Behav Res
DOI 10.3758/s13428-013-0365-7

Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers

Jesse Chandler · Pam Mueller · Gabriele Paolacci

© Psychonomic Society, Inc. 2013

Abstract Crowdsourcing services—particularly Amazon Mechanical Turk—have made it easy for behavioral scientists to recruit research participants. However, researchers have overlooked crucial differences between crowdsourcing and traditional recruitment methods that provide unique opportunities and challenges. We show that crowdsourced workers are likely to participate across multiple related experiments and that researchers are overzealous in the exclusion of research participants. We describe how both of these problems can be avoided using advanced interface features that also allow prescreening and longitudinal data collection. Using these techniques can minimize the effects of previously ignored drawbacks and expand the scope of crowdsourcing as a tool for psychological research.

Keywords Crowdsourcing · Internet research · Data quality · Longitudinal research · Mechanical Turk · MTurk

Crowdsourcing is an increasingly popular method of allocating and managing labor. Just as businesses have used the Web to outsource labor, a number of Web sites have been developed to aid specific academic projects (Gaggioli & Riva, 2008). For example, reCaptcha verifies that Web site users are human by asking them to transcribe distorted images of words, while simultaneously digitizing illegible portions of books (von Ahn, Maurer, McMillen, Abraham, & Blum, 2008), Galaxy Zoo (www.galaxyzoo.org) solicits "citizen-scientists" to view and classify astronomical images (Lintott et al., 2008), and Foldit (www.fold.it) harnesses the power of human pattern recognition to predict the shape that proteins will form (Cooper et al., 2010). For businesses with smaller or shorter-term projects, a number of crowdsourcing marketplaces (e.g., Innocentive, oDesk, CloudCrowd, Amazon Mechanical Turk) offer various ways to access large pools of workers to complete more modest tasks. These Web sites are primarily used by companies seeking to outsource business tasks, but social scientists have increasingly turned to them as a viable alternative to traditional participant pools (cf Chandler, Paolacci & Mueller, 2013).

Most crowdsourcing markets are tailored to large or highly specialized tasks, rendering them unsuitable for the relatively modest requirements of social scientists. However, Amazon Mechanical Turk ("MTurk") specializes in recruiting workers to complete tasks that are small, fast, and often repetitive. On MTurk, participants ("workers") browse batches of human intelligence tasks ("HITs") by title, keyword, reward amount, availability, and so forth and complete HITs of interest (for an overview, see Mason & Suri, 2012). They are paid by "requesters" upon successful completion of the accepted tasks at a piecework rate. Requesters can discretionarily reject submissions or assign bonuses to workers, ensuring that work is of relatively high quality. This format lends itself well to the kinds of tasks often required of social science research participants. Additionally, MTurk's large user base and preexisting payment technology make it a convenient means of data collection. Consequently, MTurk has rapidly become a popular tool in the social sciences, particularly in social psychology, linguistics, and decision science.

Paralleling initial research on participants recruited from the Internet (e.g., Gosling, Vazire, Srivastava, & John, 2004; Krantz & Dalal, 2000; Reips, 2000), initial cross-sample

Electronic supplementary material The online version of this article (doi:10.3758/s13428-013-0365-7) contains supplementary material, which is available to authorized users.

J. Chandler (✉)
Woodrow Wilson School of Public Affairs, Princeton University,
Princeton, NJ, USA
e-mail: jjchandl@umich.edu

P. Mueller
Department of Psychology, Princeton University,
Princeton, NJ, USA

G. Paolacci
Rotterdam School of Management, Erasmus University,
Rotterdam, Netherlands

Published online: 09 July 2013

investigations of crowdsourcing (all investigating MTurk in particular) have demonstrated that online populations are more diverse and produce data of equal or better quality than do more traditional participant pools in a variety of domains, including social psychology (Behrend, Sharek, Meade, & Wiebe, 2011; Berinsky, Huber, & Lenz, 2012; Summerville & Chartier, 2012), cognitive psychology (Goodman, Cryder, & Cheema, 2012; Paolacci, Chandler, & Ipeirotis, 2010; Sprouse, 2011), personality psychology (Buhrmester, Kwang, & Gosling, 2011), and clinical psychology (Shapiro, Chandler, & Mueller, 2013).

While early research has been encouraging, it has tended to treat MTurk as "only" a cheaper, faster, and more diverse participant pool that allows access to either a more representative population (e.g., Berinsky et al., 2012) or hard-to-find subpopulations (Shapiro et al., 2013; for a notable exception, see Suri & Watts, 2011). While this was a useful starting point for assessing the comparability of MTurk with the typical university participant pool, some crucial differences between the two have been overlooked. First, some MTurk workers may remain members of crowdsourcing Web sites for far longer than undergraduates remain members of traditional participant pools. Also, unlike nearly any university participant pool, they are not typically restricted in the type of studies in which they can participate either by the administrator or by most individual researchers. These features increase the likelihood that they will complete many similar studies. This concern is compounded by the ease of data collection on MTurk, which makes it possible to quickly run many iterations of an experimental paradigm with minor modifications, expanding the opportunities for research participants to repeat the same or similar studies.

Second, although the MTurk interface does not provide networking capabilities to workers, worker discussion boards have been developed that facilitate worker interaction (e.g., mturkforum.com, turkernation.com), and MTurk subcommunities that share information can be found in unrelated online networks (e.g., Reddit, Facebook). Moreover, plug-ins have been created that allow workers to complete the tasks of favored requesters (e.g., turkopticon.differenceengines.com, turkalert.com). As a result, some workers may know more about the HITs available to them—and about the requesters who posted them—than is commonly assumed. This can increase the representation of these workers in a sample and, perhaps, even lead workers to have foreknowledge of the experiment. While participant nonnaïveté can also be an issue in traditional participant pools (e.g., psychology students; Edlund, Sagarin, Skowronski, Johnson, & Kutter, 2009), MTurk workers might share information in a more systematic, permanent, and searchable manner, with more dramatic consequences for data validity.

We are finally concerned with how the practical advantages of MTurk and the presumptions about the workers that populate it might affect researchers' choices in terms of data collection and analysis. While the finite size of traditional participant pools can be all too obvious to researchers, MTurk can appear to provide an inexhaustible supply of labor, making the need to carefully consider the population of interest prior to data collection seem less pressing. Furthermore, the low cost of discarding responses may lead to data-cleaning practices that are more zealous than those used for other samples, especially when faced with an MTurk study that almost "worked." There are doubtlessly workers who are unmotivated and inattentive and, thus, provide "poor quality" data, but this is also true of participants in traditional participant pools (perhaps to a greater degree; e.g., Paolacci et al., 2010) and of people in real-life interactions (Fiske & Taylor, 1984). However, beliefs about MTurk workers ("Who would do studies for such low pay anyway?") may make it easier to exclude data that do not agree with the researcher's hypothesis.

Both nonnaïveté and taking excessive liberties in dealing with the data undermine key assumptions of experimental research methods—specifically, that observations are randomly assigned and independent from each other. Fortunately, MTurk has a Qualification system that can function much like the prescreening systems implemented in large psychology participant pools. This system, which seems largely ignored, enables researchers to selectively allow or deny workers access to research studies, offering the potential to conduct designs that are both more sophisticated and more methodologically rigorous than those typically conducted using online convenience samples. We hope that this article will contribute to moving MTurk experimentation beyond merely exploiting speed and convenience and toward a more careful consideration of both its opportunities and its threats to validity.

This article is divided into several sections that address these issues. First, we examine the prevalence of nonnaïve workers through a secondary analysis of worker behavior and a survey of workers. We demonstrate that although cross talk is a minimal concern (in that workers are more interested in instrumental features such as payment and length of a HIT, rather than its content), duplicate workers are more common than researchers may assume. Repeated participation by workers is associated with increased self-reports of exposure to common experimental paradigms, as well as changes in responses to a paradigm likely to be vulnerable to practice effects (cf. Basso, Bornstein, & Lang, 1999). Second, we illustrate the problem of arbitrary data-cleaning strategies through an examination of previously published studies and show that researchers may be too eager to exclude workers post hoc, a tendency that can be remedied by identifying exclusion criteria a priori and allowing only workers who meet researcher-specified exclusion criteria to participate. Third, we describe how to use the Qualification system in MTurk, first for its intended use of prescreening workers, and then as a tool to manage the inclusion and exclusion of workers over a program of research, allowing solutions to the methodological issues we identify.

Workers may be less naïve than researchers assume

Researchers often presume that participants encounter experimental materials for the first time as they complete them and that they have not been exposed to other related manipulations of experimental stimuli or, worse, debriefings including information about the purpose of specific measures. Since MTurk is a large and relatively new population, this is assumed to be especially true (e.g., Chilton, Horton, Miller, & Azenkot, 2009). However, workers may have previous exposure to an experimenter's research materials, as a result of either completing them earlier or hearing about them from other workers.

Are different HITs completed by different workers?

By default, workers (as identified through their "WorkerID") are prevented from completing the same HIT twice. Although workers may be able to have more than one concurrent MTurk account and, thus, more than one WorkerID, this is uncommon. Amazon actively works to identify and eliminate duplicate accounts. More important, requesters often restrict lucrative HITs to workers who have completed a large volume of high-quality work in the past, making it less likely that workers are willing to undertake the substantial investment in creating a second profile. Examinations of worker IP addresses typically reveal a small minority of workers (around 2.5 %; Berinsky et al., 2012) who submit HITs from the same IP address, which may often result from workers being separate members of a single household. A secondary analysis of a recent study that tracked demographic responses and IP addresses across time points (from Shapiro et al., 2013) similarly found that 2.8 % of respondents ($N = 14$) shared an IP address with at least one other worker. However, eight of these workers reported demographic characteristics that were consistent with being distinct individuals in a single household: Sexual orientation matched partner sex, and demographic characteristics remained consistent across different HITs 1 week apart. The remaining six observations may have been produced by two other individuals with multiple accounts. This suggests that the number of responses produced by workers with duplicate accounts is much lower than simple IP examination suggests.

