high in quality have the opportunity to be recognized. The barriers to wider experimentation are not so much technological – there are already platforms that facilitate open scientific communication (Wolfman-Arent, 2014) – but rather social, with current professional reward structures still encouraging publication via the traditional process and outlets. Only by experimenting with diverse approaches, some staying close in important respects to traditional academic review and others departing radically, can we identify the most effective ways to communicate scientific ideas and knowledge.

Replicating published findings

Among the best known uses of crowdsourcing are large-scale initiatives to directly replicate published research in psychology, biomedicine, economics, and other fields (e.g., Alogna et al., 2014; McCarthy, et al., 2017; Errington et al., 2014; O'Donnell et al. in press). In these crowdsourced projects, up to a hundred laboratories attempt to repeat the methodology of previous studies, collecting much larger samples to provide improved statistical power to detect the hypothesized effect. Aggregating across six major replication initiatives in the social sciences, examining 190 effects in total, crowdsourced teams successfully replicated 90 (47%; Camerer et al., 2018, 2016; Ebersole et al., 2016; Klein et al., 2018, 2014; Open Science Collaboration, 2015).

A crowdsourced approach reveals that high levels of heterogeneity in effect size estimates across laboratories are observed primarily for large effects, not small ones (Klein et al., 2018). In other words, effects that fail to replicate tend to do so consistently across cultures and demographic populations, casting doubt on the argument that as-yet-unidentified moderators explain why approximately half of published findings do not emerge when tested in independent laboratories. That there are no consistent laboratory differences in effect size estimates (i.e., some research teams are not "better" than others at obtaining support for the original hypothesis; Bench et al., 2017; Klein et al., 2018; 2014) suggests that cautious scaling (e.g., moving downward on the vertical axis of Figure 1 toward greater inclusiveness) ought to be considered. The Collaborative Replications and Education Project (CREP; Grahe et al., 2015; Wagge et al., in press) seeks to achieve this by organizing undergraduate experimental methods classes into research teams, an approach that promises to radically scale up data collection for replications by integrating this activity into student education (Everett & Earp, 2015; Frank & Saxe, 2012). The Psychological Science Accelerator (PSA), an international network of over 300 psychological science laboratories, have committed to contributing to large-scale collaborations on an ongoing basis, including regularly involving their students via the Accelerated CREP initiative (Moshontz et al., 2018).

Deciding what findings to pursue further

Faced with a voluminous and constantly growing research literature – more than 30 million academic papers have been published since 1965 (Pan et al., 2016) – and also evidence that many published findings are less robust than initially thought (Begley & Ellis, 2012; Errington et al., 2014; Open Science Collaboration, 2015; Prinz, Schlange & Asadullah, 2011) the question becomes how best to distribute limited replication resources. Viable options include

focusing on highly cited papers, findings covered in student textbooks, results that receive widespread media coverage, or on research with practical relevance (e.g., for government policies, or interventions to reduce demographic gaps in educational attainment). The replication value of a study might be calculated based on the impact of the finding relative to the strength of the available evidence (e.g., statistical power of the original demonstrations; Nosek et al., 2012).

Another, complementary rather than competing approach is to leverage the collective wisdom of the scientific community. The aggregated estimates of crowds perform surprisingly well at predicting future outcomes – such as election results, news and sporting events, and stock market fluctuations – since in many cases the aggregation cancels out individual errors (Galton, 1907; Mellers et al., 2014; Surowiecki, 2005). Similarly, the averaged independent predictions of scientists regarding research outcomes – based solely on examination of research abstracts and study materials – are remarkably well aligned with realized significance levels and effect sizes (Camerer et al., 2016; DellaVigna & Pope, in press, 2018; Dreber et al., 2015; Forsell et al., 2018; Landy et al., 2018). Senior academics (e.g., full professors) and junior academics (e.g., graduate students and research assistants) exhibit similar forecasting accuracy (DellaVigna & Pope, in press, 2018; Landy et al., 2018), suggesting the feasibility of an inclusive bottom-left quadrant approach. It may be reasonable to avoid allocating replication resources to findings a heterogeneous crowd of scientists consider either clearly spurious or well-established, and focus on findings about which beliefs are conflicting or uncertain.

A decision market might be used to select among the many available options for independent replication, the idea being to allocate resources as efficiently as possible. Crowdsourced replications will be most useful when a clear, widely agreed upon question of broad interest is present. Large scale efforts seem less appropriate for findings the community

considers highly unlikely to be true (e.g., extra sensory perception) or not particularly theoretically interesting if true. Such crowd-based selection might be ongoing, with attention dynamically shifting away from effects that have experienced repeated replication failures and for which the community's expectations drop below a predetermined threshold (Dreber et al., 2015). This would help prevent cases in which numerous laboratories conduct replications of an effect, collecting many thousands of participants, when fewer tests would have already led to strong inferences. Decision markets might also be used to select the most and least likely populations an effect should emerge in as an initial test of universality (Norenzayan & Heine, 2005).

Crowd science can also be used to make gradual improvements to existing research paradigms and interventions. Lai and colleagues (2014; 2016) held a series of crowdsourced contests to identify the best interventions for reducing implicit racial biases. Beginning in the upper-left quadrant of Figure 1 (low communication, high expertise), research teams submitted 17 interventions to reduce implicit biases (e.g., exposure to positive exemplars, perspective-taking, empathy). Of those interventions, 8 successfully reduced implicit intergroup bias in the short-term. Moving horizontally into the upper-right of quadrant by adding the element of information exchange, teams were able to observe and learn from each other's approaches between rounds of data collection. Several teams used this opportunity to improve their own intervention, leading to progressively greater effectiveness in reducing intergroup bias across rounds. We believe this contest model holds widespread applicability for identifying and improving upon practical interventions to address societal challenges. We envision a future scientific landscape in which forecasting surveys and decision markets are run in tandem with research contests and other large-scale empirical data collections on an ongoing basis.

Reforms to Facilitate Large-Scale Collaboration

We believe most researchers have intrinsic interest in contributing to knowledge accumulation and are not solely driven by prestige goals. At the same time, professional reward systems can be updated in ways to encourage voluntary participation in large scale collaboration, and better align intrinsic and extrinsic motives. The current culture and reward system impose pressures for researchers to act independently as opposed to collectively, and pursue initial evidence for novel findings rather than engage in systematic verification, more than is ideal for scientific progress. Further, although merit matters in science, there are also Matthew effects (Bol, De Vaan, & van de Rijt, in press; Clauset, Arbesman, & Larremore, 2015; Merton, 1968; Petersen et al., 2011; Wahls, 2018). The resulting hierarchical and network-based arrangements interfere with inclusivity for researchers with much to offer, but disadvantaged backgrounds and resources. Thus, we advocate for changes to include greater rewards for collective engagement.

Distribution of grant funding

Empirical evidence suggests that distributing grant funding more evenly would increase total return on investment in terms of scientific knowledge (Wahls, 2018). Receipt and renewal of such funds could be further linked to evidence of ongoing contributions to open science. These might include publicly posting data and materials (Simonsohn, 2013), disclosing data exclusions and stopping rules (Simmons et al., 2011), running highly powered studies (Stanley, Carter, & Doucouliagos, 2018), pre-registration of studies and analysis plans (Nosek, Ebersole, DeHaven, & Mellor, 2018; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Nosek & Lakens, 2014), conducting replications, helping develop new methods, sharing resources on platforms such as StudySwap, and participating in crowdsourced initiatives, among other options. A more equitable distribution of financial support for research could reward merit

and encourage excellence by not only providing additional opportunities for those with useful skills and knowledge to contribute (Wahls, 2018), but also directly incentivizing emerging best practices. To avoid diffusion of responsibility on projects with many collaborators, not only authorship but also grant funding might be made contingent on specific deliverables (e.g., minimum number of participants collected, provision of annotated analysis code others can reproduce, etc).

Author contribution statements

Although some especially elaborate crowd projects involve specialized sub-teams who are able to publish a separate report of their work (e.g., Dreber et al., 2015; Forsell et al., 2018), these are atypical cases. Many authored papers reporting large scale projects require reforms in how intellectual credit is allocated. Inputs can be documented through careful and detailed author contribution statements, which are increasingly required at academic journals. A good starting point for the crafting of clear contribution statements is the CRediT taxonomy (Brand, Allen, Altman, Hlava & Scott, 2015). Contributions throughout the full research life-cycle are represented in categories such as conceptualization, data curation, writing, and visualization. Providing information about which co-authors contributed to which CRediT categories allows collaborators to transparently communicate how authorship was determined and which author deserves credit for which components of a research project. This sort of detailed accounting is a necessary precursor for the acceptance of increasingly long author lists that are already commonplace in fields such as high-energy physics.

Selection and promotion criteria

In addition to traditional metrics of scholarly merit, search and promotion committees should take into account an applicant's contributions to conducting rigorous research and making

science as a whole better. In some fields, a demonstrated commitment to open science and scientific reform is already starting to be factored into selection and promotion decisions (Nosek, 2017; Schönbrodt, 2018). One way in which applicants might choose to fulfill these criteria is by participating in crowdsourced initiatives to replicate findings, reanalyze data, generate and select ideas, and so forth. Comprehensive shifts in incentives will require that hiring and tenure and promotion committees rely more on specific indicators of contribution (Brand et al., 2015), such as the author contribution statements described above, rather than heuristics of counting papers and whether the person was first, last, or somewhere in the middle of an authorship list. In this way, individuals who led an important subcomponent of a massive project (e.g., the subteam that conducted the forecasting survey, qualitative analyses, or Bayesian meta-analysis) can be more fairly recognized.

Another, more radical option is making entire project workflows open and linked to each contributor (something possible through the Open Science Framework), and for hiring and promotion committees to examine these workflows before making their decisions. In a future in which open peer review becomes commonplace, online links to feedback provided on the articles of colleagues might be formally listed on CVs, as further evidence of intellectual contribution and service to the field. If the multifold aspects of an academic's workflow are made transparent, decision makers can move beyond heuristics and use more complete information to better allocate rewards based on merit.

Integrating crowd science into pedagogy

Another way to encourage crowd science is to build such initiatives into activities scientists in many fields already do routinely, such as collecting data in methods classes for student projects, and analyzing complex datasets as part of graduate education (Everett & Earp,

2015; Frank & Saxe, 2012; Grahe et al., 2012; Mavor et al., 2016). The CREP (Grahe et al., 2015; Wagge et al., in press) and Pipeline Projects (Schweinsberg et al., 2016; Schweinsberg, Viganolla, et al., 2018) offer opportunities to leverage such activities for many-authored crowdsourced replications. In these cases, middle author on the report of an interesting initiative has only to beat no professional reward at all to make rational sense for students and course instructor alike. Crowdsourcing avoids letting the students' hard work at collecting data go to waste repeating established paradigms (e.g., the Stroop effect) in unpublishable class projects whose results are low in information gain. As a further incentive, the Pipeline Project 2 offers course instructors a free curriculum they can use in their lectures, reducing course preparation time (https://osf.io/hj9zr/). Whether graduate programs provide opportunities for experiential education and authored work on crowd science projects could potentially be factored into their rankings and accreditations.

