ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2019, Vol. 14(5) 711–733
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1745691619850561
www.psychologicalscience.org/PPS
SAGE

Scientific Utopia III: Crowdsourcing Science

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

[1]Organizational Behaviour Area, INSEAD, Singapore; [2]Department of Psychology, University of Virginia; [3]Department of Psychology, Ashland University; [4]Center for Open Science, Charlottesville, Virginia; [5]Department of Psychology, University of Utah; [6]Department of Psychological and Brain Sciences, Washington University in St. Louis; [7]Center for the Study of Family Violence and Sexual Assault, Northern Illinois University; [8]University Libraries, University of Minnesota; and [9]Department of Business and Management, University of Sussex

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 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 and 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 can accelerate the progress of science and improve the quality of scientific research.

Keywords

crowdsourcing, collaboration, teams, methodology, metascience

There is no perfect study. Scientists, in their effort to understand nature, are constrained by limited time, resources, and expertise. This constraint 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 originators of new ideas and evidence through the 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

Corresponding Authors:

Eric Luis Uhlmann, INSEAD, Organisational Behaviour Area, 1 Ayer Rajah Ave., 138676 Singapore
E-mail: eric.luis.uhlmann@gmail.com

Brian A. Nosek, University of Virginia, Department of Psychology, Box 400400, Charlottesville, VA 22904-4400
E-mail: nosek@virginia.edu

contributions, enable big science, and increase transparency and reliability. The 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

Some academic research resembles a vertically integrated business. An individual or 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 coauthors on a manuscript reporting the research. The sought-after reward is acceptance and publication in the most widely read and prominent journal possible.

This model has several 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 collect only so much data, carry out only so many analyses, and consider only 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. Scholars are evaluated on the basis of the reception of their work in the idea marketplace. Outstanding ideas and evidence 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 the productive generation and culling of ideas by the independent actions of scholars with different priors, assumptions, expertise, and interests. Small teams often work together repeatedly and develop cospecializations that enable deep dives into a methodology or phenomenon. A community of scientists then shares its work, exchanges feedback, and serially builds on each other's findings.

At the same time, for some research questions and contributions, the standard model may limit progress. Individual researchers and small teams must consider certain trade-offs 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 trade-offs 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, 2019; Simons, 2014), obtaining several different perspectives on how to analyze the same data (Silberzahn et al., 2018), and using 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 the 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). Modern science is already stretching the standard model in more collaborative directions (see Supplement 1 in the Supplemental Material available online). 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

(Valderas et al., 2007; Wuchty, Jones, & Uzzi, 2007), and working in teams is associated with greater individual career success (Kniffin & Hanks, 2018). Team-authored articles are more cited than solo-authored articles, and this gap in scholarly impact has increased over time (Valderas et al., 2007; Wuchty et al., 2007).

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 several potential benefits for science, among which are 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 or small teams 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 personal 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, 2018; Clauset, Arbesman, & Larremore, 2015). Grant funding is overallocated to elite universities, and evidence suggests that returns on investment would be greater if the funds were distributed more evenly (Wahls, 2018). 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 rich in talent, 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 a research literature filled with inaccurately estimated effect sizes as a result of 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 under way in psychology and other scientific disciplines. Multilab 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, 2018; 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 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 research from small teams 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 implemented 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.