Duplicate responses across related experiments are a more difficult problem to resolve. Duplicate respondents spread across related experiments still violate assumptions of statistical independence when the evidence offered by individual studies is considered together. This is true when evidence is considered either quantitatively (i.e., through meta-analysis) or qualitatively (through review of evidence); replication of a finding on two different populations leads to a stronger (and fundamentally different) inference than does replicating a finding on an identical sample twice, and studies that contain a mix of repeated and new participants fall somewhere in between. This distinction bears both on the generalizability of a finding, in that subsequent replications may not consist of samples of new individuals, and on the validity of a finding, in that subsequent responses may be more likely to be consistent with prior responses for reasons other than the theoretical question of interest, such as recall of a previous response or memory of the debriefing materials. Indeed, prior knowledge about the purpose of an experiment, familiarity with an experimental manipulation, or reason to suspect deception are all known to influence participant responses, albeit in unpredictable and often paradigm-specific ways (Brock & Becker, 1966; Edlund et al., 2009; Glinski, Glinski, & Slatin, 1970; Rosnow & Aiken, 1973; Sawyer, 1975; Silverman, Schulman, & Wiesenthal, 1970).

From a requester perspective, the pool of available workers can seem limitless, and Amazon declares that the MTurk workforce exceeds 500,000 users ("Amazon Mechanical Turk Requester Tour," n.d.). On the basis of this estimate, it is tempting to assume that the likelihood of recruiting the same worker for two identical experiments is low. However, MTurk workers are likely to complete many more tasks than the typical respondent in an experimental participant pool, even if it is conservatively assumed that the typical worker spends an hour per week on MTurk. Some workers actually spend hours each day on MTurk, treating it as more or less a full-time job (Ipeirotis, 2010). Thus, while the probability that the typical (modal) worker might complete two related HITs is low, a small subset of workers may complete nearly every HIT available to them. This is particularly true if, as seems likely, HITs offered by researchers are more interesting and lucrative than HITs offered by businesses, if workers who spend more time online become better at finding desirable HITs (e.g., by using applications to monitor the activity of favored requesters), or if there are more general differences in preferences across the worker population that result in workers sorting themselves into different types of HITs.

To investigate the prevalence of duplicate respondents, we pooled data from the authors and several collaborators, resulting in a sample of 16,408 HITs (i.e., individual observations) distributed across 132 batches (i.e., academic studies). Within this sample, we found substantial reason to be concerned about duplicate responses. These HITs had been completed by a total of 7,498 unique workers. The average worker was observed to have completed 2.24 HITs (*SD* =3.19), but a small minority of workers were responsible for submitting most of the HITs. The most prolific 1 % of the workers from this sample were responsible for completing 11 % of the submitted HITs, and the most prolific 10 % were responsible for completing 41 % of the submitted HITs (see Fig. 1). This resonates with preliminary evidence from Berinsky and colleagues (2012), who found that 24 % of the workers recruited for six studies participated in two or more and 1 % completed five or more.

Researchers seem largely unaware of the possibility that workers might participate in conceptually related experiments. We conducted an informal survey on the mailing lists of the Society for Judgment and Decision Making and the Society for Personality and Social Psychology (*N* =369), and only 4 % of our respondents who reported using MTurk (7 out of 182) listed repeated participants as a possible concern with MTurk research (in contrast, more than a quarter of researchers identified data quality and worker attentiveness as issues). Moreover, only 5 of a sample of 132 published papers (available in the online supplementary material to this article) reported addressing this issue.

Some care is required if one wants to prevent workers from participating in conceptually or methodologically similar studies. Researchers have tried various solutions to this problem. Some simply ask workers who have completed prior studies from a lab group to not participate, which assumes that workers are honest and will recognize the task on the basis of the lab it comes from. Others run multiple experiments via a single link within the same HIT, which prevents repeat participants but also prevents details, such as participant payment, from changing across experiments. Finally, some solutions have been developed using modifications of third party software (e.g., Qualtrics; Peer, Paolacci, Chandler, & Mueller, 2012) or Web-server-based software that can maintain a database of previous workers (Goldin & Darlow, 2013). While these methods have their own advantages, the former method relies on a paid platform, and the latter method requires advanced technical knowledge and benefits researchers only when used at a large scale. In addition to allowing researchers to prescreen workers on a wide variety of researcher-selected dimensions, MTurk's built-in Qualification system can also be used to exclude previous workers without these disadvantages. The creation and uses of Qualifications will be discussed in detail below.

Do workers share information with each other?

MTurk workers maintain online forums where they share information and opinions about MTurk and links to particularly interesting or lucrative HITs. This can also potentially lead to foreknowledge in experimental participants spreading to workers who have never completed them before. Empirical research on college undergraduate populations has demonstrated that participants do share information with each other, at least when sufficiently motivated (e.g., when incentives are offered for a correct response; Edlund et al., 2009), and similar problems have been reported with collecting samples through online forums (Buchanan, 2000). A brief review of worker-organized discussion boards (e.g., turkernation.com, mturkforum.com) suggests that similar problems could emerge within MTurk. Workers frequently share information about HITs on specific threads dedicated to the dissemination of links to "cool" research surveys, potential obstacles to receiving payment (e.g., "attention checks"), and HITs that pay above a certain threshold. However, qualitatively, workers do seem to understand that it is in poor form to explicitly discuss research hypotheses, and when potentially important information is revealed, it is

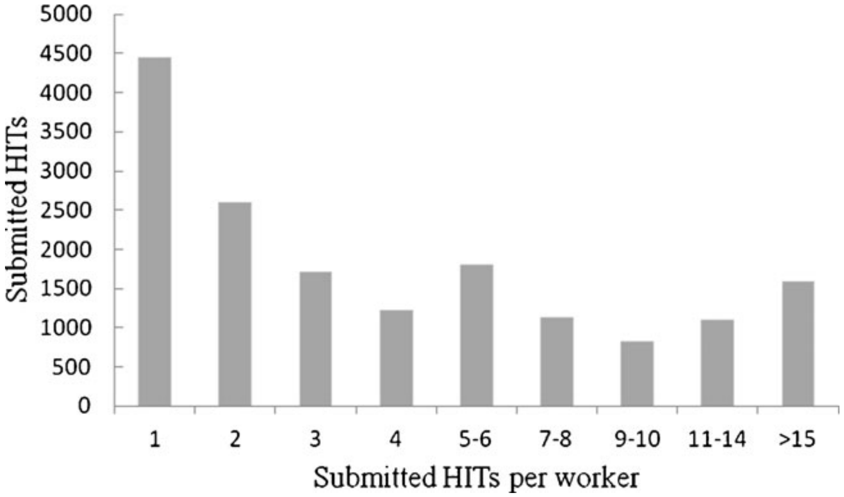

Fig. 1 Number of human intelligence tasks (HITs) completed by workers with different levels of productivity

almost always done without malice (e.g., an offhanded comment about a specific question that happens to differ across conditions). Additionally, some forums have developed specific norms against revealing pertinent study information and sanction individuals for doing so.

Study 1: Workers' reports of experience and information sharing

To gain more insight into why some workers are disproportionately likely to participate in multiple related HITs, we conducted a survey in which we asked workers about how they search for and share information about HITs. To get a better sense of the potential consequences of remaining in a common participant pool like MTurk for a long period of time, we also asked them to report whether they had participated in a number of different experimental paradigms in the past.

Method

Participants

Workers (N = 300) who reported themselves to MTurk as living at a U.S. address and whose ratio of approved/submitted HITs was higher than 95 % were recruited to take part in this survey. Since this study is descriptive and exploratory, a large sample was recruited to maximize the diversity of recruited workers. Workers were paid $0.50, and the survey took approximately 20 min to complete. Two hundred eighty-nine of these workers responded from U.S. IP addresses. We recruited participants using a RequesterID that was created specifically for this study in order to remove any reputation effects our RequesterIDs might have among the worker population.

WorkerIDs were matched to the WorkerIDs from the pool of completed HITs described earlier. Following initial data collection, this sample was supplemented with an additional 20 workers that the secondary data analysis described above identified as belonging to the most productive 1 % of workers in that sample. The final sample of highly productive workers (N = 33) represented about half of this population. Unless stated otherwise, all claims reported in this section about the MTurk population in general use only the original sample of 300 workers, while all findings that compare workers by productivity level include the supplementary sample of especially productive workers.

Our sample mirrored previously recruited samples on income, age, and education. The population was disproportionately white (80 %; 95 % CI [75.1, 84.4]) and Asian (8 %; 95 % CI [5.4, 9.6]), relative to the U.S. population as a whole (75 % and 3.6 %, respectively). Although a number of participants identified as Black (8 %, 95 % CI [5.4, 11.6], vs. 12.3 % of the population as a whole) and/or Hispanic (5.4 %; 95 % CI [3.3, 8.5]), both groups were underrepresented, as compared with the U.S. population as a whole, $\chi^2(1, N = 300) = 5.55$ and 14.09, respectively, ps < .02, ds > 0.25. Although there is nothing peculiar about the demographics of more productive workers, they tended to be somewhat older and more educated and more likely to be White than the sample as a whole.

Procedure

To gain a better understanding of workers' behavior while completing HITs, they were asked where they currently were (at work, at home, at school, in a public place like a cafe or library, or somewhere else), how many people they were with, and what other activities they were currently engaged in (listening to music, watching TV, chatting online). Workers also estimated the average time they spent using MTurk on each of the 7 previous days.

To address the issue of cross talk, we also asked participants whether they knew other workers in person and whether they knew of blogs or discussion forums dedicated to MTurk workers. If they indicated knowing other workers or participating in online forums, they estimated how often they participated in discussions of MTurk in these venues and ranked the frequency with which they discussed various topics. The purpose of these questions was to assess whether workers were directing each other to particular HITs and, if so, why. We also asked workers whether they had favorite requesters, whether these favorite requesters were academics, and whether they recalled previously participating in various commonly used research paradigms. We finally collected basic demographic data (gender, age, state of residence, education, race, and ethnicity).

