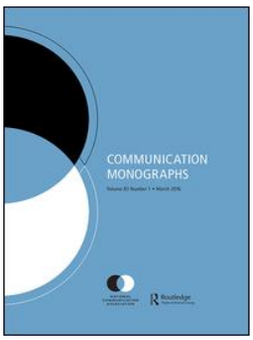

Routledge
Taylor & Francis Group

Communication Monographs

ISSN: 0363-7751 (Print) 1479-5787 (Online) Journal homepage: http://nca.tandfonline.com/loi/rcmm20

# Crowdsourcing research: Data collection with Amazon's Mechanical Turk

Kim Bartel Sheehan

**To cite this article:** Kim Bartel Sheehan (2017): Crowdsourcing research: Data collection with Amazon's Mechanical Turk, Communication Monographs, DOI: 10.1080/03637751.2017.1342043

**To link to this article:** http://dx.doi.org/10.1080/03637751.2017.1342043

Published online: 04 Jul 2017.

Submit your article to this journal

Article views: 14

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://nca.tandfonline.com/action/journalInformation?journalCode=rcmm20

**Download by:** [North Carolina State University] **Date:** 07 July 2017, At: 08:53

COMMUNICATION MONOGRAPHS, 2017
https://doi.org/10.1080/03637751.2017.1342043

Check for updates

# Crowdsourcing research: Data collection with Amazon's Mechanical Turk

Kim Bartel Sheehan

School of Journalism and Communication, University of Oregon, Eugene, OR, USA

**ABSTRACT**

Researchers in a variety of disciplines use Amazon's crowdsourcing platform called Mechanical Turk as a way to collect data from a respondent pool that is much more diverse than a typical student sample. The platform also provides cost efficiencies over other online panel services and data can be collected very quickly. However, some researchers have been slower to try the platform, perhaps because of a lack of awareness of its functions or concerns with validity. This article provides an overview of Mechanical Turk as an academic research platform and a critical examination of its strengths and weaknesses for research. Guidelines for collecting data that address issues of validity, reliability, and ethics are presented.

**ARTICLE HISTORY**
Received 7 February 2017
Accepted 18 May 2017

**KEYWORDS**
MTurk; Amazon's Mechanical Turk; crowdsourcing; research pools; guidelines; validity

In the eighteenth century, an Automaton Chess Playing machine called "The Turk" bested dozens of European chess players. Fifty years after it was introduced, "The Turk" was unmasked as a hoax with a chess master hidden inside the contraption (Morton, 2015). With apologies to IBM Deep Blue's chess playing expertise, one lesson from "The Turk" is that people are superior to machines at solving certain problems. In the twenty-first century, the essence of "The Turk" is depicted through a service from Amazon called Mechanical Turk (MTurk), which Amazon calls "artificial artificial intelligence" (Barr & Cabrera, 2006, p. 24) by connecting Internet users who are willing to accomplish small tasks for pay with companies and individuals that want to tap into the workforce.

MTurk is just one example of using what is known as crowdsourcing, or the practice of using the crowd (people participating in an online site) to complete a variety of tasks (Hitlin, 2016). Since 2006, researchers have used MTurk to pay workers to complete surveys, participate in experiments, and conduct content analyses: about a third of all available work on MTurk is in the academic realm (Hitlin, 2016). Hundreds of published studies in the social science disciplines, notably in marketing, psychology, and political science, have utilized data collected from MTurk workers. For example, in *the Journal of Consumer Research*, data collected from crowdsourcing websites have been utilized in more than 40% of the studies published in the past five years (Goodman & Paolacci, 2017). Other disciplines, such as communication and sociology, are later entrants to the

**CONTACT** Kim Bartel Sheehan ksheehan@uoregon.edu
© 2017 National Communication Association

crowdsourcing arena, although most journals in the communication realm are now publishing articles where data were gathered from MTurk workers.

As more researchers become aware of MTurk and consider its usage, numerous questions arise about its use and value for academic research. At one level, collecting data on MTurk has many of the same issues as any type of data collected online, as these samples all reflect some degree of self-selection and difficulty ensuring privacy during the research process. On a different level, the payment aspect of MTurk presents a unique set of issues. This study addresses several questions that many communication researchers, as well as researchers across multiple disciplines, have about MTurk, specifically:

(1) What is crowdsourcing and how can MTurk be used for academic research?
(2) Who are the workers at MTurk? What are the strengths and weaknesses of this respondent group?
(3) Are data collected from workers reliable and valid?
(4) What are best practices for using MTurk for communication research?
(5) What are the ethical issues surrounding MTurk?