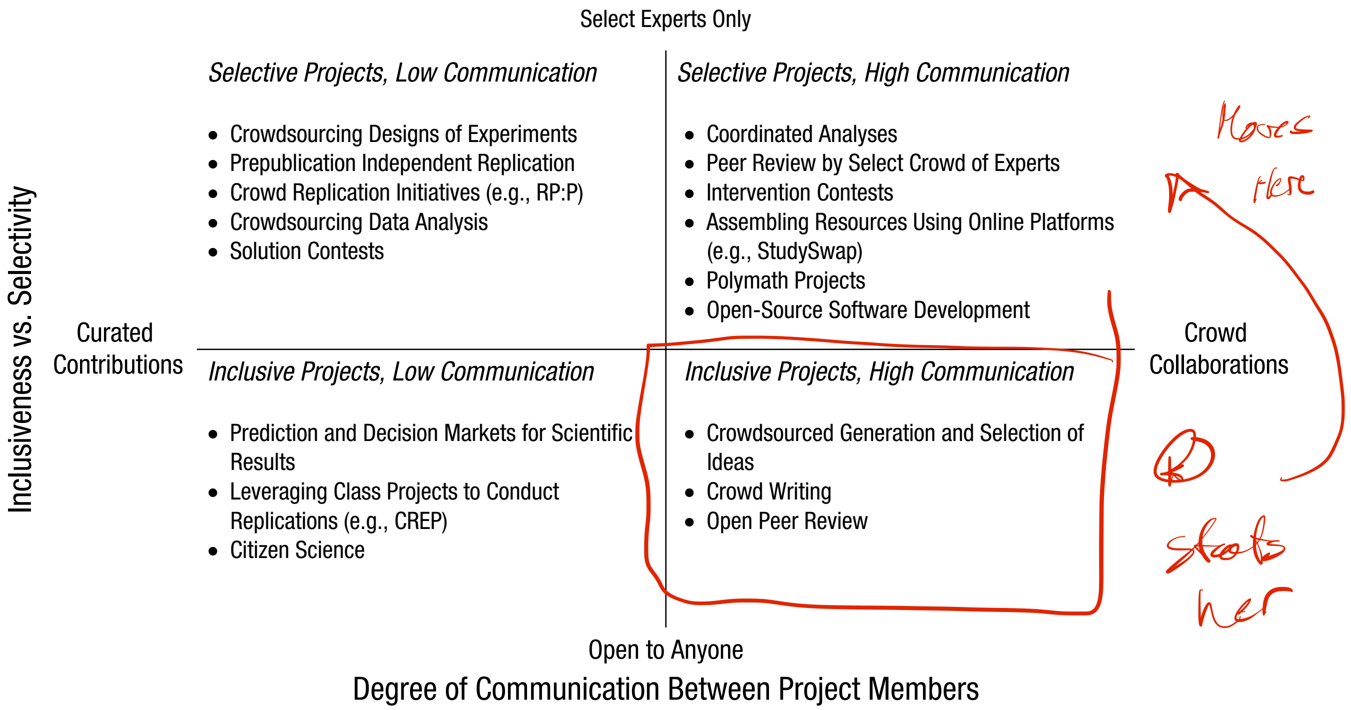
Forms of Scientific Crowdsourcing

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, Jeppesen, Lohse, & Panetta, 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. Crowd-science initiatives also 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 × 2 matrix. Examples of relevant crowdsourced projects are placed in this matrix as illustrations. These projects are described in greater detail in the next section and in Tables 1 and 2 (see also Supplements 1 and 2 in the Supplemental Material). 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 quadrant, and more inclusive forms of crowd writing (Christensen & van Bever, 2014) are in the bottom-right quadrant. Open peer review, in which anyone can publicly comment on a scientific manuscript or article, falls into the bottom-right quadrant, and crowd review by experts carefully chosen by a journal editor falls into the top-right quadrant. Traditional small-team research, with unrestricted communication and select membership, falls outside the extreme top-right corner of the matrix at the far end of both axes.

Multistage projects may operate in different locations in this space during the research life cycle. For example, to explore consensus building about disparate findings from the same data set, 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 e-mail 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 subteams 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 prone to error (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; Garcia-Berthou & Alcaraz, 2004; Salter et al., 2014; Westra et al., 2011), and the quality-quantity trade-off that can accompany scaling



Select Experts Only

Selective Projects, Low Communication

- Crowdsourcing Designs of Experiments
- Prepublication Independent Replication
- Crowd Replication Initiatives (e.g., RP:P)
- Crowdsourcing Data Analysis
- Solution Contests

Selective Projects, High Communication

- Coordinated Analyses
- Peer Review by Select Crowd of Experts
- Intervention Contests
- Assembling Resources Using Online Platforms (e.g., StudySwap)
- Polymath Projects
- Open-Source Software Development

Inclusiveness vs. Selectivity

Curated Contributions

Crowd Collaborations

Inclusive Projects, Low Communication

- Prediction and Decision Markets for Scientific Results
- Leveraging Class Projects to Conduct Replications (e.g., CREP)
- Citizen Science

Inclusive Projects, High Communication

- Crowdsourced Generation and Selection of Ideas
- Crowd Writing
- Open Peer Review

Open to Anyone

Degree of Communication Between Project Members

Fig. 1. Forms and examples of crowdsourcing. *Curated contributions* refers to projects in which project coordinators collect the individual work of a crowd of contributors whose communication with one another is limited to nonexistent. *Crowd collaborations* refers to projects in which a large group of contributors engage in regular communication regarding their shared work. CREP = Collaborative Replication and Education Project; RP:P = Reproducibility Project: Psychology.

up is potentially offset by the numerous eyes available to catch mistakes (e.g., Silberzahn et al., 2018). The available evidence suggests that 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 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

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 carries out statistical analyses to address the same research question
Replicating findings before publication	The same methodology is repeated in independent laboratories to confirm the finding before its publication
Writing research reports	A large group of contributors collectively writes a research article
Peer review	A large group of commentators writes public feedback on a scientific article
Replicating published findings	The same methods and materials from published articles 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