Changes in publication criteria

Top-down changes in publication requirements at journals (e.g., disclosure rules and open science badges) are already changing how science is done and what gets published (Everett & Earp, 2015; Nosek et al., 2015). Such systematic shifts in policies help avoid collective action problems such that only a subset of scientists engage in best practices that increase research quality but may also reduce productivity, which risks placing them at a professional disadvantage (Kidwell et al., 2016). One option, aimed at encouraging pre-publication independent replication (Schweinsberg et al., 2016) is to include independent verification of findings in another laboratory as a publication criterion at the most prestigious empirical journals (Mogil & Macleod, 2017). It is often useful to get initial evidence for a finding out there to be examined and debated by the scientific community, and individual careers should continue to advance

primarily in this way. However, it is also reasonable for those publication outlets that provide the most professional benefit to authors, and are perhaps perceived as most authoritative (e.g., *Science, Nature, Proceedings of the National Academy of Sciences*), to set the bar higher. Prominent journal outlets are also increasingly recognizing the value of meta-scientific work that relies on a crowd approach, a trend that promises to encourage future crowdsourced projects. A more general shift in emphasis towards rigorous verification, relative to novelty, as a publication criterion would incentivize high-powered crowd projects well positioned to assess the replicability and generalizability of findings.

Developing infrastructure

Another avenue is to create infrastructure and tools to make crowdsourcing easier and more efficient to do. Online platforms such as the Harvard Dataverse (https://dataverse.harvard.edu/) and Open Science Framework (https://osf.io/) are available to host data, research and teaching materials, preregistrations, and document workflows. Journal mechanisms such as Registered Reports that review methodology and accept-in-principle prior to data collection have now been adopted at scores of outlets (https://cos.io/rr/), and journals are increasingly experimenting with innovative formats such as open review, crowd review, and updatable papers. Recently introduced tools like StudySwap and standing laboratory networks such as the Psychological Science Accelerator likewise hold promise to change the landscape of everyday science.

Importantly, these approaches to encourage large scale collaboration are complements to reforms in how small team science is conducted and funded. Larger samples (Stanley, Carter, & Doucouliagos, 2018), disclosure rules (Simmons et al., 2001), preregistration (Nosek et al., 2018; Wagenmakers et al., 2012), and Registered Reports formats at journals (Chambers, 2013; Nosek

& Lakens, 2014) promise to increase the true positive rate for small studies, with scaling up for crowd projects then allowing for strong inferences about the generalizability vs. context sensitivity of particularly important findings. At the same time, crowdsourced meta-scientific investigations can help assess the effectiveness of new practices intended to improve science, but which may also have unwanted side effects. For instance, preregistration might reduce false positive results, but could also negatively impact the rate of novel discoveries by dampening creativity (Brainerd & Reyna, 2018). A crowdsourced project in progress (Ebersole et al., 2018a), will randomly assign researchers to pre-register their analyses of a complex dataset or not, to empirically assess the costs and benefits of this proposed reform. Finally, encouraging large scale collaborations to help democratize participation in research complements grants to support research at teaching institutions, addressing gender gaps in representation, and other efforts to reduce systematic inequalities in science.

Conclusion

Crowdsourcing holds the potential to greatly expand the scale and impact of scientific research. It seeks to promote inclusion in science, maximize material and human resources, and make it possible to tackle problems that are orders of magnitude greater than what could be solved by individual minds working independently. Although most commonly employed in the data collection phase of research and for conducting replications, opportunities to take advantage of a distributed, interdependent collective span the entire scientific endeavor – from generating ideas to designing studies, analyzing the data, replicating results, writing research reports, providing peer feedback, and making decisions about what findings are worth pursuing further. Crowdsourcing is the next step in science's progression from individual scholars to increasingly larger teams and now massive globally distributed collaborations. The crowdsourcing movement

is not the end of the traditional scholar nor the vertically integrated model. Rather, it seeks to complement this standard approach to provide more options for accelerating scientific discovery.

References

Afshinnekoo, E., et al. (2016). Globalizing and crowdsourcing biomedical research. *British Medical Bulletin, 120,* 27-33.

Afshinnekoo, E., Meydan, C., Chowdhury S., et al. (2015). Geospatial resolution of human and bacterial diversity from city-scale metagenomics. *Cell Systems, 1,* 72–87.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., ... Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556–578.

Allen, G.I., Amoroso, N., Anghel, C., Balagurusamy, V., Bare, C.J., Beaton, D., Bellotti, R., Bennett, D.A., Boehme, K.L., Boutros, P.C., Caberlotto, L., Caloian, C., Campbell, F., Chaibub Neto, E., Chang, Y.-C., Chen, B., Chen, C.-Y., Chien, T.-Y., Clark, T., Das, S., Davatzikos, C., Deng, J., Dillenberger, D., Dobson, R.J.B., Dong, Q., Doshi, J., Duma, D., Errico, R., Erus, G., Everett, E., Fardo, D.W., Friend, S.H., Fröhlich, H., Gan, J., St George-Hyslop, P., Ghosh, S.S., Glaab, E., Green, R.C., Guan, Y., Hong, M.-Y., Huang, C., Hwang, J., Ibrahim, J., Inglese, P., Iyappan, A., Jiang, Q., Katsumata, Y., Kauwe, J.S.K., Klein, A., Kong, D., Krause, R., Lalonde, E., Lauria, M., Lee, E., Lin, X., Liu, Z., Livingstone, J., Logsdon, B.A., Lovestone, S., Ma, T., Malhotra, A., Mangravite, L.M., Maxwell, T.J., Merrill, E., Nagorski, J., Namasivayam, A., Narayan, M., Naz, M., Newhouse, S.J., Norman, T.C., Nurtdinov, R.N., Oyang, Y.-J., Pawitan, Y., Peng, S., Peters, M.A., Piccolo, S.R., Praveen, P., Priami, C., Sabelnykova, V.Y., Senger, P., Shen, X., Simmons, A., Sotiras, A., Stolovitzky, G., Tangaro, S., Tateo, A., Tung, Y.-A., Tustison, N.J., Varol, E., Vradenburg, G., Weiner, M.W., Xiao, G., Xie, L., Xie, Y., Xu, J., Yang, H., Zhan, X., Zhou, Y., Zhu, F., Zhu, H., Zhu, S. (2016). Crowdsourced

estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimer's & Dementia, 12,* 645–653. https://doi.org/10.1016/j.jalz.2016.02.006

Anonymous. (2005). Revolutionizing peer review? *Nature Neuroscience, 8*(4), 397. http://doi.org/doi:10.1038/nn0405-397

Ball, P. (2014). Crowd-sourcing: Strength in numbers. *Nature, 506,* 422–423.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods, 43,* 666-678.

Bakker, M., van Dijk, A, & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543-554.

Bargh, J. (1997). The automaticity of everyday life. In R.S. Wyer, Ed., *Advances in social cognition*.

Bargh, J. A. (2014). Our unconscious mind. *Scientific American, 30,* 30-27

Baribault, B., Donkin, C., Little, D. R., Trueblood, J., Oravecz, Z., van Ravenzwaaij, D., White, C. N., de Boeck, P., & Vandekerckhove, J. (in press). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences.*

Baum, J.A. (2011). Free-riding on power laws: Questioning the validity of the impact factor as a measure of research quality in organization studies. *Organization ,18*(4), 449–466.

Baumeister, R.F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology*, *66*, 153-158.

Begley, C.G., & Ellis, L.M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483,* 531–533.

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*(3), 800-813.

Bench, S. W., Rivera, G. N., Schlegel, R. J., Hicks, J. A., & Lench, H. C. (2017). Does expertise matter in replication? An examination of the Reproducibility Project: Psychology. *Journal of Experimental Social Psychology, 68*, 181-184.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review, 110(2)*, 278–95.

Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research, 16,* 202–207. doi:10.1002/mpr.225

Berman, J. S. & Reich, C. M. (2010). Investigator allegiance and the evaluation of psychotherapy outcome research. *European Journal of Psychotherapy and Counselling 12*, 11–21.

Bol, T., De Vaan, M., & van de Rijt, A. (in press). The Matthew Effect in science funding. *Proceedings of the National Academy of Sciences*.

Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S.M., Hall, K. L., Keyton, J., Spring, B., Stokols, D., Trochim, W., & Uzzi, B. (2010). A multi-level systems perspective for the science of team science. *Science Translational Medicine, 2*, 49cm24.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: a multilevel meta-analysis of inter-rater reliability and its determinants. *PloS ONE*, 5(12), e14331. doi:10.1371/journal.pone.0014331

Brabham, D. C. (2013). *Crowdsourcing*. Mit Press.

Brabham, D. C. (2010). Moving the crowd at Threadless: Motivations for participation in a crowdsourcing application. *Information, Communication & Society, 13*(8), 1122-1145.

Brainerd, C. J., & Reyna, V. F. (2018). Replication, registration, and scientific creativity. *Perspectives on Psychological Science, 13*, 428–432. doi:10.1177/1745691617739421

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing, 28*(2), 151-155.

Budescu, D.V. & Chen. E. (2015). Identifying expertise to extract the Wisdom of Crowds. *Management Science, 61*, 267-280. http://dx.doi.org/10.1287/mnsc.2014.1909.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3-5.

Butcher, G.S. (1990). Audubon Christmas Bird Counts, p. 5–13. In J. R. Sauer and S. Droege [EDS.], *Survey designs and statistical methods for the estimation of avian population trends.* U.S. Fish and Wildlife Service, Biological Report 90(1).

Buttliere, B.T. (2014). Using science and psychology to improve the dissemination and evaluation of scientific work. *Frontiers in Computational Neuroscience.* https://doi.org/10.3389/fncom.2014.00082

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., … & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*, 1433–1436.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E-J., & Wu, H. (2018). Evaluating replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*.

Carp, J. (2012a). The secret lives of experiments: Methods reporting in the fMRI literature.

NeuroImage, *63*(1), 289–300. doi:10.1016/j.neuroimage.2012.07.004

Carp, J. (2012b). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, *6*, 149. doi: 10.3389/fnins.2012.00149

Cavalier, D., & Kennedy, E. (2016). *The rightful place of science: Citizen science*. Consortium for Science, Policy & Outcomes. Tempe, Arizona.

Chalmers, I., Bracken, M.B., Djulbegovic, B., Garattini, S., Grant, J., et al. (2014). How to increase value and reduce waste when research priorities are set. *Lancet, 383*, 156.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112-130.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*, 53-81.

Chargaff, E. (1978). *Heraclitean fire: Sketches from a life before nature*. New York: Rockefeller University Press.

Chartier, C., Kline, M., McCarthy, R. Nuijten, M., Dunleavy, D., & Ledgerwood, A. (2018). *The cooperative revolution is making psychological science better.* APS Observer. Available at: https://www.psychologicalscience.org/observer/the-cooperative-revolution-is-making-psychological-science-better

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*, 609–610.

Chesbrough, H.W. (2003). *Open innovation: The new imperative for creating and profiting from technology*. Boston: Harvard Business School Press.

Christensen, C. M., & van Bever, D. (2014). The capitalist's dilemma. *Harvard Business Review*,

92, 60-68.

Church, G.M. (2005). The Personal Genome Project. *Molecular Systems Biology, 1*, 2005.0030. doi:10.1038/msb4100040

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences, 14*, 119–135.

Clauset, A., Arbesman, S., & Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science Advances, 1*, e1400005.

Clemente, F. (1973). Early career determinants of research productivity. *American Journal of Sociology, 79*(2), 409-419.

Collins, K. (2014). Why researchers keep citing retracted papers. *Quartz*. Available at: https://qz.com/583497/researchers-keep-citing-these-retracted-papers/

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenen, M., Leaver-Fay, A., Baker, D., & Popovic, Z. (2010). Predicting protein structures with a multiplayer online game. *Nature*, *466*(7307), 756–760.

Costello, J.C., Heiser, L.M., Georgii, E., Gönen, M., Menden, M.P., Wang, N.J., Bansal, M., Ammad-ud-din, M., Hintsanen, P., Khan, S.A., Mpindi, J.-P., Kallioniemi, O., Honkela, A., Aittokallio, T., Wennerberg, K., Collins, J.J., Gallahan, D., Singer, D., Saez-Rodriguez, J., Kaski, S., Gray, J.W., & Stolovitzky, G. (2014). A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology 32*, 1202–1212. https://doi.org/10.1038/nbt.2877

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., et al. (in press). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*.