## What is crowdsourcing and how can it be used for academic research?

Howe (2006) defined crowdsourcing as a "business practice that means literally to outsource an activity to the crowd" (p. 2). Alonso and Lease (2011) differentiated crowdsourcing from other business practices where specific activities are assigned to an in-house employee. Crowdsourcing is a model for distributed problem-solving that utilizes a group of individuals to provide solutions to problems (Brabham, 2010). Crowdsourcing can take many forms. Wikipedia, for example, uses individuals in the crowd to create, update, and review entries on the service. News outlets encourage "the crowd" to contribute first-person video to their websites. The snack brand Doritos has also used the crowd to create television commercials that have aired on the SuperBowl.

Numerous services, including MTurk, CrowdFlower and Prolific Academic, provide small businesses, market researchers, and academics with a diverse, on-demand, and scalable workforce to complete small tasks. Depending on the service, a worker could have hundreds of tasks from which to choose and complete at his or her convenience. MTurk has become a popular site for both workers and researchers. The Amazon name is well known and fairly well trusted, and there are few barriers to register and participate for workers and researchers based in the U.S.: potential workers certify they are at least 18 years old and provide authorization for a bank account. MTurk provides an attractive alternative to other samples (such as college student samples or online panel samples) for quantitative data collection. One of the major reasons why MTurk is attractive to researchers is that data can be collected quickly. For example, a study needing 300 respondents can be completed within a few hours. MTurk's respondent pool is also much more diverse than a typical student sample (Sheehan & Pittman, 2016). In addition, the costs for data collection tend to be lower than costs charged by other sample providers because the researcher sets the cost per response. For instance, a researcher may have to pay $5 per respondent for a 10-minute survey on Opinion Outpost (a site used by Qualtrics and other firms); the same study might cost less than $1 per response on MTurk (Sheehan & Pittman, 2016). Researchers from the U.S., the United Kingdom, Australia,

France, Germany, and the Netherlands are currently able to collect data via MTurk. The appendix provides links to information on how MTurk works and to services that assist researchers outside the U.S. in collecting data on MTurk.

The process for collecting data is fairly straightforward. Once registered at the MTurk site, academic researchers (called Requesters on MTurk) post their research projects (HITs) at the site, indicating how many workers are needed to complete the task (e.g., 300 workers to complete a survey). Researchers can help ensure that the best available workers complete HITs by setting qualifications for workers in terms of location (e.g., allowing only workers in the U.S. to complete the HIT), the number of HITS a worker has completed (e.g., selecting workers who have already completed at least 100 HITs), and the percent of completed HITs accepted (e.g., selecting workers who have had the vast majority – 95% or more of their HITs – accepted by other requesters). Workers who meet the qualifications then complete HITs in the time frame specified by the researcher (e.g., within three hours) and often get paid for their work within hours or days of completion. Payments for completed tasks are transferred directly from a Requester's account to a worker's account, and a worker's account can be linked directly to his or her bank account. One does not need any specific technical or programming knowledge to participate as either a worker or Requester because the MTurk platform allows researchers to easily link to online surveys or experiments hosted on platforms such as Qualtrics and Survey Monkey.

Although Amazon does not charge Requesters to limit workers to those living in a specific geographic area (such as "only people who are in the United States" or "only people living in New Jersey"), additional charges are levied if Researchers wish to limit workers to those who fit specific characteristics based on demographics (e.g., age, employment status, gender, income, education or marital status), online usage (e.g., whether someone is a blogger, how much time is spent online, or if the worker has an account on a platform such as Facebook or Linked In), product ownership (e.g., car ownership) and other characteristics (such as military service, political affiliation, dominant hand usage). Amazon has collected these data from workers who volunteer to provide this information independent of any paid study, and not all workers have provided this information. The costs per respondent for this automated screening ranges from .05 per respondent (e.g., for whether someone has a Twitter account) to .50 (for someone who is married). It is unclear, however, how many workers have provided this information to Amazon and how it effects the respondent pool.