Source	Method	Key result(s)
	Ideation	
Sobel (2007)	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)	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)	Crowd of researchers asked to nominate hypotheses for testing with a complex data set	The crowd was able to generate interesting hypotheses for later testing
InnoCentive.com	Scientific problems are posted online, and prizes are offered for the best solution	30% of 166 scientific problems solved via crowd competitions for prizes
	Assembling resources	
Science Exchange	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
StudySwap	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
	Study design	
Landy et al. (2018)	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
	Data collection	
Olmstead (1834)	In 1833, 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, Barlow, and Gulick (2001)	Clickworkers website from the National Aeronautics and Space Administration asks volunteers to help classify images	Mapping of craters on Mars based on images from the Viking Orbiter
Church (2005)	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)	Online game *FoldIt* in which more than 50,000 players compete to fold proteins	The best human players outperform a computer in terms of determining protein structures
Price, Turner, Stencel, Kloppenborg, and Henden (2012)	Citizen sky project recruits amateur astronomers to 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)	Video game *EyeWire* in which players reconstruct part of an eye cell using three-dimensional images of microscopic bits of retinal tissue	Data from more than 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 International Consortium (2016)	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 markers

(continued)

Table 2. (Continued)

Source	Method	Key result(s)
Sørensen et al. (2016)	Video game *Quantum Moves* in which the player moves digital renditions of quantum atoms	The data produced by the more than 200,000 users has been leveraged to develop better quantum algorithms
Moshontz et al. (2018)	Psychological Science Accelerator (PSA), a network of more than 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 on the basis of valence and dominance
Zooniverse	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	Asks volunteers to help classify galaxies on the basis of images	Collection of more than 100 million classifications of galaxies based on shape, structure, and intensity; identifying supernovas and potential interactions between galaxies
Audubon Christmas Bird Count	Beginning with the Audubon Christmas Bird Count of 1900, amateur birdwatchers have been used to collect data on bird migrations	Large data set on bird migrations leveraged for scientific publications
	Data analysis	
Stolovitzky, Monroe, and Califano (2007)	In the 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 and then apply the analytic strategies to a hold-out data set 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 and Piccinin (2009)	Coordinated analysis: network of researchers use the same target constructs, model, and covariates on different longitudinal data sets 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)	42 analysts were asked to test hypotheses related to gender, status, and science using a complex data set 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)	The same data set 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
	Replicating findings before publication	
Schweinsberg et al. (2016)	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 preregistered replication criteria
	Writing research reports	
Christensen and van Bever (2014)	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

(continued)

Table 2. (Continued)

Source	Method	Key result(s)
	Peer review	
List (2017)	*Synlett* implemented a crowdsourced reviewing process to allow more than 100 referees to respond to articles 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
	Replicating published findings	
Steward, Popovich, Dietrich, and Kleitman (2012)	Initiative to replicate spinal-cord-injury research in independent laboratories	2 successful replications out of 12 targeted studies
Alogna et al. (2014)	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 article
Klein et al. (2014)	Many Labs 1: 36 laboratories attempted to replicate 13 psychology findings	10 of 13 findings replicated
Open Science Collaboration (2015)	Reproducibility project that attempted to replicate 97 original effects from top psychology journals in independent laboratories	36% of findings successfully replicated
Camerer et al. (2016)	Experimental Economics Replication Project: initiative to replicate prominent findings in experimental economics in independent laboratories	61% of findings successfully replicated
Ebersole et al. (2016)	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. (2018)	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 and Errington (2017)	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 article, 4 replicated some parts of the original article but not others, 2 were not interpretable, and 2 did not replicate the original findings
Camerer et al. (2018)	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)	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 nonreplicable effects
Cova et al. (2018)	Initiative to replicate prominent findings in experimental philosophy in independent laboratories	78% of findings successfully replicated
O'Donnell et al. (2018)	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. (2019)	Collaborative Replications and Education Project initiative to replicate social-psychology findings in student methods classes	Failure to replicate earlier findings that women are more attracted to men in photographs with red borders
	Deciding future directions	
Dreber et al. (2015)	Prediction market to see whether independent scientists could forecast the results of the Reproducibility Project: Psychology	Aggregated predictions accurately anticipated replication results

(continued)

Table 2. (Continued)

Source	Method	Key result(s)
Camerer et al. (2016)	Prediction market to see whether independent scientists could forecast replication results in experimental economics	Aggregated predictions accurately anticipated replication results
DellaVigna and Pope (2018a)	Prediction survey to see whether 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)	Prediction survey to see whether 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)	Prediction survey to see whether 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. (2018)	Prediction market to see whether independent scientists could forecast results replications of social-science articles in *Science* and *Nature*	Aggregated predictions accurately anticipated replication results
DellaVigna and Pope (2018b)	Prediction survey to see whether 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 the importance of demographic moderators; academic seniority did not moderate forecasting accuracy
Forsell et al. (2018)	Prediction market to see whether independent scientists could predict the results of the Many Labs 2 replication initiative	Aggregated predictions accurately anticipated replication results
Lai et al. (2014)	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.