Cressey, D. (2016). Bottles, bags, ropes and toothbrushes: the struggle to track ocean plastics. *Nature, 536,* 263-265.

Davis, R., Espinosa, J., Glass, C., Green, M.R., Massague, J., Pan, D., Dang, C.V., n.d. (2018). *Reproducibility Project: Cancer Biology Collection*. URL https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology (accessed 7.9.18).

DellaVigna, S., & Pope, D.G. (in press). What motivates effort? Evidence and expert forecasts. *Review of Economic Studies.*

DellaVigna, S., & Pope, D.G. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy, 126*(6), 2410–2456.

Derex, M., & Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences, 113*, 2982–2987. doi:10.1073/pnas.1518798113

Devictor, V., Whittaker, R. J., & Beltrame, C. (2010). Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, *16*, 354–362.

Donohue, J.J., & Wolfers, J. (2009). Estimating the impact of the death penalty on murder. *American Law and Economic Review, 11(2),* 249–309.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y. Nosek B.A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences, 112,* 15343-15347.

Ebersole, C.R., et al. (2018a). *Experimentally examining the consequences of preregistered analyses.* Crowdsourced research project in progress.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., Davis, W. E., Devos, T., Fletcher, M. M., German, K., Grahe, J. E., Hermann, A. D., Hicks, J. A., Honeycutt, N., Humphrey, B., Janus, B., Johnson, D. J., Joy-Gaba, J. A., Juzeler, H., Keres, A., Kinney, D., Kirshenbaum, J., Klein, R. A., Lucas, R. E., Lustgraaf, C. J. N., Martin, D., Menon, M., Metzger, M., Moloney, J. M., Morse, P. J., Prislin, R., Razza, T., Re, D. E., Rule, N. O., Sacco, D. F., Sauerberger, K., Shrider, E., Shultz, M., Siemsen, C., Sobocko, K., Sternglanz, R. W., Summerville, A., Tskhay, K. O., van Allen, Z., Vaughn, L. A., Walker, R. J., Weinberg, A., Wilson, J. P., Wirth, J. H., Wortman, J., & Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67,* 68-82.

Ebersole, C. R. , Buttrick, N., Baranski, E., Chartier, C. R., … & Nosek, B. A.. (2018b). *Many Labs 5: Can conducting formal peer review in advance improve reproducibility?* Crowdsourced replication project in progress. Retrieved from osf.io/7a6rd

Eitan, O., Viganola, D., Inbar, Y., Dreber, A., Johanneson, M., Pfeiffer, T., Thau, S., & Uhlmann, E. L. (2018). Is scientific research politically biased? Systematic empirical tests and a forecasting tournament to address the controversy. *Journal of Experimental Social Psychology*, *79*, 188-199.

Eisenman, I., Meier, W. N. & Norris, J. R. (2014). *Cryosphere, 8,* 1289–1296. http://dx.doi.org/10.5194/tc-8-1289-2014

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open

investigation of the reproducibility of cancer biology research. *eLife, 3,* Article e04333. doi:10.7554/ eLife.04333

Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology, 6(1152),* 1-4.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE, 5,* e10068.

Feyerabend, P. (1982). *Science in a free society*. London: New Left Books.

Feynman, R.P. (1974). Cargo cult science. *Engineering and Science, 37(7),* 10–13.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B.N., Johannesson, M. & Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology.*

Frank, M., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science, 7*(6) 600–604.

Freedman, L.P., Cockburn, I.M., & Simcoe, T.S. (2015). The economics of reproducibility in preclinical research. *PLoS Biology, 13(6),* e1002165. https://doi.org/10.1371/journal.pbio.1002165

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

Garcia-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and p-values in medical papers. *BMC Medical Research Methodology, 4,* 13. doi:10.1186/1471-2288-4-13.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102,* 460-465.

Giles, J. (2005). Internet encyclopedias go head to head. *Nature, 438 (7070),* 900–901.

Grahe, J. E., Brandt, M. J., IJzerman, H., Cohoon, J., Peng, C., Detweiler-Bedell, B., ... & Weisberg, Y. (2015, April 14). *Collaborative Replications and Education Project* (CREP). Retrieved from osf. io/wfc6u

Grahe, J. E., Reifman, A., Herman, A., Walker, M., Oleson, K., Nario-Redmond, M. & Wiebe, R. (2012). Harnessing the undiscovered resource of student research projects. *Perspectives on Psychological Science, 7*(6) 605–607.

Greenland, P., & Fontanarosa, P.B. (2012). Ending honorary authorship. *Science, 337*(6098), p1019. DOI: 10.1126/science.1224988

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*, 216-229.

Gura, T. (2013). Citizen science: amateur experts. *Nature*, *496*(7444), 259–261.

Haklay, M. (2015). *Citizen science and policy: A European perspective.* Woodrow Wilson International Center for Scholars.

Hand, E. (2010). Citizen science: People power. *Nature, 466,* 685–687.

Hauhart, R. C., & Grahe, J. E. (2012). A national survey of American higher education capstone practices in sociology and psychology. *Teaching Sociology, 40,* 227–241. doi:10.1177/00920 55X12441715

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33,* 61-83.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences, 104*(49), 19193–19198.

Hofer, S. M., & Piccinin, A. M. (2009). Integrative data analysis through coordination of

measurement and analysis protocol across independent longitudinal studies. *Psychological Methods, 14*(2), 150–164. https://doi.org/10.1037/a0015566

Holden, G., Rosenberg, G., Barker, K., & Onghena, P. (2006). An assessment of the predictive validity of impact factor scores: implications for academic employment decisions in social work. *Research on Social Work Practice, 16,* 613-24.

Horton, J. (2010). Online labor markets. In *Workshop on Internet and Network Economics* (pp. 515–522).

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6), Retrieved from http://www.wired.com/wired/archive/14.06/crowds.html

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine.* http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0020124

Ioannidis, J. P. A., & Trikalinos T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4,* 245-253.

Ioannidis, J.P.A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19,* 640–648.

Ioannidis, J.P.A. (2014). How to make more published research true. *PLoS Med 11*(10): e1001747. https://doi.org/10.1371/journal.pmed.1001747. Available at: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001747

Ioannidis, J.P., Tarone, R., & McLaughlin, J.K. (2011) The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*, *22*, 450–456.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin, 21,* 1161–1166.

Jia, M., Ding, I., Falcão, H., Schweinsberg, M., Chen, Y., Pfeiffer, T., Collet, F., Forsell, E.,

Johannesson, M., Dreber, A., Tierney, W.,… Parker, P…. & Uhlmann, E.L. (2018). *The crowdsourced generation, evaluation, and testing of research hypotheses.* Crowdsourced project in progress.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54-69.

Kanefsky, B., Barlow, N.G., & Gulick, V.C. (2001). Can distributed volunteers accomplish massive data analysis tasks? *Proceedings of the Lunar and Planetary Science Conference XXXII.*

Kerson, R. (1989). Lab for the environment. *MIT Technology Review, 92*(1), 11–12.

Kidwell, M.C., Lazarević, L.B., Baranski, E., Hardwicke, T.E., Piechowski, S., Falkenberg, L-S, et al. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Kim, J. S., et al. (2014). Space–time wiring specificity supports direction selectivity in the retina. *Nature*, *509*, 331–336.

Kittur, A., & Kraut, R. E. (2008). Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In Proceedings of CSCW (2008). ACM, New York.

Kittur, A., Lee, B., & Kraut, R. E. (2009). Coordination in collective intelligence: The role of team structure and task interdependence. In Proceedings of CHI (2009). ACM, New York.

Kittur, A., Smus, B., & Kraut, R. E. (2011). CrowdForge: Crowdsourcing complex work.

Technical Report CMU-HCII-11-100, http://crowdresearch.org/blog/wpcontent/.../CrowdForge_UIST_2011.pdf

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M., Joy-Gaba, J. A., Kappes, H. B., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., van't Veer, A., Vaughn, L. A., Vranka, M., Wichman, A., Woodzicka, J. A., & Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45(3),* 142–152.

Klein, R.A., Vianello, M., Hasselman, F.,.. Nosek, B.A. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*.

Kniffin, K. M., & Hanks, A. S. (2018). The trade-offs of teamwork among STEM doctoral graduates. *American Psychologist, 73*(4), 420-432. http://dx.doi.org/10.1037/amp0000288

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment, 14(10),* 551–560.

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass, 7,* 315-330.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J. L., Joy-Gaba, J. A., ... & Nosek, B. A.

(2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General, 143*, 1765-1785.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., ... & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*, 1001-1016.

Lakhani, K. R., Jeppesen, L. B., Lohse, P. A., & Panetta, J. A. (2007). *The value of openness in scientific problem solving.* Division of Research, Harvard Business School.

Landry, J., (2000). Profiting from open source. *Harvard Business Review blog*. Available at: https://hbr.org/2000/09/profiting-from-open-source. doi:10.1225/F00503.

Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Ebersole, C. R., Dreber, A., Johanneson, M., Pfeiffer, T., . . . & Uhlmann, E. L. (2018). *Crowdsourcing hypothesis tests.* Manuscript under review.

Landry, J., (2000). Profiting from Open Source. *Harvard Business Review blog*. Available at: https://hbr.org/2000/09/profiting-from-open-source. doi:10.1225/F00503.

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers in social psychology: Social judgment and decision making* (pp. 227-242). New York: Psychology Press.

LeBel, E. P., McCarthy, R., Earp, B., Elson, M., & Vanpaemel, W. (in press). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*

Lindwall, M., Cimino, C.R., Gibbons, L.E., Mitchell, M., Benitez, A., Brown, C.L., Kennison, R.F., Shirk, S.D., Atri, A., Robitaille, A., MacDonald, S.W.S., Zelinski, E., Willis, S.L., Schaie, K.W., Johansson, B., Praetorius, M., Dixon, R.A., Mungas, D.M., Hofer, S.M. &

Piccinin, A.M. (2012). Dynamic associations of change in physical activity and change in cognitive function: Coordinated analyses of four longitudinal studies. *Journal of Aging Research*. Vol 2012, Article ID 493598, 12 pages. doi:10.1155/2012/493598

List, B. (2017). Crowd-based peer review can be good and fast. *Nature, 546*, 9. https://doi.org/10.1038/546009a

Littmann, M., & Suomela, T. (2014). Crowdsourcing, the great meteor storm of 1833, and the founding of meteor science. *Endeavour, 38(2),* 130–138.

Lorenz, J., Rauhut, H., Schweitzer, F. & Helbing, D. (2011) How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108*, 9020–9025.

Luborsky, L., Diguer, L., Seligman, D.A., Rosenthal, R., Krause, E.D., Johnson, S., et al. (1999). The researcher's own therapy allegiances: A 'wild card' in comparisons of treatment efficacy. *Clinical Psychology: Science and Practice, 6*, 95–106.

Macleod, M.R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., et al. (2014). Biomedical research: increasing value, reducing waste. *Lancet, 383*, 101–104.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives in Psychological Science, 7*, 537–542.

Mannes, A.E., Soll, J.B., & Larrick, R.P. (2014) The wisdom of select crowds. *Journal of Personality and Social Psychology, 107*(2), 276–299.

Manzoli, L., Flacco, M.E., D'Addario, M., Capasso, L., DeVito, C., Marzuillo, C., et al. (2014). Non-publication and delayed publication of randomized trials on vaccines: survey. *British Medical Journal, 348*, g3058.