Amazon makes money by charging fees on top of the payments that researchers pay workers: fees range from 20% to 40% of worker payments on top of any fees for specific characteristics (see appendix for links to current fee information). Academic surveys that require more than 10 respondents incur the 40% fee. Cost estimates are therefore fairly straightforward to calculate: a study of 100 people who live in the U.S. that pays $1.00 per survey would cost a Requester $140 (100x$1 × 1.4). The same study limiting the 100 people to those living in the U.S. who are married and have Twitter accounts would cost $217 (100*1.55*1.4).

Researchers wishing to conduct repeated measures experiments and longitudinal studies requiring the use of the same respondent pool can do so with MTurk. Additionally, respondents can exclude respondents from one study in participation from additional studies (or additional time points, if collecting longitudinal data). This can be accomplished by collecting the user IDs from the first study inviting participants to take an

additional study (in order to do a follow up study) or excluding the participants from participating in an additional study (in order to exclude workers) (see Peer, Paolacci, Chandler, & Mueller, 2012, for more detailed instructions). For those who might find this procedure too complex, another alternative is to utilize the service TurkPrime (www.turkprime.com). TurkPrime connects into the MTurk programming interface and allows Researchers an increased level of control over HITs (Litman, Robinson, & Rosenzweig, 2015). Benefits of TurkPrime are the ability to easily set up longitudinal studies, automated inclusion and exclusion of workers to studies, and notification to workers of the creation of HITs (Litman et al., 2015). What remains to be seen is whether TurkPrime can generate the necessary retention rates for valid samples; Abshire et al. (2017) recommend an 80% retention rate for longitudinal studies.

TurkPrime also allows Researchers to break down a larger survey into "microbatches" of fewer than 10 respondents each with each microbatch excluding people who have already completed the study. This provides Researchers with two important benefits. First, the microbatches can be launched at different times during the day, which can increase sample representativeness. Since many workers enter in and out of MTurk during a day, collecting data at different times of the day reduces the bias from collecting data from workers only on the service on Monday mornings. In addition, utilizing the microbatch feature can reduce the Amazon overhead costs from 40% to 20% if microbatches require fewer than 10 workers, resulting in savings to the researcher. One downside of this approach, however, is that the total time for data collection will be increased.

Amazon itself takes a "hands-off" role in administering the site; its terms of agreement give power to Requesters, providing them with complete autonomy. Requesters set payments and can decide what work to pay for and what work to reject (i.e., not pay for). Researchers have rejected work for a variety of reasons. Some Researchers have rejected work if a worker misses too many attention check questions (i.e., questions that assess if the worker is paying attention to the survey items), answers a survey too quickly, or if the Requester sees a pattern of satisficing behavior. If a worker believes that he or she is unfairly rejected, Amazon will not mediate the disagreement. The uneven balance of power will be examined in the ethics section of this article.

To summarize, crowdsourcing for academic research is a practice where researchers can connect directly with, and collect data from, a global respondent pool. Data can potentially be collected quickly and relatively inexpensively. The following section provides more detail on the MTurk population.

## Who are the workers at MTurk?

Amazon has never released any data describing the size and demographic composition of the worker pool. Workers are constantly entering and exiting the pool (Huff & Tingley, 2015) and some workers may be active for a time (perhaps during a layoff from another job) and then become inactive (e.g., if other income opportunities arise). As a result, estimating the actual size of the pool is difficult. A 2011 study estimated that the global active worker pool was between 5059 and 42,912 people (Fort, Adda, & Cohen, 2011). Several years later, Kuek et al. (2015) stated that globally there were about half a million registered workers on MTurk, but not all of them were active (defined as completing HITs within the past six months). The web analytics site Alexa.com reported that

MTurk had 750 K unique visitors in December of 2015: this number reflects any global workers, Requesters, and random visitors to the site. Stewart et al. (2015) used a capture–recapture methodology from wildlife ecology to estimate that the average lab is sampling from about 7300 individuals on MTurk.

Many workers are frequent users of the site, with most people indicating that they work at least two days a week on MTurk and two-thirds reporting that they complete HITs every day (Hitlin, 2016). However, the number of hours worked varies greatly: almost a fourth spend less than five hours a week, with less than a fourth working 21 hours per week or more. At the same time, more than half of workers indicated that their hours fluctuate significantly from week to week (Hitlin, 2016).