Results

Worker productivity and survey response

The most productive workers did not report spending more time using MTurk than did less productive workers. Still, they were unusually likely to find and complete the survey: The most prolific 1 % of workers in the larger sample of HITs comprised 4 % of the initial worker survey sample (N = 13), and the most prolific 10 % (N = 72) comprised 25 % of the worker survey sample. There were significantly more of the most prolific 10 % of workers than would be expected by chance, $\chi^2(1, N = 300) = 63.79$, $p < .001$, $d = 1.04$.

Worker attention

Our survey revealed that although most workers completed the HIT from home (86 %; 95 % CI [81.6, 89.5]) and alone

(73 %; 95 % CI [52.7, 63.8]), they were often engaged in other activities simultaneously: 18 % (95 % CI [14.1, 22.7]) of them reported watching TV, 14 % (95 % CI [10.5, 18.4]) of them reported listening to music, and 6 % (95 % CI [3.8, 9.3]) of them were also instant messaging with at least one other person (see Table 1). If anything, these estimates may be conservative, since workers are likely to be motivated to underreport behaviors that call the quality of the data they provide into question (Behrend et al., 2011).

The most prolific workers tended to be somewhat more focused than the general pool of workers. When completing the demographic survey, they were more likely to be alone and less likely to be engaged in other tasks like listening to music, watching TV, or chatting online (see Table 1), suggesting that they may be particularly suitable for experiments that are sensitive to participant attention (e.g., those that rely on reaction time).

Participation in conceptually related experiments

In our survey, a substantial proportion of workers reported participating in some of the more common and easily describable experimental paradigms, such as the ultimatum game or the "trolley problems" commonly used to illustrate moral reasoning. As would be expected, the most productive workers are also the ones who are most likely to report participating in common experimental paradigms (see Table 2). This stands in contrast to earlier claims that MTurk offered a participant pool of naïve participants (Chilton et al., 2009). On the basis of these findings, it seems that, without taking steps to filter or identify nonnaïve participants, MTurk may not be appropriate for commonly used paradigms. For less-used paradigms, researchers can minimize the problem of duplicate respondents by sharing lists of worker IDs that have completed specific experimental manipulations and excluding them from future experiments.

Following favorite requesters

The majority of our participants (55 %; 95 % CI [49, 61]) reported having a list of favorite requesters that they monitored for available HITs, and 58 % (95 % CI [52, 63]) of those who followed favorite requesters (about a third of the entire sample) reported that this list included academic researchers. The most productive workers were especially likely to follow specific requesters (see Table 1).

Worker cross talk

In our survey, 26 % (95 % CI [21, 31]) of participants reported knowing someone else who used MTurk personally, and 28 % (95 % CI [23, 33]) reported reading forums and blogs about MTurk. However, when asked to rank the frequency with which they discussed or read about various aspects of MTurk, the actual purpose or contents of the HITs were far less important than pragmatic considerations such as pay rates or requesters' reputation (see Table 3). Only half of the respondents who actually read blogs (about 13 % of the overall population) reported *ever* seeing a discussion about the contents of a social science research study online.

Table 1 Distractedness and involvement among MTurk workers

	Overall	No Productivity Information	0–89th Percentile	90–98th Percentile	99th Percentile[a]	M–Hχ^{2}[b]
With other people	27 %	32 %	20 %	23 %	15 %	4.95*
Listening to music	14 %	18 %	10 %	10 %	0 %	8.85**
Watching TV	18 %	24 %	12 %	14 %	15 %	3.53†
Chatting online	6 %	9 %	3 %	0 %	3 %	5.45*
Read Mturk blogs	28 %	26 %	26 %	36 %	40 %	3.64†
Follow requesters	55 %	43 %	68 %	71 %	72 %	19.55***
Follow academic requesters	33 %	27 %	39 %	32 %	48 %	4.93*

Note. Percentages are the proportion of respondents who affirmed that they engaged in this particular behavior. Productivity percentiles were assigned to workers on the basis of the number of HITs completed in a 132 previous samples by 7,498 workers.

[a] Includes high-productivity workers who completed the initial questionnaire ($N = 13$) and a targeted supplemental sample ($N = 20$) recruited immediately after collection of the initial sample.

[b] Chi-square and significance tests for Mantel–Haenszel linear-by-linear association test.

†$p < .06$

*$p < .05$

**$p < .01$

***$p < .001$

Table 2 Previous exposure to common experimental paradigms

	Overall	No Productivity Information	0–90th Percentile	90–98th Percentile	99th Percentile[a]	M–Hχ^2[b]
Prisoner's dilemma	56 %	36 %	71 %	85 %	88 %	68.71***
Ultimatum game	52 %	32 %	65 %	78 %	94 %	69.12***
Dictator game	0 %	22 %	51 %	64 %	76 %	64.79***
Trolley problem	30 %	10 %	33 %	68 %	85 %	107.95***
p-beauty contest	7 %	5 %	10 %	10 %	9 %	6.68**

Note. Percentages are the proportion of respondents who affirmed that they engaged in this particular behavior. Productivity percentiles were assigned to workers on the basis of the number of HITs completed in a 132 previous samples by 7,498 workers.

[a] Includes high-productivity workers who completed the initial questionnaire ($N = 13$) and a targeted supplemental sample ($N = 20$) recruited immediately after collection of the initial sample.

[b] Chi-square and significance tests for Mantel–Haenszel linear-by-linear association test.

***p*<.01

****p*<.001

Viability of longitudinal data collection

Researchers who have tried to collect follow-up data from workers on MTurk by directly contacting participants and asking them to complete a follow-up study have typically obtained response rates greater than 60 % within the first few months of collecting data (Berinsky et al., 2012; Buhrmester et al., 2011; Shapiro et al., 2013). However, nobody has examined how difficult it is to recontact workers over longer time periods. We recontacted workers who responded to our survey 1 year later by sending three e-mails inviting them to complete an unrelated survey that paid $1.50 for 30 min. One hundred forty-two participants completed the survey, for a response rate of 44 %. As a comparison, a meta-analytic review of school-based longitudinal survey research found an average retention rate of 73 % (*SD* = 13 %) after 1 year (Hansen, Tobler, & Graham, 1990). Thus, attrition over very long time periods on MTurk is high, but recruitment rates are impressive when considered in light of the minimal efforts made to recontact participants (for a discussion of the effects of survey attrition and how to minimize it, see Ribisl et al., 1999).

A closer look at who responded to our follow-up survey revealed that the response rate was significantly higher (59 %) among workers who were known to have completed at least one HIT prior to completing our initial survey, as compared with workers who could not be identified as such (29 %), $\chi^2(1, N = 319) = 28.22, p < .001, d = 0.62$. Moreover, among workers who completed at least one HIT, the number of HITs they had completed previously was positively associated with the likelihood that they would complete the follow-up, B = .08, Wald = 8.16, p < .01, reaching 75 % among the top 10 % most productive workers.

Discussion

A small set of very productive workers are disproportionately likely to complete research HITs. It is important to note that this survey was posted using a new Requester ID and was available only for 3 days, so reputation effects of the authors cannot explain the overrepresentation of these workers. One possible explanation is that this survey had the keywords "survey," "research," and "experiment" associated with it

Table 3 Topics of conversation about MTurk

	In Person		Online	
	Mean Rank	Overall Rank	Mean Rank	Overall Rank
How much the HIT pays	5.64 (2.22)	1	4.41 (2.51)	1
How long the HIT takes	4.67 (1.80)	2	4.06 (1.78)	3
How fun the HIT was	4.08 (2.00)	3	3.77 (2.28)	6
How difficult it is to complete	3.94 (1.65)	4	3.90 (1.49)	4
How to successfully complete the HIT	3.53 (1.58)	5	3.82 (1.70)	5
Purpose of the HIT	2.98 (1.92)	6	3.64 (2.39)	7
Requester reputation	2.45 (2.02)	7	4.15 (2.15)	2

Note. Participants ranked all discussion topics by frequency. Mean rank scores are reversed so that a larger number denotes greater frequency. Overall rank is the rank order of aggregated means.

(per convention; Chilton et al., 2009) and that the productive workers in our sample were those who seek out social science research HITs (we return to this issue in Study 3). Alternatively, highly productive workers may be more likely to complete HITs simply because they are highly productive, spending more time searching for and completing HITs.

The presence of these "Super Turkers" in a survey is a mixed blessing. On the one hand, given their greater experience, it is likely that they have participated in many psychological experiments already, a conjecture supported by their greater familiarity with common research paradigms (Table 2). The most productive workers are also especially likely to read blogs about MTurk and follow specific requesters (see Table 1), increasing the likelihood that workers will participate in multiple potentially related HITs. Researchers should be aware that they may have a loyal following who may have completed their experiments in the past, may have read their debriefing materials, and will deliberately seek out their future experiments. We discuss how to avoid repeat participants in the Advanced Data Collection Techniques section below.

On the other hand, self-reports show that productive workers might behave as "professional participants" and be less distracted than the average worker. They are also comparatively easy to recontact and are thus ideal for longitudinal research projects. Consequently, this population may be useful for more involved projects, provided that their potential nonrepresentativeness is not relevant to the particular research question or they can be preselected in a manner that matches them on theoretically relevant traits to the population at large.

Worker cross talk seems to be a relatively minor problem for experimental validity, since few workers seem to remember discussions about the contents of surveys or experiments. This suggests that cross talk may not be an issue unless the information involved could increase financial reward. However, workers do on occasion talk about experiments on discussion boards (accompanied by links to the HIT), and they have been known to inadvertently share details that are a part of the experimental manipulation. In our own experience, this is especially true of "instructional manipulation checks" (Oppenheimer, Meyvis, & Davidenko, 2009), "gotcha" questions, and other techniques used to increase attention or deny payment. Indeed, even discussion boards that actively monitor and delete posts that reveal details of HITs (e.g., Reddit) allow workers to share information with each other about the *presence* of attention checks.