successfully replicate a given behavioral effect (Bench, Rivera, Schlegel, Hicks, & Lench, 2017; see also Klein, Vianello, Hasselman, & Nosek, 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 Fig. 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 data from fMRI across laboratories; top-left quadrant of Fig. 1). Even with an open call, potential contributors may volunteer for projects in which 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 literature 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 Fig. 1) may help to avoid the potentially biasing effect of individuals' overcommitment to intellectual claims (Berman & Reich, 2010; Luborsky et al., 1999; Manzoli 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 from sensitivity to new evidence (see Fig. 4 in 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.

Allowing information exchange and creating interdependencies between project members also comes with potential important benefits. One of the hypothesized benefits of crowd collaboration is the ability of members of the community to learn from each other (Wenger, 1998). 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, multistage 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 article, obtaining reviewer feedback, and deciding next steps for the program of research. Table 2 and Supplement 2 in the Supplemental Material summarize some recent crowdsourced scientific initiatives, organized by the respective stages on which they focused their crowd efforts.

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 other's suggestions (open communication; the far-right regions of Fig. 1). This approach has been used to great effect in the Polymath projects, resulting in several important mathematical proofs (Ball, 2014; Polymath, 2012, 2014; Tao, Croot, & Helfgott, 2012). Like how they are used in product-design contests (Poetz & Schreier, 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 data sets that 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 serves 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 data sets and collaboration opportunities that would otherwise remain closed to most scientists.

Assembling resources

Genome-wide association studies 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, Brown, McCarthy, & Yang, 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 finds they have an excess of research resources (e.g., they compensate participants for a 30-min session for completing a 15-min 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 (https://www.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 top-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., 2018; 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 because of 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 (low communication combined with high expertise; top-left corner of Fig. 1). 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 have become widely used as a source of inexpensive research participants and coders (Stewart, Chandler, & Paolacci, 2017; see Supplement 3 in the Supplemental Material). 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 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 game *EyeWire*, in which players help reconstruct eye cells (Kim et al., 2014), and the protein-folding game *FoldIt* (Cooper et al., 2010). Note that 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 data set 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 articles on fMRI 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; Simmons et al., 2011; Simonsohn, Nelson, & Simmons, 2014). One way to maximize transparency is to turn the analysis of data over to a crowd of experts. The same data set is distributed to numerous scientists who are asked to test the same theoretical hypothesis, at first without knowing the specifications used by their colleagues (high expertise combined with low communication; top-left quadrant of Fig. 1). This offers an opportunity to assess how even seemingly minor differences in choices may affect research outcomes and reduces the pressure to observe any particular outcome—at least for the 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 data sets. Crowdsourced challenges have been used by researchers for benchmarking new computational methods, as with, for instance, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) Challenge focused on predicting the 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 (a top-left-quadrant approach) and then apply the analytic strategies to a hold-out data set 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 data sets (a *coordinated analysis*; Hofer & Piccinin, 2009). This approach was pioneered by the Integrative Analysis of Longitudinal Studies on Aging 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 data sets that include the same constructs (e.g., measures of cognitive function such as memory, reasoning, and fluency) and yet measure them in disparate ways in different populations (e.g., Sweden, Austria, the Netherlands, and the United Kingdom) far more systematically assesses the generalizability of the results than relying on a single data source. Because members of this network of experts communicate extensively to agree on their shared analytic approach and measures to use from each longitudinal data set, a coordinated analysis falls into the top-right quadrant of Figure 1.

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

Replicating findings before 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, 2018) addresses power and generalizability directly. Authors can specify a priori in which replication samples and laboratories they expect their findings to emerge; for example, they might select only topic experts as their replicators and thus moving up the vertical axis of Figure 1. This approach, which thus far returns a 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 toward 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).

4

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 article requires a few group leaders who complete a disproportionate amount of the work and organize and edit the written material of others (Kittur & Kraut, 2008; Kittur, Lee, & Kraut, 2009). Our personal experience with articles with many authors 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 nonexperts (e.g., elite Mechanical Turk workers) are surprisingly capable at drafting quality summaries of scientific findings for lay readers (Kittur, Smus, Khamkar, & Kraut, 2011). Such quality raw material could be carefully vetted and included in reviews of scientific research for practitioners and lay audiences. This suggests 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 article.