Margolin, A.A., Bilal, E., Huang, E., Norman, T.C., Ottestad, L., Mecham, B.H., Sauerwine, B.,

Kellen, M.R., Mangravite, L.M., Furia, M.D., Vollan, H.K.M., Rueda, O.M., Guinney, J., Deflaux, N.A., Hoff, B., Schildwachter, X., Russnes, H.G., Park, D., Vang, V.O., Pirtle, T., Youseff, L., Citro, C., Curtis, C., Kristensen, V.N., Hellerstein, J., Friend, S.H., Stolovitzky, G., Aparicio, S., Caldas, C., & Borresen-Dale, A.-L. (2013). Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Science Translational Medicine 5,* 181re1-181re1. https://doi.org/10.1126/scitranslmed.3006112

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist, 63,* 160–168.

Mavor, D., Barlow, K., Thompson, S., Barad, B.A., Bonny, A.R., Cario, C.L., Gaskins, G., Liu, Z., Deming, L., Axen, S.D., Caceres, E., Chen, W., Cuesta, A., Gate, R.E., Green, E.M., Hulce, K.R., Ji, W., Kenner, L.R., Mensa, B., Morinishi, L.S., Moss, S.M., Mravic, M., Muir, R.K., Niekamp, S., Nnadi, C.I., Palovcak, E., Poss, E.M., Ross, T.D., Salcedo, E.C., See, S.K., Subramaniam, M., Wong, A.W., Li, J., Thorn, K.S., Conchúir, S.Ó., Roscoe, B.P., Chow, E.D., DeRisi, J.L., Kortemme, T., Bolon, D.N., Fraser, J.S., 2016. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife 5*. https://doi.org/10.7554/eLife.15802

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., Orthey, R.,.... & Yildiz, E. (2017). Registered Replication Report: Srull & Wyer (1979). *Advances in Methods and Practices in Psychological Science.* Manuscript under review.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... & Murray, T. (2014).

Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25,* 1106-1115.

Merton, R.K. (1968). The Matthew effect in science. *Science, 159*(3810), 56–63.

Mogil, J.S., & Macleod, M.R. (2017). No publication without confirmation. *Nature, 542*(7642), 409–411. pmid:28230138

Monin, B., Pizarro, D., & Beer, J. (2007). Deciding vs. reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology, 11*(2), 99-111.

Monin, B., & Oppenheimer, D.M. (2014). The limits of direct replications and the virtues of stimulus sampling [Commentary on Klein et al., 2014]. *Social Psychology, 45*, 299-300.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … & Chartier, C. R. (in press). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science.*

Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G.G. (in press). Replication studies in economics—How many and which papers are chosen for replication, and why? *Research Policy*.

Muffatto, M. (2006). *Open source: A multidisciplinary approach.* Imperial College Press.

Mynatta, C.R., Dohertya, M.E., & Tweneya, R.D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*, 85–95.

Nielson, M. (2011). *Reinventing discovery: The new era of networked science*. Princeton University Press.

Nickerson, R.S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of*

General Psychology, 2, 175–220.

Norenzayan, A., & Heine, S. J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin, 135*, 763-784.

Nosek, B. A., (2017). Are reproducibility and open science starting to matter in tenure and promotion review? *Center for Open Science blog.* Retrieved: https://cos.io/blog/are-reproducibility-and-open-science-starting-matter-tenure-and-promotion-review/

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T. A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Levy Paluck, E., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., & Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 1422-1425. Doi: 10.1126/science.aab2374

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Math = male, Me = female, therefore math = Me. *Journal of Personality and Social Psychology, 83,* 44–59.

Nosek, B.A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry 23*, 217–223.

Nosek, B.A., Errington, T.M. (2017). Making sense of replications. *eLife 6*. https://doi.org/10.7554/eLife.23383

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*, 137-141.

Nosek, B. A., Ebersole, C. R., DeHaven, A., & Mellor, D. M. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 201708274. Doi: 10.1073/pnas.1708274114

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615-631.

O'Donnell, M., et al. (in press). Registered Replication Report: Dijksterhuis & van Knippenberg (1998). *Perspectives on Psychological Science.*

Olmsted, D. (1934). Observations on the Meteors of November 13th, 1883. *American Journal of Science and Arts, 26*(1), 354-411.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). DOI: 10.1126/science.aac4716

Pan, R.K., Petersen, A.M., Pammolli, F., & Fortunato, S. (2016). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics, 12*(3), 656–678.

Panagiotou, O.A., Willer, C.J., Hirschhorn, J.N., & Ioannidis, J.P. (2013). The power of meta-analysis in genome-wide association studies. *Annual Review of Genomics and Human Genetics, 14*, 441–465.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*, 411-419.

Petersen, A.M., Jung, W.-S., Yang, J.-S., & Stanley, H. E. (2011). Quantitative and empirical demonstration of the Matthew Effect in a study of career longevity. *Proceedings of the National Academy of Sciences, 108*(1), 18–23.

Piccinin, A. M., Muniz, G., Clouston, S.A., Reynolds, C.A., Thorvaldsson, V.Deary, I., Deeg,

D., Johansson, B., MacKinnon, A., Spiro, A. III, Starr, J. M., Skoog, I. & Hofer, S. M. (2013). Integrative analysis of longitudinal studies on aging: Coordinated analysis of age, sex, and education effects on change in MMSE scores. *Journal of Gerontology: Psychological Sciences, 68(3),* 374-390. doi: 10.1093/geronb/gbs077.

Poetz, M. K., & Schreier, M. (2012). The value of crowdsourcing: Can users really compete with professionals in generating new product ideas? *Journal of Product Innovation Management, 29*(2), 245-256.

Polymath, D. H. J. (2012). A new proof of the density Hales-Jewett theorem. *Annals of Mathematics*, *175*(3), 1283–1327.

Polymath, D. H. J. (2014). New equidistribution estimates of Zhang type. *Algebra & Number Theory*, *9*(8), 2067–2199.

Price, A., Turner, R., Stencel, R.E., Kloppenborg, B.K., & Henden, A.A. (2012). The origins and future of the citizen sky project. *Journal of the American Association of Variable Star Observers*, *40*, 614–617.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews. Drug Discovery, 10*, 712.

Raymond, E.S. (1999). *The cathedral and the bazaar: Musings on Linux and open source by an accidental revolutionary.* O'Reilly Media.

Reuter, M., et al. (2017). The Personal Genome Project Canada: findings from whole genome sequences of the inaugural 56 participants. *Canadian Medical Association Journal*, *190*(5), E126–E136. doi:10.1503/cmaj.171151

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331-363.

Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly*, *5*, 91–92.

Saez-Rodriguez, J., Costello, J.C., Friend, S.H., Kellen, M.R., Mangravite, L., Meyer, P., Norman, T., & Stolovitzky, G. (2016). Crowdsourcing biomedical research: Leveraging communities as innovation engines. *Nature Reviews Genetics*, *17*, 470-486.

Sakaluk, J.K., Williams, A.J., & Biernat, M. (2014). Analytic review as a solution to the misreporting of statistical results in psychological science. *Perspectives on Psychological Science*, *9*(6), 652 –660.

Salganik, M.J. (2017). *Bit by bit: Social research in the digital age*. Princeton: Princeton University Press.

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., & Walker, A.W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*, 87.

Schönbrodt, F. (2018). Hiring policy at the LMU Psychology Department: Better have some open science track record. *Nicebread*. Retrieved at: http://www.nicebread.de/open-science-hiring-policy-lmu/

Schooler, J. (2014). Metascience could rescue the 'replication crisis'. *Nature*, *515*, 9.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L., Diermeier, D., Heinze, J., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., Du Plessis, C. Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., Ly, A., Marsman, M., Murase, T., Qureshi, I., Schaerer, M., Thornley, N., Tworek, C. M., Wagenmakers, E-J., Wong, L., Anderson, T., Bauman, C. W., Bedwell, W. L., Brescoll, V., Canavan, A., Chandler, J. J., Cheries, E., Cheryan, S., Cheung, F.,

Cimpian, A., Clark, M., Cordon, D., Cushman, F., Ditto, P. H., Donahue, T., Frick, S. E., Gamez-Djokic, M., Hofstein Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B. E., Hartwich, N. J., Hein, K., Inbar, Y., Jiang, L., Kellogg, T., Kennedy, D. M., Legate, N., Luoma, T. P., Maibeucher, H., Meindl, P., Miles, J., Mislin, A., Molden, D. C., Motyl, M., Newman, G., Ngo, H. H., Packham, H., Ramsay, P. S., Ray, J. L., Sackett, A. M., Sellier, A-L., Sokolova, T., Sowden, W., Storage, D., Sun, X., Van Bavel, J. J., Washburn, A. N., Wei, C., Wetter, E., Wilson, C., Darroux, S-C., & Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55-67.

Schweinsberg, M., Feldman, M., Staub, N., Prasad, V., Ravid, A., van den Akker, O., van Aert, R., van Assen, M., Goldstein, P., Tierney, W., … & Uhlmann, E. (2018). *Crowdsourcing data analysis: Gender, status, and science.* Manuscript in preparation.

Schweinsberg, M., Viganola, D., Prasad, V., Dreber, A., Johannesson, M., Pfeiffer, T., Tierney, W.T., Eitan, O…& Uhlmann, E.L. (2018). *The pipeline project 2: Opening pre-publication independent replication to the world.* Data analysis phase of large scale crowdsourced project.

Sears, D.O. (1986). College sophomores in the laboratory: Influences of a narrow data base on psychology's view of human nature. *Journal of Personality and Social Psychology, 51*, 515-530.

Sears, D.O., Sidanius, J., & Bobo, L. (2000). *Racialized politics: The debate about racism in America.* Chicago: University of Chicago Press.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal, 314*, 498-502.

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., Fong, N., Gamez-Djokic, M., Glenz, A., Gordon-McKeon, S., Heaton, T. J., Hederos, K., Heene, M., Hofelich Mohr, A. J., Högden, F., Hui, K., Johannesson, M., Kalodimos, J., Kaszubowski, E., Kennedy, D., Lei, R., Lindsay, T. A., Liverani, S., Madan, C. R., Molden, D., Molleman, E., Morey, R. D., Mulder, L. B., Nijstad, B. A., Pope, N. G., Pope, B., Prenoveau, J. M., Rink, F., Robusto, E., Roderique, H., Sandberg, A., Schlüter, E., Schönbrodt, F. D., Sherman, M. F., Sommer, S., Sotak, K., Spain, S., Spörlein, C., Stafford, T., Stefanutti, L., Tauber, S., Ullrich, J., Vianello, M., Wagenmakers, E., Witkowiak, M., Yoon, S., & Nosek, B.A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science, 9*(1), 76 –80.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False– positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24*(10), 1875-1888.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *P-Curve: A key to the file drawer*. *Journal of Experimental Psychology: General*, *143*, 534-547.

Sobel, D. (2007). *Longitude: The true story of a lone genius who solved the greatest scientific*

problem of his time. Bloomsbury Publishing.

Sørensen, J. J., Pedersen, M. K., Munch, M., Haikka, P., Jensen, J. H., Planke, T., ... & Sherson, J. F. (2016). Exploring the quantum speed limit with computer games. *Nature, 532*(7598), 210-213.

Srinarayan, S., Sugumaran, V., & Rajagopalan, B. (2002). A framework for creating hybrid-open source software communities. *Information Systems Journal, 12*, 7-25.

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (in press). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin.*

Steward, O., Popovich, P.G., Dietrich, W.D., & Kleitman, N. (2012). Replication and reproducibility in spinal cord injury research. *Experimental Neurology, 233(2),* 597-605.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences, 21*(10), 736-748.

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making, 10*(5), 479-491.