However, workers are quick to share information about requesters' speed of payment and tendency to reject work, both of which can have serious reputational consequences for researchers. These practices are unlikely to directly affect data quality but may impact the speed with which data can be collected or possibly influence the characteristics of workers who choose to complete the HIT. Thus, it remains a good practice for researchers to ask workers how they found the HIT at the end of their survey and to systematically monitor discussion boards that refer a lot of respondents.

Study 2: The consequences of nonnaiveté

Previous experience with research studies can have varying, and perhaps unpredictable, effects on the diagnosticity of the data provided by workers. One straightforward prediction is that foreknowledge should improve performance on tasks for which sufficient thought can lead to a verifiably correct answer (Basso et al., 1999). The cognitive reflection task (CRT; Frederick, 2005) is likely to be one such task and also happens to be commonly used on MTurk (e.g., Goodman et al., 2012; Mata, Fiedler, Ferreira, & Almeida, 2013; Paxton, Ungar, & Greene, 2012; Shenhav, Rand, & Greene, 2012; West, Meserve, & Stanovich, 2012). It consists of questions such as the following: "A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost?" Each question elicits an intuitive response that can be recognized as wrong with some additional thought. Completing this task several times or reading about the task on discussion forums increases the likelihood of discovering the correct response or at least being aware that there is a "trick" that necessitates that the question receive additional scrutiny. This is of particular importance when the task (like the CRT) is presumed to measure a stable individual difference in cognitive orientation.

Method

One hundred workers were recruited to take part in a study on "impression formation." The primary analysis involves comparing two correlations. A power analysis using G*Power indicated that a sample size of 83 would be sufficient to detect a difference in correlations of .1 and 0 between worker experience and responses on two self-report measures, assuming that the self-report measures were correlated at .6 or higher (Faul, Erdfelder, Buchner, & Lang, 2009). The same RequesterID was used as in Study 1. These studies were posted and completed more than 2 years apart. Outside of these studies, the RequesterID was used only for a small number of HITs 1 year after the first study reported here and 1 year prior to this study. To ensure an even distribution of workers across various levels of experience, two HITs were created (using the Qualification method outlined below). Workers who completed at least one and not more than four HITs (according to the data reported earlier) received a Qualification that allowed them to complete one HIT (N = 50), and workers who completed more than four HITs (approximately the top 10 % of the sample) were eligible for a separate HIT. Both HITs paid \$0.60. All WorkerIDs associated with submitted HITs satisfied these eligibility requirements, indicating that the Qualification procedure worked as intended.

Participants were first asked to rate 15 trait adjectives (0, *least favorable*, to 6, *most favorable*) or, alternatively, mark an X to indicate that a letter string was not a word. Twelve of the traits were randomly selected from a larger list. One letter string was a pseudoword ("maltated"), one word was strongly positive ("sincere"), and one word was strongly negative ("cruel"; Anderson, 1968). Participants then completed the short Need for Cognition (NFC) scale (Cacioppo, Petty, & Kao, 1984), followed by a "new" version of the CRT that contained items that workers were unlikely to have encountered before (from Finucane & Gullion, 2010) and the "original" CRT (Frederick, 2005).

Results

Workers who respond randomly should be less likely to correctly identify a nonword letter string and, on average, should also show a smaller difference in the reported valence of two words that elicit opposite and extreme evaluations (i.e., "sincere" and "cruel"). Most workers ($N = 77$) correctly identified "maltated" as a nonword and showed a large difference in the reported positivity of the words "cruel" and "sincere" ($M = 4.83$, $SD = 1.63$). We found no evidence that worker experience predicts the tendency of workers to respond randomly to surveys. The number of HITs a worker had previously completed did not predict their likelihood of correctly identifying a letter string as a nonword, $B = .06$, Wald = 2.54, $p > .1$, nor did it predict the spread between ratings of "sincere" and "cruel," $\beta = .11$, $t = 1.13$, $p = .26$.

Workers who respond randomly should also show an attenuated relationship between the positive and reverse scored items on the NFC scale. A GLM analysis with positive item NFC score, the number of previously completed HITs, and their interaction as predictors was conducted to investigate whether work experience attenuated the relationship between positive and negative NFC items. Unsurprisingly, the positive item NFC score predicted the negative item NFC score, $F(1, 96) = 141.6$, $p < .001$, $\eta^2_p = .59$. However, this effect was not moderated by the number of previously completed HITs, $F < 1$. Finally, there was no main effect of previous experience, $F(1, 96) = 1.89$, $p = .17$.

To examine how prior experience with a task might affect data validity, we examined performance on the new CRT questions and the original CRT questions. These questions are structurally identical and differ only in terms of their content. However, if workers recognize items, they may be able to "correctly" answer them, without genuine "reflection" on the correct answer. A repeated measures ANOVA (mean correct responses on new and original CRT items), with the number of known previously completed HITs as a continuous moderator, revealed a significant interaction between the CRTs and previously completed HITs, $F(1, 98) = 10.35$, $p < .001$, $\eta^2_p = .10$. Expressed in correlational terms, while the new and old CRT measures were highly correlated, $r(98) = .79$, the number of previously completed HITs did not predict performance on the new CRT items, $r(98) = .04$, n.s., but did predict performance on the old CRT items, $r(98) = .23$, $p = .02$, and these correlations were significantly different, Hotelling's $t(97) = 3.05$, $p < .01$, $d_z = 0.31$. Thus, prior experience seems to increase performance on this commonly used individual-difference measure, but not on novel but conceptually identical items.

Contrary to Frederick (2005), we did not observe a relationship between performance on either CRT and the NFC scale, rs < .13, ps > .20.

Discussion

The quality of worker data as measured by gold standard responses and consistent responding does not appear to vary as a function of prior experience. However, correct responses to a measure that anecdotally is quite common on MTurk are correlated with the number of previously completed HITs. Importantly, performance on structurally identical questions that are unlikely to have appeared on MTurk are *not* correlated with prior experience, suggesting that experienced workers' superior performance on the standard CRT indexes greater experience with the questions themselves, rather than greater reflectivity. This difference is observed even though we have no knowledge of whether specific workers have completed the CRT in the past and, instead, relies merely on the conjecture that workers who have completed many HITs are likely to encounter commonly used measures. This finding is particularly notable because the CRT is commonly treated as measuring a stable individual difference. The influence of prior experience on performance could thus conceivably inflate CRT scores observed in online samples and undermine the predictive accuracy of the CRT. While, as was noted before, the precise consequences of nonnaïveté are context- and item-specific and well beyond the scope of the present article, this finding builds on the initial differences in self-reported exposure to measures by demonstrating that general worker experience predicts performance on potentially familiar psychological measures.

Prevalence of post hoc data cleaning

A second issue arises from the strategies that researchers use to clean data collected from MTurk. Despite numerous replications of classic findings using MTurk (Buhrmester et al., 2011; Paolacci et al., 2010; Rand, 2012; Shapiro et al., 2013), worker "quality" seems a matter of persistent concern among researchers who use MTurk to collect data. In the same informal survey of researchers described earlier, more than two thirds listed worker attentiveness or data quality as their single greatest concern with MTurk. While it is clearly important to make sure that workers possess the theoretically relevant attributes necessary for an effect to emerge and effect sizes may be attenuated if workers are not sufficiently attentive or careful to provide "quality" data, it is not clear that this is more of a problem on MTurk than in any other convenience samples or than in the behavior of individuals in their daily life.

What is particularly worrying is that even a casual inspection of the papers that use data collected from MTurk reveals that workers are frequently excluded for a wide variety of reasons and that these exclusion criteria are often applied post hoc. To illustrate this, we conducted an exhaustive search of all MTurk papers published prior to December 31, 2011. According to Google Scholar, over 3,400 papers, dissertations, and conference proceedings were published that contained the words "Mechanical Turk" or "MTurk." Articles from this initial sample were selected if they met the following criteria: (1) They were classified within the social sciences by Google Scholar, and (2) they were peer-reviewed articles or conference proceedings. "Online first" publications for paper journals appearing after the search period were excluded. A full list of the selected articles is reported in the online supplemental materials.

Studies that use MTurk worker data often exclude a large number of workers who provide "questionable" data. Of the published articles, 44/132 (33.3 %) reported excluding workers on the basis of the quality of the data they provided. In these papers, on average, 15 % of the sample was excluded (range 3 %–37 %). Worse, circumstantial evidence suggests that some researchers do not report having excluded participants when, in fact, they did. Since workers are effectively unlimited and resource constraints are low, one would expect researchers to disproportionately recruit round numbers of participants (e.g., Pope & Simonsohn, 2011; Rosch, 1975). Indeed, among the studies that report excluding workers for any reason, 43 % of the initial samples are a multiple of ten. However, only 25 % of the studies that did not report excluding workers had a sample that was a multiple of ten [a significant difference; $\chi^2(1, N = 216) = 6.1, p = .01$]. Thus, it is likely that more than a third of all papers drop workers post hoc for one reason or another.

One of the most common methods used to exclude workers is to include questions with verifiable responses, such as "catch trials" that identify workers who agree with unlikely or even impossible statements (Downs, Holbrook, Sheng, & Cranor, 2010; Kittur, Chi, & Suh, 2008; for a discussion, see Goodman et al., 2012), gold standard questions with factually correct responses, or "factual manipulation checks" that ask participants to remember key details of a manipulation. However, the methods used by researchers, including the threshold number of questions a worker must correctly answer, are heterogeneous. A few researchers took the approach of monitoring the consistency of responses to the same question (e.g., Munson & Resnick, 2010), while others adopted more arbitrary exclusion criteria, including discarding outliers in responses or response latency and looking for "suspicious" responses (e.g., selecting the same answer for all questions; for a discussion, see Johnson, 2005).

Whether each of these exclusion criteria is justifiable is beyond the scope of this article and surely dependent on the specific nature of the task and sample (cf. Goodman et al., 2012; Kittur et al., 2008). However, preliminary research has cast doubt on whether any of these approaches actually improve MTurk data quality or, for that matter, what data "quality" even means (Downs, Holbrook, & Peel, 2012). Moreover, researchers are probably more likely to search for reasons why data do not support their initial intuition than to search for reasons why their intuitions appear true for spurious reasons, leading to the possibility that increasing data "quality" may inadvertently inflate researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011).