Peer review

In the current system of academic peer review, an unpublished manuscript is submitted to a journal and evaluated by the editor and usually two to five external referees, each of whom provides 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 ("Revolutionizing Peer Review?" 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, Mutz, & Daniel, 2010; Cicchetti, 1991; Marsh, Jayasinghe, & Bond, 2008) and that there is 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 article. 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, 1997).

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 data sets, code, and materials and reanalyses, replications, and critiques by other scientists. Peer review would be open, credited, and citable, and prominent articles that attract attention would be evaluated 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, the posting of data and code that others find useful, insightful feedback on others' work, and the curation of content related to specialized topic areas (e.g., replicability of the effects of mood on helping behaviors; LeBel et al., 2018). 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), top-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, whereas 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; Errington et al., 2014; McCarthy et al., 2018; O'Donnell et al., 2018). In these crowdsourced projects, up to 100 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., 2016, 2018; Ebersole et al., 2016; Klein et al., 2014, 2018; Open Science Collaboration, 2015).

A crowdsourced approach to replicability 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 be replicated tend to consistently fail to be replicated across cultures and demographic populations, which casts doubt on the argument that as-yet-unidentified moderators explain the disappointing results. The lack of 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., 2014, 2018) 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., 2013; Wagge et al., 2019) 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, an international network of more than 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 articles have been published since 1965 (Pan et al., 2016)—and 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), researchers must determine how best to distribute limited replication resources. Viable options include focusing on highly cited articles, 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 on the basis of 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—because in many cases, the aggregation cancels out individual errors (Galton, 1907; Mellers et al., 2014; Surowiecki, 2005). Likewise, the averaged independent predictions of scientists regarding research outcomes—based solely on examinations of short summaries of the findings, research abstracts, or study materials—are remarkably well aligned with realized significance levels and effect sizes (Camerer et al., 2016; DellaVigna & Pope, 2018a, 2018b; 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, 2018a, 2018b; 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 considered either clearly spurious or well-established by a heterogeneous crowd of scientists 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-on question of broad interest is present. Large-scale efforts seem less appropriate for findings the community considers highly unlikely to be true (e.g., extrasensory 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 (Lai et al., 2014, 2016) held a series of crowdsourced contests to identify the best interventions for reducing implicit racial biases. Beginning in the top-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 top right of the 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 an intrinsic interest in contributing to the accumulation of knowledge and are not solely driven by prestige. 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 et al., 2018; Clauset et al., 2015; Merton, 1968; Petersen et al., 2011; Wahls, 2018). The resulting hierarchical and network-based arrangements interfere with inclusivity for researchers who have much to offer but come from disadvantaged backgrounds and/or lack 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 the total return on investment in terms of scientific knowledge (Wahls, 2018). The 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), preregistering studies and analysis plans (Nosek, Ebersole, DeHaven, & Mellor, 2018; Nosek & Lakens, 2014; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012), conducting replications, helping to 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, not only by providing additional opportunities for those with useful skills and knowledge to contribute (Wahls, 2018) but also by directly incentivizing emerging best practices. To avoid the 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).

Author contribution statements

Although some especially elaborate crowd projects involve specialized subteams 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. Articles with many authors that report large-scale projects require reforms in how intellectual credit is allocated. Input can be documented through careful and detailed author contribution statements, which academic journals increasingly require. A good starting point for the crafting of clear contribution statements is the CRediT taxonomy (Brand, Allen, Altman, Hlava, & Scott, 2015), in which contributions throughout the full research life cycle are represented in categories such as conceptualization, data curation, writing, and visualization. Providing information about which coauthors 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 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 articles 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; http://osf.io/) 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 one's curriculum vitae (CV) 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 on the basis of merit.

Integrating crowd science into pedagogy

Another way to encourage crowd science is to build such initiatives into activities that scientists in many fields already do routinely, such as collecting data in methods classes for student projects and analyzing complex data sets as part of graduate education (Everett & Earp, 2015; Frank & Saxe, 2012; Grahe et al., 2012; Mavor et al., 2016). The CREP (Grahe et al., 2013; Wagge et al., 2019) and Pipeline Projects (Schweinsberg et al., 2016; Schweinsberg, Viganolla, et al., 2018) offer opportunities to leverage such activities for articles with many authors that report crowdsourced replications. In these cases, for both students and course instructors, being the middle author on a report of an interesting initiative is better than no author credit at all. Crowdsourcing avoids letting the students' hard work collecting data go to waste through repeating established paradigms (e.g., the Stroop effect) in unpublishable class projects the results of which are low in information gain. As a further incentive, the second Pipeline Project 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 to 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 prepublication 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 metascientific work that relies on a crowd approach, a trend that promises to encourage future crowdsourced projects. A more general shift in emphasis toward 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. Online platforms such as the Harvard Dataverse and Open Science Framework are available to host data, research and teaching materials, and preregistrations and document workflows. Journal mechanisms such as Registered Reports that review methodology and accept articles in principle before 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 articles. Recently introduced tools such as StudySwap and standing laboratory networks such as the Psychological Science Accelerator likewise hold promise to change the landscape of everyday science.