Stolovitzky, G.A., Monroe, D., & Califano, A. (2007). Dialogue on reverse-engineering assessment and methods. *Annals of the New York Academy of Sciences, 1115*, 1–22.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor Books.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.

Tao, T., Croot, E., & Helfgott, H. (2012). Deterministic methods to find primes. *Mathematics of Computation, 81*(278), 1233–1246.

Tetlock, P.E., Mellers, B., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools

for increasing transparency and the quality of debate. *Current Directions in Psychological Science, 23*(4), 290-295.

The MetaSUB Consortium (2016). The metagenomics and metadesign of the subways and urban biomes (MetaSUB) international consortium inaugural meeting report. *Microbiome*, *4*, 24-38.

Thelen, B.A., & Thiet, R. K. (2008). Cultivating connection: Incorporating meaningful citizen science into Cape Cod National Seashore's estuarine research and monitoring programs. *Park Science, 25*, 74-80.

Tierney, W., Schweinsberg, M., & Uhlmann, E. L. (in press). Making prepublication independent replication mainstream. *Behavioral and Brain Sciences* (Commentary).

Tierney, W., Schweinsberg, M., Jordan, J., Kennedy, D. M., Qureshi, I., Sommer, S. A., Thornley, N., Madan, N., Vianello, M., Awtrey, E., Zhu, L., Diermeier, D., Heinze, J., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., Du Plessis, C. Gronau, Q. F., Hafenbrack, A. C., Liao, E. Y., Ly, A., Marsman, M., Murase, T., Schaerer, M., Tworek, C. M., Wagenmakers, E-J., Wong, L., Anderson, T., Bauman, C. W., Bedwell, W. L., Brescoll, V., Canavan, A., Chandler, J. J., Cheries, E., Cheryan, S., Cheung, F., Cimpian, A., Clark, M., Cordon, D., Cushman, F., Ditto, P. H., Donahue, T., Frick, S. E., Gamez-Djokic, M., Hofstein Grady, R., Graham, J., Gu, J., Hahn, A., Hanson, B. E., Hartwich, N. J., Hein, K., Inbar, Y., Jiang, L., Kellogg, T., Legate, N., Luoma, T. P., Maibeucher, H., Meindl, P., Miles, J., Mislin, A., Molden, D. C., Motyl, M., Newman, G., Ngo, H. H., Packham, H., Ramsay, P. S., Ray, J. L., Sackett, A. M., Sellier, A-L., Sokolova, T., Sowden, W., Storage, D., Sun, X., Van Bavel, J. J., Washburn, A. N., Wei, C., Wetter, E., Wilson, C., Darroux, S-C., & Uhlmann, E. L.

(2016). Data from a pre-publication independent replication initiative examining ten moral judgment effects. *Nature: Scientific Data*, 3, 160082.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293–315.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five Years of GWAS Discovery. *American Journal of Human Genetics*, *90*(1), 7–24. http://doi.org/10.1016/j.ajhg.2011.11.029

Wagge, J.R. et al. (in press). A demonstration of the Collaborative Replication and Education Project: Replication attempts of the red-romance effect. *Collabra*.

Wahls, W.P. (2018). High cost of bias: *Diminishing marginal returns on NIH grant 3 funding to institutions.* Unpublished manuscript.

Wen, H., Wang, H.-Y., He, X., Wu, C.-I. (2018). On the low reproducibility of cancer studies. *National Science Review*. https://doi.org/10.1093/nsr/nwy021

Westra, H.-J., Jansen, R. C., Fehrmann, R. S. N., te Meerman, G. J., van Heel, D. et al. (2011). MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics, 27*, 2104–2111.

Williamson, I.O., & Cable, D.M. (2003). Predicting early career research productivity: The case of management faculty. *Journal of Organizational Behavior, 24*, 25–44.

Wolfers, J., & Stevenson, B. (2013). Six ways to separate lies from statistics. *Bloomberg View*. Available at:

http://www.bloombergview.com/articles/2013-05-01/six-ways-to-separate-lies-from-statistics

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling in social psychological experimentation. *Personality and Social Psychology Bulletin, 25,* 1115-1125.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge university press.

Wenneras, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature, 387,* 341–343.

Wolfman-Arent, A. (2014). Frustrated scholar creates new way to fund and publish academic work. *Chronicle of Higher Education*. Retrieved at: https://www.chronicle.com/blogs/wiredcampus/frustrated-scholar-creates-new-route-for-funding-and-publishing-academic-work/53073?cid=at&utm_medium=en&utm_source=at&cid=at&utm_medium=en&utm_source=at

Wuchty, S., Jones, B. F., & Uzzi, B. (2007a). The increasing dominance of teams in the production of knowledge. *Science, 316,* 1036–1038.

Wuchty, S., Jones, B., & Uzzi, B. (2007b). Why do team authored papers get cited more. *Science, 317,* 1496-1497.

Zastrow, M. (2014). Error discovered in Antarctic sea-ice record. *Nature*. Available at: http://www.nature.com/news/error-discovered-in-antarctic-sea-ice-record-1.15605

Table 1

Crowdsourcing different stages of the research process

Stage of Research	*How crowds are leveraged*
Ideation	Crowds are used to generate novel research ideas and solutions to problems
Assembling resources	Online exchanges are used to match investigators with needs with partner laboratories who have that resource
Study design	The same research hypothesis is given to different scientists, who independently design studies to test it
Data collection	Numerous collaborators aid in obtaining research participants, observations, or samples
Data analysis	A network of researchers carry out statistical analyses to address the same research question
Replicating findings prior to publication	The same methodology is repeated in independent laboratories to confirm the finding prior to its publication
Writing research reports	A large group of contributors collectively writes a research article
Peer review	A large group of commentators write public feedback on a scientific paper
Replicating published findings	The same methods and materials from published papers are repeated in independent laboratories to assess the robustness of the findings
Deciding future directions	Crowd predictions about future research outcomes are factored into decisions about how to allocate research resources for maximum impact

Table 2

Examples of Crowdsourced Scientific Initiatives

Citation	*Crowdsourced Stage*	*Method*	*Key Result(s)*
Sobel (2007)	Ideation	Starting in 1714, the British Parliament launched an open competition to solve how to calculate the longitude of a ship at sea	Development of the marine chronometer
Polymath (2012, 2014)	Ideation	Mathematical challenges are posted online for open crowd collaboration	A new combinatorial proof to the density version of the Hales–Jewett theorem, among other solved mathematical problems
Schweinsberg, Feldman, et al. (2018)	Ideation	Crowd of researchers asked to nominate hypotheses for testing with a complex dataset	The crowd was able to generate interesting hypotheses for later testing
InnoCentive.com	Ideation	Scientific problems are posted online and prizes are offered for the best solution	30% of 166 scientific problems solved via crowd competitions for prizes
Science Exchange	Assembling resources	Online marketplace that enables scientists to identify and outsource specific research needs	Program to independently validate antibodies; partnership with the Center for Open Science to conduct the Reproducibility Project: Cancer Biology
Study Swap	Assembling resources	Platform for posting brief descriptions of resources available for use by others, or needed resources another researcher may have	Used to gather resources for both crowdsourced and small team projects
Landy et al. (2018)	Study design	Independent research teams separately design experiments to test the same hypothesis; research participants are then randomly assigned to different study versions	Different study designs associated with widely dispersed effect size estimates for the same research question; for four out of five hypotheses examined, the materials from different teams returned significant effects in opposite directions
Olmstead (1834)	Data collection	In 1833 Professor Denison Olmsted used letter correspondence to recruit citizen scientists to help document a meteor shower	Detailed documentation of the great meteor storm of 1833; birth of citizen science movement
Kanefsky et al. (2001)	Data collection	Clickworkers website from NASA asks volunteers to help classify images	Mapping of craters on Mars based on images from the Viking Orbiter

Citation	*Crowdsourced Stage*	*Method*	*Key Result(s)*
Church (2005)	Data collection	The Personal Genome Project recruits everyday people willing to publicly share their personal genome, health, and trait data as a public research resource	Collection of data from 10,000 volunteers; full analyses of the genomes of 56 participants with identification of potential health impacts in 25% of cases; ongoing project to link genetics, memory, and attention
Cooper et al. (2010)	Data collection	Online game Foldit in which over 50,000 players compete to fold proteins	The best human players outperform a computer in terms of determining protein structures
Price et al. (2012)	Data collection	Citizen sky project recruits amateur astronomers help professionals gather observations of the planets, moons, meteors, comets, stars, and galaxies	Gathering observations of Epsilon Aurigae, an unusual multiple star system, among other targets
Kim et al. (2014)	Data collection	Video game EyeWire in which players reconstruct part of an eye cell using three dimensional images of microscopic bits of retinal tissue	Data from over 2,000 elite gamers used to collectively map neural connections in the retina, contributing to a better understanding of how the eye detects motion
MetaSUB Consortium (2015)	Data collection	Commuters are enlisted to obtain samples from surfaces in subways and other public areas	Identification of new species and novel biosynthetic gene clusters; global maps of antimicrobial resistance (AMR) markers
Sørensen et al. (2016)	Data collection	Video game Quantum Moves in which the player moves digital renditions of quantum atoms	The data produced by the over 200,000 users has been leveraged to develop better quantum algorithms
Moshontz et al. (2018)	Data collection	Psychological Science Accelerator (PSA), a network of over 300 laboratories to conduct replications and collect other data for crowdsourced projects	The first large scale PSA project will seek to replicate earlier findings that people rate faces based on valence and dominance
Zooniverse	Data collection	Online platform where citizen volunteers assist professional researchers with projects	Enables citizen science initiatives such as "Mapping Prejudice" in which project volunteers identify racially restrictive property deeds
Galaxy Zoo	Data collection	Galaxy Zoo website asks volunteers to help classify galaxies based on images	Collection of over 100 million classifications of galaxies based on shape, structure, and intensity; identifying supernovas and potential interactions between galaxies
Audubon Christmas Bird Count	Data collection	Beginning with the Audubon Christmas Bird Count of 1900, amateur birdwatchers have been used to collect data on bird migrations	Large dataset on bird migrations leveraged for scientific publications

