Supplement 1: The growth of crowdsourcing

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

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

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

Supplement 2: More detailed descriptions of specific crowdsourced projects

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

Ideation

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

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

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

Assembling resources

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

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

Study design

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

Data collection

Opening participation in projects to the public via citizen science initiatives has had the most impact in biology, astronomy, ecology, and conservation, but is spreading to other fields. Amateur astronomers help professionals gather observations of the planets, moons, meteors, comets, stars, and galaxies (Price, Turner, Stencel, Kloppenborg, & Henden, 2012), and members of the general public aid in classifying images in huge research databases (e.g., NASA's Clickworkers and Galaxy Zoo; Hand, 2010; Kanefsky, Barlow, & Gulick, 2001). Over a quarter million amateur bird watchers and butterfly watchers are relied on to document animal populations and migrations (ebird.org; Cavalier & Kennedy, 2016; Devictor, Whittaker, & Beltrame, 2010), a mobile app is used by thousands of boat-goers to track water debris (Cressey, 2016), commuters are enlisted to obtain samples from surfaces in subways and other public areas in order to map a city's microbiome (Afshinnekoo et al., 2015; The MetaSUB Consortium, 2015), and armies of volunteers collect rain samples to facilitate research on pollution (Haklay, 2015; Kerson, 1989).

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

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

Data analysis

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

In a second crowdsourcing data analysis initiative, 42 analysts were asked to test hypotheses related to gender, status, and science using a complex dataset on academic debates (Schweinsberg, Feldman et al., 2018). The first hypothesis posited that female scientists participate more in intellectual conversations with a greater number of women, and the second that higher status academics are more verbose than are lower status academics. Each researcher decided not only her or his preferred statistical approach, but also how to operationalise key variables. For example, volubility could be operationalised as number of words spoken or number of times speaking; status could be measured using citation counts, job rank, university rank, or some combination. Under these conditions, which arguably more closely mimic those of the typical research project, effect size estimates proved radically dispersed, with different analysts in some cases reporting significant effects in opposite directions for the same hypothesis tested with the same data. This raises the unsettling possibility that even in the absence of perverse incentives and directional motives, analytical choices may have as great an effect on research conclusions as whether the hypothesis is true. Only a crowdsourced approach can make transparent the full extent to which research conclusions are contingent on the subjective decisions made by different analysts.

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

Replicating findings prior to publication

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

Writing research reports

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

Peer review

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

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

Replicating published findings

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

These efforts have generally yielded disappointing results. In the Reproducibility Project: Psychology, 35 (36%) of the original 97 effects from top psychology journals produced significant effects (p <.05) in the expected direction in the more highly powered replications. Although original and replication effect sizes were significantly correlated, replication effect sizes were also systematically lower than in the original papers. Earlier efforts by pharmaceutical companies to replicate a total of 120 landmark biomedical studies (53 by Amgen and 67 by Bayer) obtained reproducibility rates of 11-25% (Begley & Ellis, 2012; Prinz, Schlange & Asadullah, 2011). In the ongoing Reproducibility Project: Cancer Biology, 12 replications have been published to date, with editors at the publishing journal *eLife* determining that 4 replicated important parts of the original paper, 4 replicated some parts of the original paper but not others, 2 were not interpretable, while 2 did not replicate the original findings (Davis et al., 2018; cf. Wen et al., 2018). An effort among academics to replicate spinal cord injury research obtained six null results, 3 mixed results, an inconclusive outcome, and two successful outcomes out of 12 studies (Steward, Popovich, Dietrich, & Kleitman, 2012). Although direct comparisons cannot be made with any confidence due to differences in sampling and methodology, other replication initiatives obtained reproducibility rates of 61% in experimental economics (Camerer et al., 2016), and 78% in experimental philosophy (Cova et al., in press).

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

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

One particularly promising model for facilitating crowdsourced research, whether to conduct replications, novel studies, or intervention contests, is the development of a standing, international network of psychological science laboratories that have committed to contributing to large-scale collaborations. The Psychological Science Accelerator (PSA) is a distributed laboratory network, currently numbering 346 laboratories in 53 countries, that aims to crowdsource every step of the research life-cycle, from idea generation and experimental design through to drafting and dissemination (psysciacc.org; Moshontz et al., 2018). Thus far the PSA has selected 5 studies that are at various stages of preparation and all have large numbers of labs committed to data collection, ranging from just over 30 to 160. The PSA has secured 1 in-principle acceptance for a study and begun data collection. Two studies are currently under review as registered reports, and two more are in preparation. Finally, in a collaboration with the CREP project called the Accelerated CREP, students from PSA labs will conduct one CREP replication project per year. The Accelerated CREP aims to greatly reduce the amount of time typically required to complete a CREP replication.

Deciding what findings to pursue further

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

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

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

In the case of ongoing scientific debates, a tournament-based approach can be employed (Tetlock, Mellers, Rohrbaugh, & Chen, 2014). Scientists with a diverse range of opinions first make *a priori* predictions regarding the results of a high-powered empirical study relevant to the controversy. They are subsequently presented with the obtained evidence and provided the opportunity to either update their beliefs or counter-argue the results. Eitan et al. (2018) carried out a prediction survey to see if scientists could forecast the extent to which coded research abstracts from a social psychology conference would exhibit political biases. Forecasters accurately predicted than conservatives would be the focus of explanation more than liberals, and explained in more negative terms than liberals, in the scientific abstracts. They also significantly overestimated the size of both these explanatory and evaluative differences, and updated their general beliefs about politics in science in light of the new empirical evidence.

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

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

Supplement 3: Data quality and online studies

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

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