These approaches to encourage large-scale collaboration are important complements to reforms in how small-team science is conducted and funded. Larger samples (Stanley et al., 2018), disclosure rules (Simmons et al., 2011), preregistration (Nosek et al., 2018; Wagenmakers et al., 2012), and Registered Report 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 versus context sensitivity of particularly important findings. At the same time, crowdsourced metascientific investigations can help to assess the effectiveness of new practices intended to improve science but that may also have unwanted side effects. For instance, preregistration might reduce false-positive results but could also negatively affect the rate of novel discoveries by dampening creativity (Brainerd & Reyna, 2018). A crowdsourced project in progress (Ebersole et al., 2018) will randomly assign researchers to preregister their analyses of a complex data set to empirically assess the costs and benefits of this proposed reform. Finally, the encouragement of large-scale collaborations to help democratize participation in research is a complement to supporting research at teaching institutions through grants, 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 used 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 or of the vertically integrated model. Rather, it seeks to complement this standard approach to provide more options for accelerating scientific discovery.

Action Editor

Timothy McNamara served as action editor for this article.

Author Contributions

The study was outlined and the literature review conducted through a crowdsourced process to which all authors contributed. The third through ninth authors are listed alphabetically in the byline. E. L. Uhlmann, C. R. Ebersole, C. R. Chartier, T. M. Errington, C. K. Lai, R. J. McCarthy, A. Riegelman, R. Silberzahn, and B. A. Nosek drafted the body of the manuscript. C. R. Ebersole created the figure and tables. All of the authors provided critical edits and revisions and approved the final manuscript for publication.

ORCID iD

Timothy M. Errington https://orcid.org/0000-0002-4959-5143

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by the Center for Open Science, John Templeton Foundation, Templeton World Charity Foundation, Templeton Religion Trust, Arnold Ventures, James McDonnell Foundation, and a research and development grant from INSEAD.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/1745691619850561

References

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.

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

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

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

Baribault, B., Donkin, C., Little, D. R., Trueblood, J., Oravecz, Z., van Ravenzwaaij, D., . . . Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences, USA*, *15*, 2607–2612. doi:10.1073/pnas.1708285114.

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*, 449–466.

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

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*, 278–295.

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. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences, USA*, *15*, 4887–4890. doi:10.1073/pnas.1719557115.

Börner, K., Contractor, N., Falk-Krzesinski, H. J., Fiore, S. M., Hall, K. L., Keyton, J., . . . Uzzi, B. (2010). A multi-level systems perspective for the science of team science. *Science Translational Medicine*, *2*(49), Article 49cm24. doi:10.1126/scitranslmed.3001399

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), Article e14331. doi:10.1371/journal.pone.0014331

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*, 151–155.

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*, 267–280. doi:10.1287/mnsc.2014.1909

Buttliere, B. T. (2014). Using science and psychology to improve the dissemination and evaluation of scientific work. *Frontiers in Computational Neuroscience, 8*, Article 82. doi: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., . . . Wu, H. (2018). Evaluating replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, *2*, 637–644.

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

Carp, J. (2012b). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*, 289–300. doi:10.1016/j.neuroimage.2012.07.004

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

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

Chargaff, E. (1978). *Heraclitean fire: Sketches from a life before nature*. New York, NY: Rockefeller University 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*, Article 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 & Brain Sciences*, *14*, 119–135.

Clauset, A., Arbesman, S., & Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science Advances*, *1*(1), Article e1400005. doi:10.1126/sciadv.1400005

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

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

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., . . . Zhou, X. (2018). Estimating the reproducibility of experimental philosophy. *Review of Philosophy and Psychology*. Advance online publication. doi:10.1007/s13164-018-0400-9

Davis, R., Espinosa, J., Glass, C., Green, M. R., Massague, J., Pan, D., & Dang, C. V. (2018). *Reproducibility project: Cancer biology*. Retrieved from https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology

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

DellaVigna, S., & Pope, D. G. (2018b). What motivates effort? Evidence and expert forecasts. *The Review of Economic Studies*, *85*, 1029–1069.

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

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

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . 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., et al. (2018). *Experimentally examining the consequences of preregistered analyses*. Manuscript in preparation.

Eisenman, I., Meier, W. N., & Norris, J. R. (2014). A spurious jump in the satellite record: Has Antarctic sea ice expansion been overestimated? *Cryosphere*, *8*, 1289–1296. doi:10.5194/tc-8-1289-2014

Eitan, O., Viganola, D., Inbar, Y., Dreber, A., Johanneson, M., Pfeiffer, T., . . . 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.