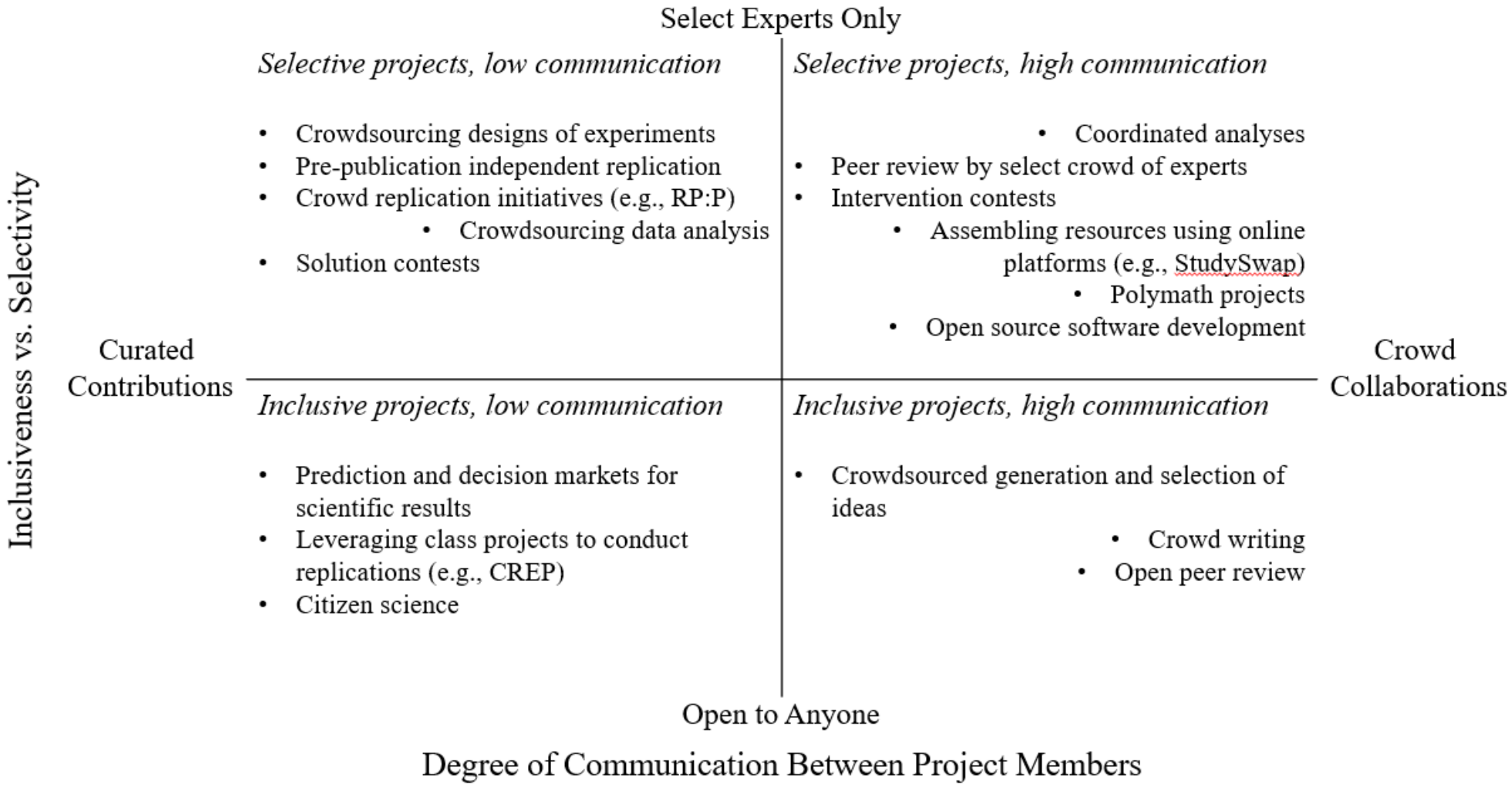
Citation	*Crowdsourced Stage*	*Method*	*Key Result(s)*
Stolovitzky et al. (2007)	Data analysis	In the DREAM (Dialogue for Reverse Engineering Assessments and Methods) Challenges, organizers provide a test data set and a particular question to be addressed to many independent analysts, then apply the analytic strategies to a hold-out dataset to evaluate their robustness	Improved prediction of survival of breast cancer patients, drug sensitivity in breast cancer cell lines, and biomarkers for early Alzheimer's disease cognitive decay
Hofer & Piccinin (2009)	Data analysis	Coordinated analysis: Network of researchers use the same target constructs, model, and covariates on different longitudinal datasets to address the same research question	Changes in physical activity over time affect cognitive function; education may not be a protective factor against cognitive decline
Schweinsberg, Feldman, et al. (2018)	Data analysis	42 analysts were asked to test hypotheses related to gender, status, and science using a complex dataset on academic debates	Radical effect size dispersion, with analysts in some cases reporting significant effects in opposite directions for the same hypothesis tested with the same data
Silberzahn et al. (2018)	Data analysis	Same dataset was distributed to 29 analysis teams, who separately analyzed it to address the same research question (do soccer referees give more red cards to dark skin toned players than light skin toned players?)	Effect size estimates ranging from slightly negative to large positive effects; 69% of analysts reported statistically significant support for the hypothesis and 31% reported nonsignificant results
Schweinsberg et al. (2016)	Replicating findings prior to publication	25 independent laboratories attempted to replicate 10 unpublished findings from one research group	6 of 10 findings were robust and generalizable across cultures according to the pre-registered replication criteria
Christensen & van Bever (2014)	Writing research reports	Online collaboration platform used to collect ideas and comments regarding why companies often do not invest in innovations that create new markets	The article "The Capitalist's Dilemma" which argues this occurs because companies incentivize their managers to find efficiency innovations that eliminate jobs and pay off fast, rather than market innovations that pay off years later
List (2017)	Peer review	*Synlett* implemented a crowdsourced reviewing process to allow over 100 referees to respond to papers after they were posted to an online forum for reviewers	The crowd review was faster and provided more comprehensive feedback than the traditional peer-review process
Steward et al. (2012)	Replicating published findings	Initiative to replicate spinal cord injury research in independent laboratories	2 successful replications out of 12 targeted studies

Citation	*Crowdsourced Stage*	*Method*	*Key Result(s)*
Alogna et al. (2014)	Replicating published findings	Registered Replication Report: Attempt by many laboratories to replicate the verbal overshadowing effect	Verbal overshadowing successfully replicated, but with a smaller effect size than in the original paper
Klein et al. (2014)	Replicating published findings	Many Labs 1: 36 laboratories attempted to replicate 13 psychology findings	10 of 13 findings replicated
Open Science Collaboration (2015)	Replicating published findings	Reproducibility Project: Psychology attempted to replicate 97 original effects from top psychology journals in independent laboratories	36% of findings successfully replicated
Camerer et al. (2016)	Replicating published findings	Experimental Economics Replication Project: Initiative to replicate prominent findings in experimental economics in independent laboratories	61% of findings successfully replicated
Ebersole et al. (2016)	Replicating published findings	Many Labs 3: 20 laboratories attempted to replicate 10 psychology findings at different times of the semester	3 of 10 findings replicated; most unaffected by time of semester
McCarthy, et al. (2017)	Replicating published findings	Registered Replication Report: Attempt by many laboratories to replicate the effects of priming hostility on impression formation	Failure to replicate the hostility priming effect, with low heterogeneity in effect sizes across laboratories
Nosek & Errington (2017)	Replicating published findings	Reproducibility Project: Cancer Biology, an initiative to replicate prominent findings in cancer biology	Of 12 replications thus far, 4 reproduced important parts of the original paper, 4 replicated some parts of the original paper but not others, 2 were not interpretable, and 2 did not replicate the original findings
Camerer et al. (2018)	Replicating published findings	Social Sciences Replication Project, an initiative to replicate 21 social science findings in *Science* and *Nature*	13 (62%) of findings successfully replicated
Klein et al. (2018)	Replicating published findings	Many Labs 2: 28 psychology findings replicated across 125 sites	14 of 28 findings replicated; heterogeneity in effect size estimates was highest for large effect sizes, and low for non-replicable effects
Cova et al. (in press)	Replicating published findings	Initiative to replicate prominent findings in experimental philosophy in independent laboratories	78% of findings successfully replicated

Citation	*Crowdsourced Stage*	*Method*	*Key Result(s)*
O'Donnell et al. (in press)	Replicating published findings	Registered Replication Report: Attempt by many laboratories to replicate the effect of priming professors on intellectual performance	Failure to replicate the professor priming effect, with low heterogeneity in effect sizes across laboratories
Wagge et al. (in press)	Replicating published findings	Collaborative Replications and Education Project (CREP) initiative to replicate social psychology findings in student methods classes	This project fails to replicate earlier findings that women are more attracted to men in photographs with red borders
Dreber et al. (2015)	Deciding future directions	Prediction market to see if independent scientists could forecast the results of the Reproducibility Project: Psychology	Aggregated predictions accurately anticipated replication results
Camerer et al. (2016)	Deciding future directions	Prediction market to see if independent scientists could forecast replication results in experimental economics	Aggregated predictions accurately anticipated replication results
DellaVigna & Pope (2018)	Deciding future directions	Prediction survey to see if forecasters could anticipate the effects of treatment conditions on worker productivity	Aggregated predictions anticipated research outcomes; expert behavioral scientists, doctoral students, and Mechanical Turk workers similarly accurate
Eitan et al. (2018)	Deciding future directions	Prediction survey to see if scientists could forecast the size of political biases in scientific abstracts, and to gauge their reactions to the research results	Forecasters accurately predicted that conservatives would be explained more, and explained in more negative terms, in scientific abstracts in social psychology. They also significantly overestimated the size of both effects, but updated their beliefs in light of the new evidence.
Landy et al. (2018)	Deciding future directions	Prediction survey to see if independent scientists could predict the results of conceptual replications	Aggregated predictions accurately anticipated overall outcomes, including variability in results across different study designs testing the same hypothesis
Camerer et al. (in press)	Deciding future directions	Prediction market to see if independent scientists could forecast results replications of social science papers in *Science* and *Nature*	Aggregated predictions accurately anticipated replication results
DellaVigna & Pope (in press)	Deciding future directions	Prediction survey to see if forecasters could anticipate the effects of treatment conditions on worker productivity as well as moderation by their demographic characteristics	Aggregated predictions anticipated treatment effects, but overestimated importance of demographic moderators; academic seniority did not moderate forecasting accuracy

Citation	*Crowdsourced Stage*	*Method*	*Key Result(s)*
Forsell et al. (in press)	Deciding future directions	Prediction market to see if independent scientists could predict the results of the Many Labs 2 replication initiative	Aggregated predictions accurately anticipated replication results
Lai et al. (2014, 2016)	Deciding future directions	Contest to identify the most effective intervention to reduce implicit preferences for Whites over Blacks	8 of 17 interventions effective in the short term, but none effective a day or more after the intervention. Teams were able to iteratively improve their interventions between rounds

Figure 1 – Forms and Examples of Crowdsourcing



Note: *Curated Contributions* refer to projects where project coordinators collect the individual work of a crowd of contributors whose communication with one another is limited to nonexistent. *Crowd Collaborations* refer to projects in which a large group of contributors engage in regular communication regarding their shared work.

Supplement 1: The growth of crowdsourcing

Crowdsourcing can involve combining the work of many individuals addressing small components of a larger problem, posing an open call for solutions to a challenge, scaling up data builds by distributing the work of collecting observations across numerous contributors, or aggregating the predictions or recommendations of a large group of people (Salganik, 2017; Surowiecki, 2005). Some examples of scientific work distribution to a large set of individuals date back well over a century. In 1714, the British Parliament announced an open competition for the best method to determine the longitude of a ship at sea – the winner, the marine chronometer, was the invention of John Harrison, a previously unknown clockmaker (Sobel, 2007). Professor Denison Olmsted used letter correspondence to carry out a collective effort to document the great meteor storm of 1833 (Littmann & Suomela, 2014; Olmsted, 1934). The Audubon Christmas Bird Count of 1900 organized an army of amateur bird watchers, a practice that continues to this day (Butcher, 1990). More recently, crowdsourcing activities in the for-profit and not-for-profit sectors have grown exponentially as the internet has eroded barriers to global communication and collaboration (Brabham, 2013; Chesbrough, 2003; Muffatto, 2006; Raymond, 1999). For example, nonprofit initiatives have organized volunteers to create common goods, such as encyclopedias (e.g., Wikipedia.org) and searchable genealogy databases (e.g., FamilySearch.org). Crowdsourcing science is part of a global movement towards expanded online collaborative networks.

Open competitions have been used by private companies to generate ideas (Poetz & Shreier, 2012) and solve scientific problems (Brabham, 2010). Websites such as InnoCentive.com organize contests in which a preset payment is awarded to the best solution to a problem. A study of 166 unsolved discrete scientific problems posted at InnoCentive.com (such as finding "a stable form of tetrasodium pyrophosphate") found that 30% of these problems were effectively solved, challenges that large and well-known R&D-intensive firms had been unsuccessful in solving internally. Notably, intrinsic motivation to crack a tough problem turned out to be an even stronger predictor of being a winning solver than the desire to win the award (Lakhani et al., 2007).