As was discussed earlier, workers can be prescreened for various attributes. We discuss how to create and assign Qualifications below, but for now, note that this includes screening workers not only for demographic characteristics or desired psychological attributes, but also for attentiveness or ability to complete tasks judged to be crucial for completion of the research HIT. Consequently, workers who provide "poor quality" data (however researchers choose to define this term) can be identified *before* completing a research HIT, rather than excluded after the fact.

The kinds of research conducted on MTurk are so varied that it is probably inappropriate to impose a one-size-fits-all set of restrictions on how samples should be restricted. However, since the justification for these measures is often described in dispositional terms (e.g., identifying native English speakers or people who take experiments seriously), we believe that these restrictions are less subjective if they are applied prior to data collection as a part of an initial prescreening task, with only those workers who qualify invited to participate in data collection. While this does not eliminate questions about the generalizability of the final sample, it does reduce concerns about excessive researcher degrees of freedom.

Using Qualifications to prescreen workers or collect longitudinal data

At the most general level, Qualifications are filters that researchers can use to ensure that only the workers they want are included in an experiment. The primary function of Qualifications is to identify desirable workers and restrict the availability or visibility of HITs to them. MTurk automatically assigns a small number of Qualifications to workers that, for instance, allow all requesters to restrict HITs to workers from specific countries or to those who have successfully completed a specified number of HITs. These Qualifications are available through dropdown menus and are frequently used by requesters.

Requesters can also create their own Qualifications. Once a requester has created a Qualification, HITs can be restricted only to those workers who possess it, just like MTurk's preexisting Qualifications. If a requester knows that workers meet specific criteria (e.g., from prior survey responses), they can automatically assign a Qualification to them. Qualifications can also be configured so that workers can request to receive them. A Qualification can be set so that it is automatically awarded to anyone who requests it, automatically to workers who respond to Qualification questions in a specific manner, or manually pending requester review.

Using the Qualification system for prescreening offers several important advantages over other means of restricting a HIT to a specific sample, such as recruiting everybody and filtering out ineligible participants, asking in the public HIT description that only participants with certain characteristics complete the HIT, or administering screening questions to participants after they have accepted the HIT. First, it is cost effective. Prescreening surveys are often short, and the cost of administering them is pennies per worker, allowing highly specific populations to be targeted without undue expense. Second, restrictions are imposed a priori, eliminating concerns that a specific subgroup became of particular interest only after the results were known. Third, to workers, there is no obvious connection between responses to a prescreening or Qualification questionnaire and eligibility for subsequent studies. In the case of requester-assigned Qualifications, workers may not even know why they are eligible for a particular study. The ability to obscure inclusion criteria may be of particular importance in cases where awareness of the inclusion criteria may influence participants' willingness to participate or their responses (e.g., race and standardized test performance; Danaher & Crandall, 2008; Steele & Aronson, 1995). A researcher can assign Qualifications that cannot be overridden by workers requesting the Qualification. Also, if a Qualification questionnaire is used, a researcher can restrict workers from attempting it more than a predetermined number of times. Together, these features prevent workers from discovering the correct eligibility criteria by answering prescreening questions multiple times.

The easiest but least versatile method for using Qualifications is through MTurk's Web interface. The Web interface is best for simple Qualifications that are manually granted to workers who have worked for the requester before. The second interface, the Command Line Tools (CLT), is a set of downloadable programs that balance functionality and simplicity. It speeds up the management of Qualifications and allows them to be assigned to workers who have never worked for the requester before. It also allows for the creation of Qualifications that workers can request themselves and that can be automatically granted. The third option is the Amazon Web Services API, which requires the use of an outside interface. Although the API is the most flexible tool for working with MTurk, we will not discuss this approach further, because it is not superior to the CLT (and is more complicated) for creating and managing Qualifications. Should those without significant programming experience want to explore the API further, we have found that boto (retrievable on the Python Programming Language Web site) is a relatively straightforward tool with which to access MTurk.

Using the Web interface to assign Qualifications

Basic Qualifications can be created without any coding knowledge from the "Manage" > "Qualification Types," tab. To create a Qualification, select "Create New Qualification Type." Name and describe the Qualification to distinguish it from other Qualifications you might create. Note that workers can see the name of the Qualification.

Once created, the Qualification can then be assigned by first selecting "Manage" > "Workers" and downloading a .csv file containing all workers who have completed previous HITs for the requester (if the Qualification is not visible, click on "Customize View" and add it). The .csv file contains columns labeled *WorkerID*, *Link to Individual Worker Page*, *Number of HITs approved or rejected* (for you), *Number of HITs approved* (for you), *Your approval rate*, *CURRENT Blocked Status*, and *UPDATE Blocked Status*.

If a requester were to create a Qualification named "Gender," two columns labeled *CURRENT–Gender* and *UPDATE–Gender* would also be visible. *UPDATE–Gender* is where the requester would assign new values to the workers. For instance, all women could be assigned the value "1." To change the value associated with a worker for a particular Qualification, place this value in the update column (or type "revoke" to assign no value), then save the file and upload it to MTurk using the "Upload CSV" button on the same page. Qualifications can be assigned to individual workers without downloading the .csv file by clicking "Manage" > "Workers" and then selecting individual WorkerIDs.

Once this Qualification is created, subsequent HITs can be restricted to women only by requiring that all eligible workers have a value of 1 assigned for the gender Qualification. These

HITs can be set to be viewable only by workers who are qualified to participate or by all workers. Note that when the Web interface is used, the population of women who can participate in the HIT is restricted to those who have previously worked for the requester. This will generally limit the available pool of workers and makes it an unviable strategy for new requesters who have conducted only a few studies.

After assigning a Qualification, two strategies can be used to recruit workers. The HIT can be posted without contacting workers, and the researcher can wait for qualified workers to find and complete the HIT. This method is relatively slow and has a somewhat lower response rate than directly contacting workers, since it requires that the worker select the HIT from all available alternatives. However, a potential advantage of this method is that it prevents workers from realizing why a HIT is visible to them or even that it might not be visible to others, minimizing demand characteristics. Alternatively, a researcher can directly contact eligible workers, either by e-mailing them individually through the Web interface or *en masse* using the API (for a tutorial on how to do this, see Mueller & Chandler, 2012).

Using Command Line Tools to assign Qualifications

CLT offers additional flexibility. In particular, CLT allows you to create Qualifications that can be assigned automatically by MTurk. Another advantage of using CLT is that these Qualifications can be assigned to all workers, even those who have not completed any HITs for the Qualification's creator. Thus, researchers can share a common list of workers to include in (or exclude from) their research with the creator of a given Qualification.

Installing the command line tools

CLT can be downloaded from Amazon (http://aws.amazon.com/developertools/694) by following the installation instructions in the MTurk documentation available online (http://bit.ly/1153tf3). On a Mac, make sure the Java and CLT settings are as indicated in the installation guide. Then, within the terminal, navigate to the "bin" folder located in the same directory as the CLT (e.g., by typing *cd /Applications/aws-mturk-clt-1.3.0/bin/*). This folder contains files that correspond with many commands you might wish to use (i.e., *grantBonus*, *createQualificationType*, etc.). You can view the files in the terminal by typing the *ls* command in a Mac Finder window. Installing CLT on a PC is simpler, since you can download the necessary Java environment in the package with the CLT.

In subsequent sections, all CLT commands will be presented for the Mac/Unix framework for clarity. Mac/Unix commands take the form: *./grantBonus.sh*. An identical command for Windows would merely include the text: *grantBonus*. We describe Mac commands because their formatting provides additional clarity about where each command begins.

Creating a prescreening Qualification that is automatically scored

CLT allows a requester to create a prescreening questionnaire that can be automatically scored. Workers are assigned a value on the basis of their responses and, importantly, with the proper command, cannot retake the questionnaire to override their initial responses (to see whether this makes better surveys visible). Thus, for example, a Qualification could be created and awarded to everyone who reports that they are a parent.

A Qualification with a questionnaire requires three text files to be created and saved in the /bin folder, which is a subfolder of the main CLT folder that is created when you download the tools. These will specify the question that will be asked to participants, the values that will be assigned for specific answers to the question, and the properties of the Qualification. Assume that a researcher wants to create a Qualification to identify parents. The question could be placed in an XML file that we will call "Parent.question." The values associated with different answers that MTurk will reference will be saved in an XML file called "Parent.answer." The properties of the Qualification will be saved in a text file called "Parent.properties." There are many potential question types (multiple choice, text entry, etc.) one can use for Qualifications, so we will not include them all here. Samples of question and answer files that can be used as templates can be found in the "samples" folder installed with CLT. The MTurk Developer Guide provides additional resources for creating Question files (http://bit.ly/1be3Rwc) and Answer files (http://bit.ly/10FVDb9).

A Qualification that is automatically granted upon request requires only the .properties text file. The .properties file specifies additional features of the Qualification. This is the file that is required for automatically granted Qualifications or any other Qualification type that does not have a questionnaire attached (e.g., a Qualification you wish to assign to your prior workers on the basis of prior survey responses). Note that, throughout, the examples that follow commands listed within parentheses are optional and the list of properties discussed here is not exhaustive. Parentheses and brackets should not be included in the text files that will be used by CLT.

The properties file for our example of the "Parent" Qualification would contain the following syntax:

```
name=[Parent]
description=[workers with children]
```

`(keywords=[parent, children] )` *—words workers might search to find this Qualification*

`(autogranted=[false] )` *—if true, it immediately awards the Qualification to anyone who requests it, though prior assigned values will not be overridden if –noretry is specified when the Qualification is created.*

`(autograntedvalue=[5] )`*—the Qualification value autogranted to workers*

`(sendnotification=[false] )` *—if true, alerts the workers that they have successfully been awarded the Qualification.*

Once these files are created, the Qualification is uploaded using the command `/createQualificationType.sh –properties Parent.properties (-question Parent.question -answer Parent.answer –noretry)`; the -noretry argument prevents a worker from attempting a Qualification more than once. This command will output a file named "Parent.properties. success."