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*, Article 1152. doi:10.3389/fpsyg.2015.01152

Fanelli, D. (2010). "Positive" results increase down the Hierarchy of the Sciences. *PLOS ONE*, *5*(4), Article e10068. doi:10.1371/journal.pone.0010068

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

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., . . . Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*. Advance online publication. doi:10.1016/j.joep.2018.10.009

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

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*, Article 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*, 900–901.

Grahe, J. E., Brandt, M. J., IJzerman, H., Cohoon, J., Peng, C., Detweiler-Bedell, B., . . . Weisberg, Y. (2013). *Collaborative Replications and Education Project (CREP)*. Retrieved from the Open Science Framework website: https://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*, 605–607.

Greenland, P., & Fontanarosa, P. B. (2012). Ending honorary authorship. *Science*, *337*, 1019. doi:10.1126/science.1224988

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

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

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

Hirsch, J. E. (2007). Does the *h* index have predictive power? *Proceedings of the National Academy of Sciences, USA*, *104*, 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*, 150–164. doi: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*, 613–624.

Horton, J. (2010). Online labor markets. In A. Saberi (Ed.), *Workshop on internet and network economics* (pp. 515–522). Basel, Switzerland: Springer.

Howe, J. (2006, June 1). The rise of crowdsourcing. *Wired Magazine*. Retrieved from http://www.wired.com/wired/archive/14.06/crowds.html

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

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), Article e124. Retrieved from http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0020124

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

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

Jia, M., Ding, I., Falcão, H., Schweinsberg, M., Chen, Y., Pfeiffer, T., . . . Uhlmann, E. L. (2018). *The crowdsourced generation, evaluation, and testing of research hypotheses*. Manuscript in preparation.

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, March). *Can distributed volunteers accomplish massive data analysis tasks?* Poster presented at the Proceedings of the 32nd Annual Lunar and Planetary Science Conference, Houston, TX.

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

Kim, J. S., Greene, M. J., Zlateski, A., Lee, K., Richardson, M., Turaga, S. C., . . . Seung, H. S. (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 the 2008 ACM Conference on Computer Supported Cooperative Work* (pp. 37–46). New York, NY: Association for Computing Machinery.

Kittur, A., Lee, B., & Kraut, R. E. (2009). Coordination in collective intelligence: The role of team structure and task interdependence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1495–1504). New York, NY: Association for Computing Machinery.

Kittur, A., Smus, B., Khamkar, S., & Kraut, R. E. (2011). CrowdForge: Crowdsourcing complex work. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (pp. 43–52). New York, NY: Association for Computing Machinery.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*, 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*, *1*, 443–490.

Kniffin, K. M., & Hanks, A. S. (2018). The trade-offs of teamwork among STEM doctoral graduates. *American Psychologist*, *73*, 420–432. doi: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*, 551–560. doi:10.1002/fee.1436

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* (Working Paper 07-050). Retrieved from the Harvard Business School website: https://www.hbs.edu/faculty/Publication%20Files/07-050.pdf

Landry, J. (2000, September/October). Profiting from open source. *Harvard Business Review*. Retrieved from https://hbr.org/2000/09/profiting-from-open-source

Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Ebersole, C. R., . . . Uhlmann, E. L. (2018). *Crowdsourcing hypothesis tests*. Manuscript submitted for publication.

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, NY: Psychology Press.

LeBel, E. P., McCarthy, R., Earp, B., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*, 389–402.

Lindwall, M., Cimino, C. R., Gibbons, L. E., Mitchell, M., Benitez, A., Brown, C. L., . . . 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*, *2012*, Article 49359812. doi:10.1155/2012/493598

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

Littmann, M., & Suomela, T. (2014). Crowdsourcing, the great meteor storm of 1833, and the founding of meteor science. *Endeavour*, *38*, 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, USA*, *108*, 9020–9025.

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

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*, 276–299.

Manzoli, L., Flacco, M. E., D'Addario, M., Capasso, L., DeVito, C., Marzuillo, C., . . . Ioannidis, J. P. (2014). Non-publication and delayed publication of randomized trials on vaccines: Survey. *British Medical Journal*, *348*, Article g3058. doi:10.1136/bmj.g3058

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., . . . Fraser, J. S. (2016). Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *eLife*, *5*, Article e15802. doi:10.7554/eLife.15802

McCarthy, R. J., Skowronski, J. J., Verschuere, B., Meijer, E. H., Jim, A., Hoogesteyn, K., . . . Yildiz, E. (2018). Registered replication report on Srull & Wyer (1979). *Advances in Methods and Practices in Psychological Science*, *1*, 321–336.