A model case of an ecosystem that has embraced the value of open collaboration and innovation is the open source software community. In contrast to traditional proprietary software, software and code is made available to anyone for modification and use, with new developments happening online publicly through an open collaboration process (Muffatto, 2006). The movement towards open software has roots in projects from the 1980s (Raymond, 1999) and gained prominence in the late 1990's. Examples include Netscape communicator, Mozilla Firefox, Android, the iOS Software Development Kit, the Apache HTTP Server, and Linux. A model in which users not only access and distribute the software for free, but can even help create it, aims to increase adoption and loyalty and can speed up innovation and improvement of the software. Preliminary versions are often released early in the development process to find collaborators and solve problems (Srinarayan, Sugumaran, & Rajagopalan, 2002). Even large for-profit companies including Microsoft, IBM Google, and Hewlett-Packard have developed an open source presence, with the goals of promoting the company's image and lowering marketing costs (Landry, 2000). The open source software community may preview science's future.

Supplement 2: More detailed descriptions of specific crowdsourced projects

Below are more detailed descriptions of some specific crowdsourced projects (see also Table 2). These are organized by the stage of the research process the crowd's efforts were focused on (see Table 1).

Ideation

In 2009, Cambridge University mathematician Tim Gowers experimented with crowdsourced idea generation by posting a mathematical challenge on his blog – "find a new combinatorial proof to the density version of the Hales–Jewett theorem" – and soliciting suggestions from anyone on how to solve it (Ball, 2014). After seven weeks and over 1000 comments from more than 40 colleagues, Gowers declared the problem largely solved, although some additional work was needed prior to the completion of the proof and publication of the article in the *Annals of Mathematics* (credited to "Polymath, 2012"). The ongoing Polymath Project poses further unsolved mathematical challenges online for crowd collaboration, resulting in more published articles (e.g., Polymath, 2014), and even when unsuccessful at producing a full solution, sometimes generating ideas that contribute to other proofs (e.g., Tao, Croot, & Helfgott, 2012).

Schweinsberg, Feldman, et al. (2018) asked a group of colleagues, recruited via an open call online, to nominate hypotheses for testing with a complex dataset on the role of gender, status, and science in intellectual debates. A second survey then asked scientists to rate each idea for its likelihood of finding empirical support, theoretical interest value if true, and overall scientific worth. Hypotheses generated by the crowd were rated as just as high in quality as those the project coordinators had initially planned to test with the data.

An in progress initiative to Crowdsource the Generation, Evaluation, and Testing (CGET) of research ideas will leverage a proprietary dataset that cannot be distributed beyond the project coordination team. A data descriptor will be posted online and an open call made for interesting hypotheses that could be tested with the available variables. A decision market will then be used to select which hypotheses to pursue. The analyses will then be carried out by the project coordinators with the hypothesis-proposers as coauthors on the final report (Jia et al., 2018).

Assembling resources

Science Exchange (scienceexchange.com) is an online marketplace of research services that enables scientists to outsource their research and development. Researchers can search from thousands of qualified service providers, such as university shared facilities or commercial contract research organizations (CROs), to identify and outsource specific experimental needs. The marketplace has been used to independently validate antibodies, and (in a partnership with the Center for Open Science) to conduct the Reproducibility Project: Cancer Biology.

StudySwap (http://osf.io/view/StudySwap/) is a platform for posting brief descriptions of resources available for use, or needed resources another researcher may have. Examples of research resources that could be exchanged are the capacity to collect data for another researcher,

access to a hard-to-reach population of participants, or access to specialized equipment. In its first year, StudySwap has been used for a diverse set of research resource exchanges. Coordinators of the Pipeline Project 2 (Schweinsberg, Tierney et al., 2018) and Many Labs 5 (Ebersole et al., 2018b) successfully recruited numerous labs to join these crowdsourcing replication initiatives. A researcher in Malaysia found a collaborating lab in the Netherlands to test the cultural generalizability of an educational psychology finding. As another example, a researcher in the United Kingdom, who was without data collection capacity for a time period, found a lab in the United States to collect data for an idea. This active and eclectic opening year and a half bodes well for the potential of StudySwap to facilitate widespread research resource exchange.

Study design

Landy et al. (2018) compiled five unpublished effects related to moral judgments, negotiations, and intergroup attitudes that had used a single operationalization each. The associated research questions were then posed to up to a dozen additional research teams who independently designed studies to test each question (e.g., "Is a utilitarian vs. deontological moral orientation related to personal happiness?", "Are people aware of their automatic prejudices?", "Does working despite no material need to do so elicit moral praise?"). Participants were randomly assigned to one of the multiple operationalizations testing the same question. Variability in estimated effect sizes attributable to design choices was substantial. Operationalizations for four out of the five questions elicited significant effect sizes in opposite directions. Aggregating across different study versions via meta-analysis revealed strong support for two hypotheses and a lack of overall support for three hypotheses. Contrary to the concept of researcher "flair" or talent leading some investigators to obtain empirical support for predictions where others fail (Baumeister, 2016), no team produced consistently larger effect sizes than any other. Rather, variability in effect sizes was attributable to whether the hypothesis was supported overall or not and subjective design choices by the researchers. Notably, all five target hypotheses directly replicated using the original study materials (Landy et al., 2018). If the standard approach to science had been applied, all five hypotheses, rather than the two supported in the crowdsourced conceptual replications, would have been considered supported.

Data collection

Opening participation in projects to the public via citizen science initiatives has had the most impact in biology, astronomy, ecology, and conservation, but is spreading to other fields. Amateur astronomers help professionals gather observations of the planets, moons, meteors, comets, stars, and galaxies (Price, Turner, Stencel, Kloppenborg, & Henden, 2012), and members of the general public aid in classifying images in huge research databases (e.g., NASA's Clickworkers and Galaxy Zoo; Hand, 2010; Kanefsky, Barlow, & Gulick, 2001). Over a quarter million amateur bird watchers and butterfly watchers are relied on to document animal populations and migrations (ebird.org; Cavalier & Kennedy, 2016; Devictor, Whittaker, & Beltrame, 2010), a mobile app is used by thousands of boat-goers to track water debris (Cressey, 2016), commuters are enlisted to obtain samples from surfaces in subways and other public areas in order to map a city's microbiome (Afshinnekoo et al., 2015; The MetaSUB Consortium,

2015), and armies of volunteers collect rain samples to facilitate research on pollution (Haklay, 2015; Kerson, 1989).

The Personal Genome Project is a coalition of projects around the globe aimed at everyday people willing to publicly share their personal genome, health, and trait data as a public research resource (personalgenomes.org; Church, 2005; Reuter et al., 2017; see also the uBiome and American Gut projects; Afshinnekoo et al., 2016). This approach has been expanded to Open Humans (openhumans.org), a platform that allows citizen volunteers to upload and privately store their personal data (e.g. genetic, activity, or social media), which can be shared publicly or with specific research projects. Zooniverse, launched in 2009, is another platform where citizen volunteers assist professional researchers. The platform hosts various research projects, and users are able to select and participate. As of July 2018, there were 88 active and 11 finished projects. One example of a Zooniverse crowdsourced project is "Mapping Prejudice" where project volunteers view Minneapolis property deeds, identify racially restrictive deed covenants, and affiliated covenant addresses are then mapped.

In addition to helping collect scientific observations, citizen scientists can aid in cracking scientific problems. In the video game Quantum Moves, the player moves digital renditions of quantum atoms, with the data produced by the over 200,000 users and 8 million plays leveraged to develop better quantum algorithms (Sørensen et al., 2016). Over 2,000 elite gamers on the website EyeWire reconstructed part of an eye cell using three dimensional images of microscopic bits of retinal tissue, collectively mapping neural connections in the retina and contributing to a better understanding of how the eye detects motion (Kim et al., 2014). In the online game Foldit, over 50,000 players compete to fold proteins, with the best players outperforming a computer in terms of determining protein structures (Cooper et al., 2010). Such gamification of science holds the potential to recruit armies of online volunteers to facilitate discoveries.

Data analysis

Silberzahn et al. (2018) distributed the same archival dataset to 29 analysis teams, asking them each to test whether dark skin toned football (soccer) players were more likely than light skin toned players to receive red cards from referees. No two specifications were exactly alike, with the crowd of analysts employing diverse statistical perspectives and choices of covariates. The range of effect sizes from different teams of scientists spanned from directionally negative and non-significant, to positive, large, statistically significant effects. If the analysis and presentation of the results were handled by a single vertically integrated research team, there would have been a 69% probability of significant support for the hypothesis being reported, and a 31% chance of a nonsignificant effect.

In a second crowdsourcing data analysis initiative, 42 analysts were asked to test hypotheses related to gender, status, and science using a complex dataset on academic debates (Schweinsberg, Feldman et al., 2018). The first hypothesis posited that female scientists participate more in intellectual conversations with a greater number of women, and the second that higher status academics are more verbose than are lower status academics. Each researcher decided not only her or his preferred statistical approach, but also how to operationalise key

variables. For example, volubility could be operationalised as number of words spoken or number of times speaking; status could be measured using citation counts, job rank, university rank, or some combination. Under these conditions, which arguably more closely mimic those of the typical research project, effect size estimates proved radically dispersed, with different analysts in some cases reporting significant effects in opposite directions for the same hypothesis tested with the same data. This raises the unsettling possibility that even in the absence of perverse incentives and directional motives, analytical choices may have as great an effect on research conclusions as whether the hypothesis is true. Only a crowdsourced approach can make transparent the full extent to which research conclusions are contingent on the subjective decisions made by different analysts.

The DREAM (Dialogue for Reverse Engineering Assessments and Methods) Challenges have been used to evaluate model predictions and pathway inference algorithms in systems biology and medicine (dreamchallenges.org). These include predicting survival of breast cancer patients based on clinical information about the patient's tumor and genome-wide molecular profiling data (Margolin et al., 2013), integrating multiple-omics measurements and predicting drug sensitivity in breast cancer cell lines (Costello et al., 2014), and predicting the best biomarkers for early Alzheimer's disease cognitive decay from genetic or structural imaging data (Allen et al., 2016).

Replicating findings prior to publication

In the first Pipeline Project, twenty-five independent laboratories attempted to replicate 10 unpublished findings from one research group, collecting over eleven thousand research participants from half a dozen countries (Schweinsberg et al., 2016). Six of the findings were robust and generalizable across cultures according to the pre-registered replication criteria. This modest reproducibility rate even under the best of conditions suggests that failed replications are an unavoidable aspect of science. It also shows that organizing independent replications of unpublished work is a pragmatically achievable goal.

Writing research reports

In one recent initiative, 150 Harvard MBA students and alumni from Professor Clayton Christensen's course "Building and Sustaining a Successful Enterprise" used an online collaboration platform to post and comment on ideas regarding why companies often do not invest in innovations that create new markets. The end result is the well-cited article "The Capitalist's Dilemma" in *Harvard Business Review*, which argues this occurs because companies incentivize their managers to find efficiency innovations that eliminate jobs and pay off fast (in 1-2 years), rather than market creating innovations that bring in new types of customers and open novel markets but take 5 to 10 years to have impact (Christensen & van Bever, 2014). The published version features a visual map of how ideas emerged, merged, and diverged in the crowd before they arrived at the final article.

Peer review

Experimentation with peer review is emerging with some staying close in important respects to traditional peer review and others departing radically. The chemical-synthesis journal *Synlett* implemented a crowdsourced reviewing process to allow over 100 highly qualified referees, mostly suggested by the editorial board, to respond to papers after they were posted to a protected online forum for reviewers. The crowd review was faster – three days versus weeks – and collectively provided more comprehensive feedback than the traditional peer-review process (List, 2017). The *Living Reviews* group of journals in physics allow authors to update their articles in response to peer review feedback (https://www.springer.com/gp/livingreviews). An innovative multi-stage approach at *Atmospheric Chemistry and Physics* begins with an open crowd review, and then moves on to assessments by select reviewers invited by the editor.