Within this file, and also printed onscreen, is the Qualification type ID number (qualtypeid), signifying that the Qualification has been created and is ready for use. Either the qualtypeid or the .success file can be used with the proper command to assign a Qualification to an individual worker or list of workers. A HIT can then be created that requires a specific value for this Qualification, and workers will be informed that they need to complete it prior to accepting this HIT. Workers who complete a Qualification successfully or who are assigned a Qualification by a requester will automatically gain access to the HIT.

Assigning a Qualification to workers on the basis of survey response

As with the Web interface, CLT allows requesters to limit the availability of a HIT to a subset of workers for follow-up or longitudinal studies. For example, a requester may wish to follow up with parents who reported particularly high levels of parenting stress in an initial survey. While this kind of task is easy to do through the Web interface, we include the CLT description for completeness. Since these data were not collected in a Qualification, qualifying workers will need to be identified on the basis of their survey responses and assigned a new Qualification, "highlevelparents," by the requester (again, workers can see the name, so make sure you choose appropriate Qualification names). To do so, the requester needs the WorkerID codes [*workerid*], which can be found in the .csv file associated with the HIT, and a way to identify the individual workers of interest (such as a unique verification code that is included in their survey response and submitted as proof that they completed the HIT). The syntax for assigning this Qualification to an individual worker is then:

```
./assignQualification.sh -qualtypeid
[qualtypeid] -workerid[workerid] (-score
2 -donotnotify)
```

If a score is not included in the command, workers will be assigned the default value 1.

The -donotnotify option allows a Qualification that, by default, sends a notification to workers when it is granted to, instead, be assigned to workers without their awareness (note that the Qualification will still be visible if a worker views his list of Qualifications, so choose a title accordingly). This is useful for researchers who want to disguise the relationship between a Qualification and a HIT to prevent workers from knowing why they were eligible to participate in a particular HIT. If the Qualification was created with the parameter `sendnotification=false`, do not also use the `–donotnotify command`.

The Qualification can be assigned to eligible workers *en masse* by creating a tab-delimited file [*workers.txt*] containing a column for the WorkerIDs [workerid] and a column for the scores you wish to assign to them [*score*]. The syntax is then:

```
./assignQualification.sh -input
highlevelparents.properties.success -
scorefile workers.txt.
```

As was mentioned above, one important advantage of this method over the Web interface is that workers can be assigned a Qualification by a requester who has not worked with them before. Another researcher interested in stressed parents could ask the creator of the Qualification to assign it to his qualified workers as well, increasing the size of the qualified population for both researchers.

Qualifications are not binary

Qualifications earned by workers can be overwritten with a new value by the creator of the Qualification, and eligibility for specific HITs can be restricted to a single Qualification value or a range of values. In other words, Qualifications are not binary. This feature is potentially useful to requesters who want to present several different surveys that workers complete in sequence. For example, a requester could ask workers to complete ostensibly unrelated measures in a particular order (e.g., a parenting stress questionnaire and a questionnaire about child educational outcomes) and use a single Qualification to manage access to all measures and assign different values to workers who are at different stages in the sequence of questionnaire completion (e.g., "1" for workers eligible to complete the first questionnaire and "2" for workers who have completed the first questionnaire and are thus eligible to complete the second questionnaire). Alternatively, this feature can be useful for managing access to a number of different kinds of HITs that target the same population, with different HITs restricted to workers assigned different values on the Qualification.

Using Qualifications to exclude workers

Within MTurk, a system to prevent duplicate workers can be created by extending the logic of assigning Qualifications described in the section above. This can be used to prevent workers from completing conceptually related HITs or, in other circumstances, where a requester wants to avoid a particular set of workers. For example, a requester conducting research on a hard-to-access subpopulation may wish to pilot materials on ordinary workers first and ensure that the target population does not see or complete the pilot materials. To do so, first a Qualification must be created that automatically grants a numerical value (e.g., "1") to anyone who requests it. Workers who want to complete the HIT will see that the Qualification is required and request it. An automatically granted Qualification is created with syntax similar to that described in "Creating a Prescreening Question" (above), except that autogrant is set to [true] and autogrant value is set to whatever value you wish—for example, 1. You *must* use the -noretry argument to create this Qualification to prevent workers from requesting it when they already have a value assigned to them. Second, all HITs of a given type must require that a worker have a value of 1 for this Qualification in order to complete them. Third, the workers you wish to exclude (e.g., workers who complete the experiment, workers who have already completed a related study, or workers belonging to a hard-to-access subpopulation) need to be assigned a new Qualification value (e.g., "2") using either the Web interface or CLT updating methods described above. Since the HITs require a Qualification value of "1," workers who have completed one study (earning a value of 2) or workers you wish to exclude for other reasons (who have been assigned a value of 2) will be unable to participate. Finally, known workers who are eligible to complete the HIT can be preassigned a value that will make them eligible to complete the HIT without requesting the autogranted Qualification. A value assigned by the researcher cannot be overridden when the -noretry argument is used, so an excluded worker cannot simply request the automatically granted Qualification to gain access.

Sharing Qualifications with other researchers

While a requester cannot currently assign other requesters' Qualifications, all that is required to grant a worker a Qualification is their WorkerId. WorkerIds can be easily exported from MTurk and shared with other researchers. The creator of the Qualification can then assign it to these workers. Consequently, a group of researchers interested in studying a particular population could develop and maintain a common participant pool of workers who meet the criteria of interest or who should be excluded as a result of having completed similar experiments.

Study 3: Effect of Qualifications on completion time and sample characteristics

One legitimate concern is that requiring Qualifications may slow down the speed of data collection. Since completion time is constrained by the total number of workers who can potentially complete a HIT, the data collection speed of a batch that requires a requester-granted Qualification depends on the number of workers that can be assigned this Qualification—that is, the number of known WorkerIDs regardless of workers' prior work history. However, it is less clear what effect requiring workers to request an autogranted Qualification will have on their willingness to accept a HIT. Finally, a related question is whether including terms related to academic research may accelerate the rate at which workers accept a task. We divided another research study into four batches to address these questions and examine the rate at which HITs are accepted by workers.

Method

Four conditions were created, each consisting of 100 HITs (total $N = 400$). Since this study is descriptive and exploratory, a large sample was recruited to maximize the diversity of recruited workers. All HITs were restricted to U.S. workers with an approval rate above 95 %. Workers were told that the HIT took less than 10 min to complete and were paid $0.60. The HITs were launched simultaneously to eliminate the influence of time or day of the week.

The *no-keyword* condition had no further restrictions as to who could participate and no keywords to aid workers in their search. It was posted last, so that workers searching for recently posted HITs would see it prior to the other three. The *keyword* condition was identical to the no-keyword condition, except that it was tagged with the keywords "psychology," "survey," "academic," "research," and "university" to assist workers in finding the HIT. The *previous workers* condition was identical to the keyword condition, except that eligibility to complete the HIT was granted prior to launch to workers who were known to have completed at least one HIT and not more than four HITs in the past ($N = 13{,}715$). Other workers had to request a Qualification that was automatically granted. The *autogrant-only* condition required that all workers request an automatically granted Qualification before they were eligible to complete the HIT. For logistical reasons, workers were allowed to participate in as many of the four HITs as they wanted to.

Results

Fifteen HITs were submitted by a worker who had already completed at least one of the other three HITs, leaving a total sample of 385 workers.

Search completion time

For all conditions except the autogrant-only condition, HIT acceptance peaked early and trailed off from hour 1 to hour 5, with completion rates picking up again after this point in time (Fig. 2). The no-keyword HIT (the first one visible to workers in the default search setting) was completed fastest, although this was thanks in part to being posted by workers on MTurk forums. A single HIT remained unclaimed until 6 hours later, but it is likely that this HIT was accepted and then returned uncompleted by a worker, making it unavailable for at least part of this time. Workers completed HITs in the keyword condition at a steadier pace, also finishing around hour 6. HITs that required a Qualification to complete took somewhat longer, but all four HITs were effectively completed by hour 8. The HITs accepted the following day were likely accepted by workers on the first day but, ultimately, not completed, since a visual inspection after hour 8 revealed that there were no HITs available to workers, despite these HITs remaining uncompleted. Thus, all four approaches led to the relatively rapid recruitment of participants, albeit at different rates.

Effects of HIT design on completion time

In general, conditions were completed in the order in which HITs were visible to requesters by default, and surprisingly few workers completed more than one HIT. Thus, it is possible that workers completed the first HIT they saw and avoided subsequent HITs that appeared identical and that this contributed to the slower speed at which the Qualification-required HITs were completed. However, there is additional evidence that suggests that autogranted Qualifications are a disincentive to workers. Workers who were preassigned the necessary Qualification to complete the previous workers condition were far more likely to complete this HIT (82 %) than they were to completed the condition where they needed to request a Qualification to participate (autogrant only; 9 %), a significant difference, $\chi^2(1, N = 193) = 98.2$, $p < .001$, $d = 2.1$.

Effects of HIT design on sample composition

Replicating Study 1, the most productive workers were disproportionately represented in the sample. The top 10 % most productive workers made up 20 % of the keyword, no-keyword, and autogrant-only conditions (conditions in which workers were treated equally regardless of experience), significantly higher than chance, $\chi^2(1, N = 287) = 21.02$, $p < .001$, $d = 0.56$.