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*, 56–63.

MetaSUB International Consortium. (2016). The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. *Microbiome*, *4*, Article 24. doi:10.1186/s40168-016-0168-z

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

Monin, B., & Oppenheimer, D. M. (2014). The limits of direct replications and the virtues of stimulus sampling. *Social Psychology*, *45*, 299–300.

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

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., . . . Chartier, C. R. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*, 501–515.

Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2019). Replication studies in economics—How many and which papers are chosen for replication, and why? *Research Policy*, *48*, 62–83.

Muffatto, M. (2006). *Open source: A multidisciplinary approach*. London, England: 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.

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, July 14). *Are reproducibility and open science starting to matter in tenure and promotion review?* Retrieved from the Center for Open Science website: 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., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425. doi:10.1126/science.aab2374

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

Nosek, B. A., Ebersole, C. R., DeHaven, A., & Mellor, D. M. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA*, *115*, 2600–2606. doi:10.1073/pnas.1708274114

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *eLife*, *6*, Article e23383. doi: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., 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., Nelson, L. D., Ackermann, E., Aczel, B., Akhtar, A., Aldrovandi, S., . . . Zrubka, M. (2018). Registered replication report: Dijksterhuis & van Knippenberg (1998). *Perspectives on Psychological Science*, *13*, 268–294. doi:10.1177/1745691618755704

Olmsted, D. (1834). Observations on the meteors of November 13th, 1833. *American Journal of Science and Arts*, *26*, 354–411.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. 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*, 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.

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, USA*, *108*, 18–23.

Piccinin, A. M., Muniz, G., Clouston, S. A., Reynolds, C. A., Thorvaldsson, V., Deary, 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*, 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*, 245–256.

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

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

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

Saez-Rodriguez, J., Costello, J. C., Friend, S. H., Kellen, M. R., Mangravite, L., Meyer, P., . . . 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*, 652–660.

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

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., . . . Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*, Article 87. doi:10.1186/s12915-014-0087-z

Schönbrodt, F. (2018, June 25). Hiring policy at the LMU Psychology Department: Better have some open science track record. *Nicebread*. Retrieved from http://www.nicebread.de/open-science-hiring-policy-lmu

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

Schweinsberg, M., Feldman, M., Staub, N., Prasad, V., Ravid, A., van den Akker, O., . . . Uhlmann, E. (2018). *Crowdsourcing data analysis: Gender, status, and science*. Manuscript in preparation.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., . . . 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., Viganola, D., Prasad, V., Dreber, A., Johannesson, M., Pfeiffer, T., . . . Uhlmann, E. L. (2018). *The pipeline project 2: Opening pre-publication independent replication to the world*. Manuscript in preparation. Retrieved from https://osf.io/skq2b/

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.

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., . . . Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*, 337–356. doi:10.1177/2515245917747646

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.

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

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 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*. London, England: 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*, 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. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*, 1325–1346.

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

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

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

Surowiecki, J. (2005). *The wisdom of crowds*. New York, NY: Anchor Books.

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

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., Jordan, J., Kennedy, D. M., Qureshi, I., Sommer, S. A., . . . Uhlmann, E. L. (2016). Data from a pre-publication independent replication initiative examining ten moral judgment effects. *Scientific Data*, *3*, Article 160082. doi:10.1038/sdata.2016.82

Tierney, W., Schweinsberg, M., & Uhlmann, E. L. (2018). Making prepublication independent replication mainstream [Commentary]. *Behavioral & Brain Sciences*, *41*, Article e153. doi:10.1017/S0140525X18000894

Valderas, J. M., Buckley, R., Wray, K. B., Wuchty, S., Jones, B. F., & Uzzi, B. (2007). Why do team authored papers get cited more? *Science*, *317*, 1496–1498.

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

Wagenmakers, E.–J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 627–633.

Wagge, J. R., Baciu, C., Banas, K., Nadler, J. T., Schwarz, S., Weisberg, Y., . . . Grahe, J. (2019). A demonstration of the Collaborative Replication and Education Project: Replication attempts of the red-romance effect. *Collabra*, *5*(1), Article 5. doi:10.1525/collabra.177

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

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, England: Cambridge University Press.

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

Westra, H.-J., Jansen, R. C., Fehrmann, R. S. N., te Meerman, G. J., van Heel, D., Wijmenga, C., & Franke, L. (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.

Wolfman-Arent, A. (2014, June 5). Frustrated scholar creates new way to fund and publish academic work. *Chronicle of Higher Education*. Retrieved from https://www.chronicle.com/blogs/wiredcampus/frustrated-scholar-creates-new-route-for-funding-and-publishing-academic-work/53073

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