Some aspects of an open commenting system are also emerging, such as the integration of the annotating service Hypothesis with the journal *eLife*, as well as PsyArXiv (http://psyarxiv.org/), SocArXiv (http://socarxiv.org/), and other preprint servers hosted on the Open Science Framework (OSF).

Replicating published findings

In the Many Labs and Registered Replication Report initiatives, a dozen laboratories or more each attempt to replicate published findings such as heuristics and biases in judgment, gender differences in attitudes towards mathematics, and nonconscious priming effects on behaviour (e.g., Alogna et al., 2014; Klein et al., 2014). Another approach, designed to capture a greater number of original studies, is to assign each original study to only one other laboratory, as in the Reproducibility Project: Psychology (Open Science Collaboration, 2015), Reproducibility Project: Cancer Biology (Errington et al., 2014), and the Social Sciences Replication Project (Camerer et al., 2018) typically collecting a larger sample in the replication study to provide improved statistical power to detect the effect.

These efforts have generally yielded disappointing results. In the Reproducibility Project: Psychology, 35 (36%) of the original 97 effects from top psychology journals produced significant effects ($p < .05$) in the expected direction in the more highly powered replications. Although original and replication effect sizes were significantly correlated, replication effect sizes were also systematically lower than in the original papers. Earlier efforts by pharmaceutical companies to replicate a total of 120 landmark biomedical studies (53 by Amgen and 67 by Bayer) obtained reproducibility rates of 11-25% (Begley & Ellis, 2012; Prinz, Schlange & Asadullah, 2011). In the ongoing Reproducibility Project: Cancer Biology, 12 replications have been published to date, with editors at the publishing journal *eLife* determining that 4 replicated important parts of the original paper, 4 replicated some parts of the original paper but not others, 2 were not interpretable, while 2 did not replicate the original findings (Davis et al., 2018; cf. Wen et al., 2018). An effort among academics to replicate spinal cord injury research obtained six null results, 3 mixed results, an inconclusive outcome, and two successful outcomes out of 12 studies (Steward, Popovich, Dietrich, & Kleitman, 2012). Although direct comparisons cannot be made with any confidence due to differences in sampling and methodology, other replication

initiatives obtained reproducibility rates of 61% in experimental economics (Camerer et al., 2016), and 78% in experimental philosophy (Cova et al., in press).

Some previously celebrated findings in psychology, such as demonstrations of nonconscious priming effects on judgments and behaviors (see Bargh, 1997, 2014, for reviews), have consistently yielding effect size estimates close to zero in replication studies (e.g., Klein et al., 2014; O'Donnell et al. in press; McCarthy, et al., 2017). Earlier findings that unscrambling sentences related to hostility leads a target person to be perceived as hostile, exposure to images of the national flag impacts political attitudes, and activating thoughts about professors increases performance on general knowledge questions were not obtained in independent laboratories. There are many reasons why an effect may fail to replicate other than it being a false positive – replicator error, lack of fidelity to the original study, and unidentified moderators, among others – yet these accumulating null findings suggest that, if the original effects are true positives, the eliciting conditions are not yet understood and reliably demonstrable. At the same time, other well known findings – such as anchoring (Jacowitz & Kahneman, 1995), gain vs. loss framing (Tversky & Kahneman, 1981), question framing (Rugg, 1941), and gender differences in implicit and explicit math attitudes (Nosek, Banaji, & Greenwald, 2002) – have been consistently confirmed, albeit in some cases with effect sizes smaller than in the original work (e.g., Alogna et al., 2014).

The Collaborative Replications and Education Project (CREP; Grahe et al., 2015; Wagge et al. in press) is a crowdsourced initiative to organize undergraduate experimental methods classes into research teams. Consider that in the United States alone, 70% of the more than 80,000 students who graduate each year with a bachelor's degree in psychology complete a class requiring conducting an empirical data collection (Hauhart & Grahe, 2012). Only one in ten of these class projects, often direct replications of classic and well-established findings such as the Stroop effect (Stroop, 1935), are ever presented at conferences or submitted to a journal (Perlman & McCann, 2005). The CREP is leveraging such student projects to replicate published findings whose robustness is less well established, such as the effects of color on attraction, disgust on moral judgment, and desire for social status on conservation behaviors. The focus is on simple studies within the technical abilities of students, the kind that would in at least some cases be delegated to undergraduate research assistants if conducted in a traditional laboratory context. In the collaborative replication and education model, the student truly becomes a junior scientist, with quality work aggregated with the results from other student projects and submitted for publication to peer-reviewed journals (Everett & Earp, 2015; Frank & Saxe, 2012).

One particularly promising model for facilitating crowdsourced research, whether to conduct replications, novel studies, or intervention contests, is the development of a standing, international network of psychological science laboratories that have committed to contributing to large-scale collaborations. The Psychological Science Accelerator (PSA) is a distributed laboratory network, currently numbering 346 laboratories in 53 countries, that aims to crowdsource every step of the research life-cycle, from idea generation and experimental design through to drafting and dissemination (psysciacc.org; Moshontz et al., 2018). Thus far the PSA has selected 5 studies that are at various stages of preparation and all have large numbers of labs committed to data collection, ranging from just over 30 to 160. The PSA has secured 1 in-principle acceptance for a study and begun data collection. Two studies are currently under

review as registered reports, and two more are in preparation. Finally, in a collaboration with the CREP project called the Accelerated CREP, students from PSA labs will conduct one CREP replication project per year. The Accelerated CREP aims to greatly reduce the amount of time typically required to complete a CREP replication.

Deciding what findings to pursue further

Generally supporting the idea that crowd inputs are useful in deciding what scientific findings are worth pursuing further, studies consistently show that the aggregated predictions of scientists accurately anticipate replication outcomes (realized effect sizes and significance levels). The first such demonstration was from Dreber et al. (2015), who carried out a prediction market allowing scientists to bet money on the results of the ongoing Reproducibility Project: Psychology. Collectively, participants in the prediction market accurately anticipated the project results, with aggregated bets closely tracking replication outcomes. Similar results were obtained for predicting replications of experimental results in economics, social science articles published in *Science* and *Nature,* and the Many Labs 2 initiative in social psychology (Camerer et al., 2016; 2018; Forsell et al., 2018).

DellaVigna & Pope (in press, 2018) examined whether a diverse crowds of individuals, from expert behavioral scientists to doctoral students to Mechanical Turk workers, could predict the results of experimental manipulations designed to improve task performance. Interventions such as different levels of piece rate pay, telling workers better performance would lead to a donation to charity, and encouraging social comparisons to other workers were used in the context of simple tasks (e.g., pressing the 'a' or 'b' on a keyboard, coding World-War II conscription cards). The forecasting results again revealed substantial accuracy, although crowds systematically overestimated the extent to which demographic characteristics such as gender, age, and education would moderate the effectiveness of the treatments. Remarkably, senior scientists were no more accurate than junior scientists and online workers at forecasting research outcomes. (See also Landy et al., 2018 and Eitan et al., 2018 for similar null and mixed effects of academic seniority in forecasting contexts).

Landy et al. (2018) provided a crowd of scientists with 64 sets of materials from unpublished experiments designed to test five distinct hypotheses related to moral judgments, negotiations, and implicit cognition. Forecasters were asked to predict the significance levels and effect sizes that would emerge when online participants were run in each study design. Aggregated estimates accurately anticipated not only the overall results, but also variability in results across different sets of study materials designed to test the same hypothesis. In other words, forecasters were able to predict, from the materials alone, how design choices would affect the degree of empirical support for a given hypothesis.

In the case of ongoing scientific debates, a tournament-based approach can be employed (Tetlock, Mellers, Rohrbaugh, & Chen, 2014). Scientists with a diverse range of opinions first make *a priori* predictions regarding the results of a high-powered empirical study relevant to the controversy. They are subsequently presented with the obtained evidence and provided the opportunity to either update their beliefs or counter-argue the results. Eitan et al. (2018) carried out a prediction survey to see if scientists could forecast the extent to which coded research

abstracts from a social psychology conference would exhibit political biases. Forecasters accurately predicted than conservatives would be the focus of explanation more than liberals, and explained in more negative terms than liberals, in the scientific abstracts. They also significantly overestimated the size of both these explanatory and evaluative differences, and updated their general beliefs about politics in science in light of the new empirical evidence.

That crowds both exhibit considerable accuracy at forecasting future findings and rationally update their beliefs bodes well for leveraging them to select what directions to head in next. For instance, scientific claims the crowd considers either highly unlikely (e.g., extrasensory perception) or clearly proven (e.g., anchoring bias) might be deprioritized in favor of findings about which controversy exists and predictions are mixed. Crowd surveys might also be used to identify which findings the scientific community regards as especially important if true, for instance due to their theoretical or social policy implications. These complementary criteria (likelihood of being true, and interest value if true) might be used in conjunction to allocate research resources for maximum impact and information gain.

As discussed in the main text, crowds can be mobilized to help identify the most robust research paradigms and then improve upon them. Lai and colleagues (2014; 2016) held intervention contests to identify the most effective strategies for reducing implicit preferences for Whites over Blacks. This approach allowed for direct quantitative comparison between interventions that would not have occurred if studies were conducted under a singular contribution model. Research teams submitted 17 interventions to the contest with substantial diversity of theoretical mechanisms including imagined positive contact, exposure to counter-stereotypical exemplars, evaluative conditioning, perspective-taking, and appeals to egalitarian values. Eight were effective in reducing implicit White preference immediately after the intervention, but none were effective a day or more after the intervention. Through systematic comparisons, the contest revealed what approaches were most effective at shifting implicit preferences, and showed that changing implicit cognitions is more difficult than previously understood. Teams were able to observe each others' approaches and results between rounds, which a number of them used to improve their own experimental intervention.

Supplement 3: Data quality and online studies

On Amazon's Mechanical Turk (MTurk), employers hire workers to complete simple tasks a computer cannot do effectively, such as transcribing text. A researcher can hire a small subset of the site's half a million workers to complete her research study, converting the platform into an expedient, low-cost source of data. MTurk samples are more representative of the general population than convenience samples of university students, scales exhibit similar reliabilities as when administered in the laboratory, and the magnitude of well-established experimental effects (e.g., base rate neglect; Tversky & Kahneman, 1981, 1983) is likewise comparable (Behrend, Sharek, Meade, & Wiebe, 2011; Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). Researchers can screen participants for clinical and cross-cultural comparison studies, and contact the same respondent repeatedly to collect longitudinal observations (Chandler & Shapiro, 2016; Paolacci et al., 2010). Similar online labor platforms to Mechanical Turk include clickworker.com, crowd-works.com, figure-eight.com, ttv.microworkers.com, and prolific.ac.

Although they have significant limitations, online platforms for crowdsourced labor have succeeding in reducing some research areas' over-reliance on university subject pools, providing access to more demographically diverse samples (Sears, 1986). One shortcoming of the MTurk workforce as a data source is that a subset of workers complete far more than their share of the posted online studies, and therefore may not represent naive participants for some widely studied effects (Chandler, Mueller, & Paolacci, 2014; Stewart et al., 2015). There is also a risk some participants will fake their geographic locations to participate in studies not open to them, and subsequently provide low quality data. Some measures to address data quality that researchers can consider include only recruiting workers with a 99% acceptance rate and more than 1000 hits approved, screening out duplicate GPS coordinates, and removing any participants who provide incoherent written statements or statements that are word-for-word identical to another participant.