Although more productive workers reported deliberately seeking out academic HITs in Study 1, worker productivity did not predict whether workers accepted the keyword or no-keyword HIT, $B = .19$, Wald = 1.13, $p = .29$. However, we found that more productive workers did tend to avoid HITs for which they did not possess the Qualification. Comparing

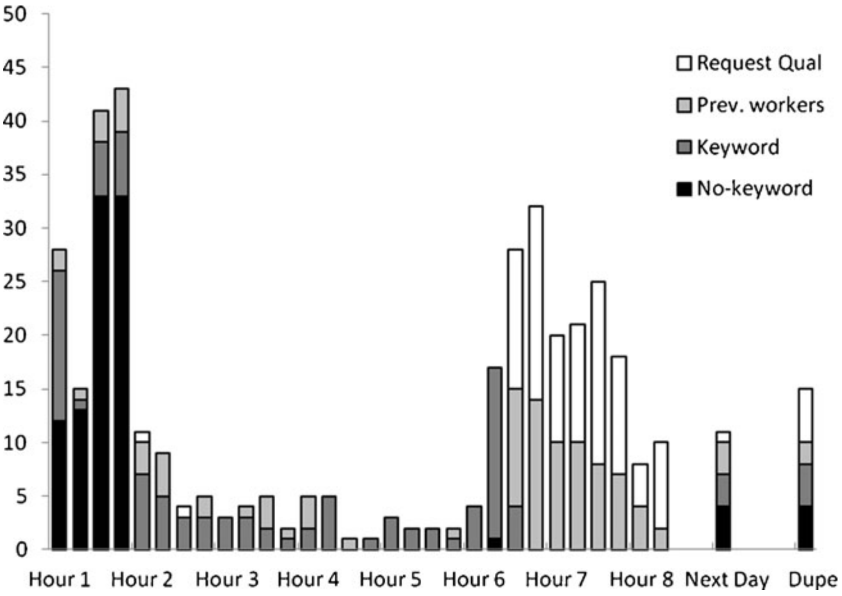

Fig. 2 Human Intelligence Tasks (HITs) completed by condition over time

the keyword, no-keyword, and autogrant-only conditions revealed that increased worker productivity was associated with a more general tendency to avoid the autogrant-only condition, $B = -.14$, Wald = 7.98, $p < .01$. This cannot be explained by moderately productive workers (one to four previous HITs) having the opportunity to complete an additional HIT (the previous workers condition) without requesting a Qualification, because this effect remains significant when only highly productive (more than five previous HITs) and new workers are compared, $B = -.13$, Wald = 8.48, $p < .01$. Thus, the marginally slower speed associated with Qualifications may be worthwhile in situations where highly productive workers are not desired.

One unexpected insight was the importance of forums as a source of workers. At the end of the survey, we asked workers to indicate how they had found it. Thirty-one workers in the no-keyword condition and 20 workers in the keyword condition reported seeing the HIT on a forum post, with the earliest mentions of forum posts within the first hour. The previous workers and autogrant-only conditions had proportionately larger populations who reported finding the HIT in forums (Ns = 50 and 82, respectively), perhaps reflecting the initially slow completion rates by workers using the traditional interface and the longer time period for which they were available. Supporting this interpretation, the large uptick in responses observed in hour 6 is largely attributable to workers who found the HITs on Reddit.

One possibility is that productive workers find surveys through sharing information through forums; after all, as was noted in Study 1, they tend to report reading more forum posts. However, an examination of the keyword, no-keyword, and autogrant-only conditions (conditions in which workers are treated equally regardless of experience) revealed that more productive workers were *less* likely to come from forums, $B = -.08$, Wald = 6.47, $p = .01$, suggesting that forum posts cannot explain the unusually high proportion of highly productive workers. Furthermore, we should note that workers who found the HIT through forums were younger ($M_{\mathrm{age}} = 28.9$, $SD = 8.7$) and predominantly male (66 %), as compared with workers recruited from MTurk at large ($M_{age} = 34.5$, $SD = 12.4$; 48 % male). Most of these workers came from Reddit.

Discussion

In sum, while all four conditions resulted in reasonably fast completion rates, there are differences in how quickly they allow workers to be recruited. Allowing workers to autogrant themselves access to a Qualification will allow previously uncontacted workers the opportunity to complete a HIT. However, workers are unlikely to request autogranted Qualifications unless other workers have vouched for the worthiness of the HIT on a forum. This is a mixed blessing, since it slows data collection (although in this case, not to a detrimental degree) and may bias the sample toward a specific subset of MTurk workers, but it also seems to act as a deterrent to the most productive workers. There are two reasons why workers may avoid HITs that require Qualifications. First, the autogrant function adds an additional step to completing a HIT, making task completion less efficient. Second, relatively few requesters use Qualifications, and thus it is likely that workers are currently unfamiliar with them. It is possible, although not certain, that workers will become more willing to accept autogranted Qualifications as they become more prevalent.

For researchers who wish to use Qualifications, data collection can be accelerated in two ways. Since workers are not returned directly to a HIT after completing a Qualification, requesters should make HITs that require an autogranted Qualification easily searchable and should inform workers of the terms to use in order to find it again once they receive the Qualification. Requesters should also prequalify as many workers as possible by using a list of their own workers who have previously completed unrelated HITs, supplemented by lists drawn from other researchers.

The potential necessity of worker word of mouth for autogrant-only Qualifications is problematic for two reasons. First, data collection is dependent on HITs being posted within a forum. Second, forum populations may differ qualitatively from the population as a whole. Although the precise differences are beyond the scope of this article, it is illustrative that in this study, workers recruited from forums were much younger and mostly male. They may also differ in undesirable ways in other demographic characteristics, motivations, and mindsets, raising more general questions about the consistency of MTurk samples, since their final composition may be dependent on details that are only partially under researcher control. Future research on MTurk, particularly research on the effects of payment characteristics on sample type, could fruitfully address this issue. Additionally, research that uses MTurk samples should take care to collect and report demographic information, rather than relying exclusively on prior research about the representativeness of MTurk as a whole.

Finally, the source of the most productive workers remains something of a mystery. While it is possible that there is a weak tendency for productive workers to complete HITs containing academic keywords and that this tendency can become more apparent in larger samples, it alone cannot explain their overrepresentation in this study. Likewise, productive workers are no more likely to report finding a survey in a forum than are less productive workers. Finally, requester reputation effects could not have influenced participation rates in Study 1 and could have had only a minimal influence in Study 3. It is possible that other, unexamined factors contribute to the prevalence of the most productive workers. Alternatively, productive workers may be more likely to appear in HITs simply because they are more productive.

Concluding comments

MTurk is widely perceived as a cheaper, faster, and more convenient surrogate for traditional convenience samples. While it does possess all of these attributes, thinking about it in only these terms may obscure important differences between MTurk and other methods of data collection, while producing undesirable trade-offs. In particular, among researchers, there seems to be a lack of awareness about the possible nonindependence of observations and a perception that MTurk data quality is poor despite several empirical investigations that have failed to support this claim.

While cross talk between workers seems to be minimal and focused on the payment and general appeal of HITs (rather than their specific contents), as was demonstrated in Study 3, it can introduce sampling issues, since some forums may not be made up of populations that are as representative of the general population as MTurk is as a whole. The potential for forum posts to drive traffic surprised us and highlights issues that could benefit from future investigation. However, it suggests that, in principle, samples recruited from MTurk should not be assumed to be as representative as those in larger previous studies, particularly when features make HITs more or less attractive (e.g., Qualifications, payment level). This finding also suggests that making a HIT too attractive may, in fact, make the final sample less desirable.

The prevalence of nonnaïve respondents is another issue that is currently not appreciated by researchers. The seeming vastness of the pool of available workers can lead researchers to assume that workers recruited from MTurk are "less experienced as participants" than traditional participant pools (e.g., Horton, Rand, & Zeckhauser, 2011). However, it turns out that many workers have completed dozens, and likely hundreds, of experiments and surveys. As one reviewer pointed out, high participation rates are an issue for all participant pools that allow members to participate in multiple research studies. This is true, and the concerns we highlight are of more general applicability to other samples. However, this problem is of particular interest within MTurk, because workers have fewer practical constraints on the number of research studies they participate in (college campuses are not open 24 h a day, and students eventually graduate), and MTurk has a built-in infrastructure that can regulate worker participation that is largely unknown and, thus, underused.

These findings highlight that although large, the pool of workers available is not limitless. The ease and low effort of data collection enabled by crowdsourcing Web sites such as MTurk may make it tempting for researchers to quickly collect data, with little thought about the underlying quality of the methods used. It is beyond the scope of this article to discuss the deontological aspects connected to crowdsourcing. Instead, we merely note that there are practical reasons why the research community should avoid overusing shared participant pools such as MTurk. For more commonly used methods and measures, the pool of MTurk workers presents a "commons dilemma" for researchers: It should not be assumed that respondents are naïve, and groups of researchers would be better off if they could coordinate their recruitment efforts.

In addition to the general tendency for some workers to complete more research studies than others, each individual requester is likely to have yet more workers who are particularly likely to participate in their specific HITs, because workers follow individual requesters. Thus, the central concern is not whether to permit or deny workers who are generally prolific, as much as it is to minimize the participation of workers who have completed the specific tasks in which the researcher is interested. For individual researchers, several concrete steps can be taken to reduce the problem of nonnaïve participants. Knowledge of some popular experimental designs has saturated the population of those who quickly respond to research HITs. The value of off-the-shelf experimental manipulations and measurements of psychological variables is highly suspect in the best of circumstances (for a discussion, see Ellsworth & Gonzalez, 2003) and becomes more so if participants have completed them umpteen times before. Thus, we recommend that researchers who care about participant naïveté avoid commonly used paradigms and, at minimum, make an effort to measure whether participants have participated in similar experiments before. Furthermore, researchers using unique paradigms should track who has previously participated in their experiments, make efforts to exclude participants from future similar experiments, and ideally should coordinate worker recruitment with other researchers conducting similar experiments on MTurk.

Turning to the issue of "data quality," while we are sympathetic to researchers' concerns about data quality and agree that numerous contextual features (perceived anonymity, financial motivations, etc.) may tempt some workers to invest little effort into the HITs they complete, the current norms around data exclusion are troubling. Most of the techniques currently used involve excluding workers post hoc for reasons that vary in their arbitrariness and that exclude significant quantities of workers. Although, on their own, each of the various data quality measures used by researchers appears defensible, in aggregate, the frequency with which they are used and the number of workers they exclude are troubling. Moreover, from an objective standpoint that does not rest on the detection of hypothesized results, there is no strong evidence that they improve data quality (Downs et al., 2012). From a theoretical perspective, worker attentiveness seems more useful as a moderator variable, since it is not clear that inattentiveness necessarily reduces data quality for all phenomena of interest (e.g., Bodenhausen, 1990), nor is it clear that people are especially attentive in any other aspect of their day-to-day life (Fiske & Taylor, 1984). However, there are situations under which it may be justifiable to exclude workers for either practical or theoretical reasons. In these cases, the perception that these methods are arbitrary and motivated can be eliminated by using them as a prescreening tool, rather than through post hoc filtering.

Most important, we believe that the single most important advantage of MTurk over other online recruitment tools is often ignored. Although the identity of workers is typically not known, their responses across different HITs can be monitored, integrated, and managed. This offers a potential solution for the a priori management of workers according to their "quality" or other attributes. It also presents an opportunity to try more interesting methods that require prescreening participants and spreading measures across multiple HITs that are temporally separate. These features can further be combined with other unique attributes of the MTurk platform, such as the ability to pay variable rate bonuses and quickly assemble groups of workers (e.g., Suri & Watts, 2011) to produce experimental designs that are more elaborate than surveys or text-based experiments.

Author Note Jesse Chandler, Postdoctoral Research Associate, Woodrow Wilson School of Public Policy, Princeton University (jjchandl@umich.edu), Pam Mueller, Graduate Student, Department of Psychology, Princeton University (pamuelle@princeton.edu); Gabriele Paolacci, Assistant Professor, Department of Marketing Management, Rotterdam School of Management, Erasmus University (gpaolacci@rsm.nl).

Jesse Chandler is now at PRIME Research, Ann Arbor, MI and The Institute for Social Research, University of Michigan.

The authors wish to thank John Myles White for help developing and testing the API syntax and Elizabeth Ingriselli for her help coding data.

Correspondence concerning this article can be addressed to any of the authors.

References

Amazon Mechanical Turk Requester Tour. (n.d.). Retrieved from https://requester.mturk.com/tour

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology, 9*(3), 272

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist, 13*(3), 283–292. doi:10.1076/clin.13.3.283.1743

Behrend, T., Sharek, D., Meade, A., & Wiebe, E. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*(3), 800–813. doi:10.3758/s13428-011-0081-0

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*(3), 351–368. doi:10.1093/pan/mpr057

Bodenhausen, G. V. (1990). Stereotypes as judgmental heuristics: Evidence of circadian variations in discrimination. *Psychological Science, 1,* 319–322. doi:10.1111/j.1467-9280.1990.tb00226.x

Brock, T. C., & Becker, L. A. (1966). 'Debriefing' and susceptibility to subsequent experimental manipulations. *Journal of Experimental Social Psychology, 2,* 3–5. doi:10.1016/0022-1031(66)90087-4

Buchanan, T. (2000). Potential of the Internet for personality research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 121–140). San Diego: Academic Press.

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. doi:10.1177/1745691610393980

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306–307. doi:10.1207/s15327752jpa4803_13

Chandler, J., Paolacci, G., & Mueller, P. (2013). Risks and rewards of crowdsourcing marketplaces. In P. Michelucci (Ed.) *Handbook of Human Computation*. New York: Sage.

Chilton, L. B., Horton, J. J., Miller, R. C., & Azenkot, S. (2009). Task search in a human computation market. In Proceedings of the ACM SIGKDD workshop on human computation (pp. 1–9). In P. Bennett, R. Chandrasekar, M. Chickering, P. Ipeirotis, E. Law, A. Mityagin, F. Provost, & L. von Ahn (Eds.), *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation (77–85)*. New York: ACM. doi:10.1145/1837885.1837889

Cooper, S., Khatib, F., Treuille, A., Barbero, J., Lee, J., Beenan, M., . . . Foldit Players (2010). Predicting protein structures with a multilayer online game. Nature, 466, 756–760. doi:10.1038/nature09304

Danaher, K., & Crandall, C. S. (2008). Stereotype threat in applied settings re–examined. *Journal of Applied Social Psychology, 38*(6), 1639–1655. doi:10.1111/j.1559-1816.2008.00362.x

Downs, J. S., Holbrook, M., & Peel, E. (2012). *Screening Participants on Mechanical Turk: Techniques and Justifications*. Vancouver: Paper presented at the annual conference of the Association for Consumer Research. October 2012.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? Screening Mechanical Turk workers. In *Proceedings of the 28th international conference on Human factors in computing systems* (pp. 2399–2402). New York: ACM. doi:10.1145/1753326.1753688

Edlund, J. E., Sagarin, B. J., Skowronski, J. J., Johnson, S. J., & Kutter, J. (2009). Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk. *Personality and Social Psychology Bulletin, 35,* 635–642. doi:10.1177/0146167208331255

Fiske, S. T., & Taylor, S. E. (1984). Social cognition. New York: Random House

Ellsworth, P. C., & Gonzalez, R. (2003). Questions and comparisons: Methods of research in social psychology. In M. Hogg & J. Cooper (Eds.), *The Sage Handbook of Social Psychology* (pp. 24–42). London: Sage Publications, Ltd.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. doi:10.3758/BRM.41.4.1149

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging, 25*(2), 271. doi:10.1037/a0019106

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Gaggioli, A., & Riva, G. (2008). Working the Crowd. *Science, 12,* 1443. doi:10.1126/science.321.5895.1443a

Glinski, R. J., Glinski, B. C., & Slatin, G. T. (1970). Nonnaivety contamination in conformity experiments: sources, effects, and implications for control. *Journal of Personality and Social Psychology, 16,* 478–485. doi:10.1037/h0030073

Goldin, G., Darlow, A. (2013). TurkGate (Version 0.4.0) [Software]. Available from, http://gideongoldin.github.com/TurkGate/

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*.

Gosling, S., Vazire, S., Srivastava, S., & John, O. (2004). Should we trust web-based studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist, 59,* 93–104. doi:10.1037/0003-066X.59.2.93

Hansen, W. B., Tobler, N. S., & Graham, J. W. (1990). Attrition in Substance Abuse Prevention Research. *Evaluation Review, 14,* 677–685. doi:10.1177/0193841X9001400608

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 4,* 399–42. doi:10.1007/s10683-011-9273-9

Ipeirotis, P. (2010). Demographics of Mechanical Turk. *CeDER-10–01 working paper*, New York University.

Johnson, J. A. (2005). Ascertaining the validity of Web-based personality inventories. *Journal of Research in Personality, 39,* 103–129. doi:10.1016/j.jrp.2004.09.009

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the ACM conference on human factors in computing systems* (pp. 453–456). New York: ACM.

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35–60). New York: Academic Press.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., . . . Vandenberg, J. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. Monthly Notices of the Royal Astronomical Society, 389(3), 1179-1189

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23. doi:10.3758/s13428-011-0124-6

Mata, A., Fiedler, K., Ferreira, M. B., & Almeida, T. (2013). Reasoning about others' reasoning. *Journal of Experimental Social Psychology.*

Mueller, P., & Chandler, J. (2012). Emailing Workers Using Python (March 3, 2012). Available at SSRN: http://ssrn.com/abstract=2100601

Munson, S. A., & Resnick, P. (2010). Presenting diverse political opinions: How and how much. In E. Mynatt, G. Fitzpatrick, S. Hudson, K. Edwards, & T. Rodden (Eds.), *Proceedings of the 28th International Conference on Human Factors in Computing Systems* (pp. 1457–1466). New York: Association for Computing Machinery. doi:10.1145/1753326.1753543

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872. doi:10.1016/j.jesp.2009.03.009

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

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177.

Peer, E., Paolacci, G., Chandler, J., & Mueller, P. (2012). Selectively Recruiting Participants from Amazon Mechanical Turk Using Qualtrics (May 2, 2012). Available at SSRN: http://ssrn.com/abstract=2100631

Pope, D., & Simonsohn, U. (2011). Round numbers as goals: Evidence from baseball, SAT takers, and the lab. *Psychological Science, 22*(1), 71–79.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299,* 172–179. doi:10.1016/j.jtbi.2011.03.004

Reips, U. D. (2000). The Web experiment method: Advantages, disadvantages and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 89–114). San Diego: Academic Press.

Ribisl, K. M., Walton, M. A., Mowbray, C. T., Luke, D. A., Davidson, W. S., & Bootsmiller, B. J. (1999). Minimizing participant attrition in panel studies through the use of effective retention and tracking strategies: Review and recommendations. *Evaluation and Program Planning, 19,* 1–25. doi:10.1016/0149-7189(95)00037-2

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7*(4), 532–547.

Rosnow, R. L., & Aiken, L. S. (1973). Mediation of artifacts in behavioral research. *Journal of Experimental Social Psychology, 9*(3), 181–201. doi:10.1016/0022-1031(73)90009-7

Sawyer, A. G. (1975). Demand artifacts in laboratory experiments in consumer research. *Journal of Consumer Research, 1*(4), 20–30. doi:10.1086/208604

Shapiro, D. N., Chandler, J. J., & Mueller, P. A. (2013). Using Mechanical Turk to Study Clinical and Subclinical Populations.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology. General, 141*(3), 423.

Silverman, I., Shulman, A. D., & Wiesenthal, D. L. (1970). Effects of deceiving and debriefing psychological subjects on performance in later experiments. *Journal of Personality and Social Psychology, 14*(3), 203–212. doi:10.1037/h0028852

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43*(1), 155–167. doi:10.3758/s13428-010-0039-7

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811.

Summerville, A., & Chartier, C. R. (2012). Pseudo-dyadic "interaction" on Amazon's Mechanical Turk. *Behavior Research Methods*, 1-9. doi:10.3758/s13428-012-0250-9

Suri, S., & Watts, D. J. (2011). Cooperation and Contagion in Web-Based, Networked Public Goods Experiments. *PLoS One, 6*(3), e16836. doi:10.1371/journal.pone.0016836

von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science, 321,* 1465–1468. doi:10.1126/science.1160379

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology, 103*(3), 506–519.