Researchers may not know the number of workers on MTurk at any given time, but there is evidence that the demographic composition of MTurk workers is much more diverse than that of traditional student samples and can be fairly representative of larger populations. In recent years, student samples have been characterized as "WEIRD" – an acronym for samples drawn from Western, Educated, Industrialized, Rich and Democratic populations (Henrich, Heine, & Norenzayan, 2010). College student participants rarely reflect the population of interest as a whole, yet researchers have for many years generalized results from student samples to larger populations (Henrich et al., 2010). Even if researchers state that their samples should not be generalized to the larger population, a substantial portion of social scientific research has been conducted with student populations and the findings from these studies are often assumed to represent the reality of the broader population. Does utilizing MTurk for data collection avoid the WEIRD? As stated earlier, MTurk workers are more diverse demographically than traditional student samples often used in academic research (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Hitlin, 2016; Huff & Tingley, 2015). MTurk workers, however, are still slightly different from the larger U.S. population. About 80% of the MTurk workforce is based in the U.S.. MTurk workers are also younger than the U.S. population as a whole (about 88% of workers are under 50, compared to 66% of U.S. working adults) and are somewhat better educated (51% of respondents have college degrees, compared to 36% of working adults in the U.S.). In addition, about three-fourths of U.S. workers on MTurk are Caucasian (compared to two thirds of the U.S. population) (Hitlin, 2016). Early studies of MTurk showed a predominance of men in the pool; yet, more recent studies show an almost equal representation of men and women (Chambers, Nimon, & Anthony-McMann, 2016; Hitlin, 2016; Huff & Tingley, 2015). Depending on how the Researcher chooses to limit the pool, respondents can still reflect some of the WEIRD characteristics identified earlier.

Ipeiritos conducts ongoing tracking of MTurk worker demographics at the website MTURKtracker.com; data collected during 2016 reflect similarities to the Pew Center study (Hitlin, 2016), indicating that about 85% of workers are under 50 with a fairly equal distribution of men and women (Ipeirotis, 2010). Workers who are not U.S. based are primarily from India and Iperitos' data show that Indian workers tend to be under 40 and male. About 50% of Indian workers make less than $15,000 per year, compared to only 13% of U.S. Workers. Many researchers limit their HITs to U.S. workers because there have been numerous concerns regarding fraudulent overseas accounts and low-quality data from workers outside the U.S. (Sheehan & Pittman, 2016).

Numerous workers participate on MTurk – at least at the beginning of their tenure – because of monetary motivations (Hitlin, 2016). This is one of the significant differences between an MTurk population and a student population, with students often exhibiting different motivations for participation in academic research (such as to complete course requirements or to earn extra credit or prize opportunities). In contrast, many U.S. workers use MTurk to supplement their incomes, and 25% indicate that all or most of their income comes from MTurk (Hitlin, 2016). Because of the payment aspect, MTurk workers may take their participation in studies more seriously than student respondents. MTurk workers report other motivations for participating on MTurk that are rarely seen in student populations, including personal growth, skill building, and contributing to knowledge and society (Deng & Joshi, 2013).

In general, the workers on MTurk represent are a more diverse sample than typical student samples utilized by academic researchers. Although some MTurk workers are motivated by the income they earn from participation, compensation alone is rarely the sole reason for participation. The next section will address how the data collected from MTurk compare to data from other samples.

## Are data reliable and valid?

Data quality is a concern for all data collected online given that the researcher is physically separated from the individual participating in the study. This lack of monitoring can lead to problems with data quality. Unlike other types of online data collection opportunities, MTurk has incentives in place to help ensure that data quality is high (Goodman & Paolacci, 2017). In particular, Requesters have the ability to reject work and not pay workers, as well as block workers from future work. As a result, workers tend to be motivated to follow instructions and pay attention to the research study, particularly if they know that the study contains questions that serve as attention checks to assess their concentration on the task at hand (Hauser & Schwarz, 2016). Additionally, because Requesters often require workers to have a high approval rate for HITs previously done, workers are even more motivated to not be rejected in order to keep up their high approval rates and have access to a range of future HITs.

Assessing data reliability (the idea that results must be inherently repeatable) and validity (particularly external validity, or the examination of any other possible causal relationship or unknown factors that affect the findings) is important to have theoretically sound data. To determine reliability and validity, several studies have compared MTurk respondents with respondents from existing pools that have been recognized as valid. Berinsky et al. (2012) compared an MTurk sample to that of the online American National Election 2008–2009 Panel Study (ANEPS). Even though it is not perfectly representative of the U.S., ANEPS is generally regarded as a high-quality Internet survey. The same study compared the MTurk sample to two face-to-face probability-based samples: the Current Population Survey and the American National Election Studies (ANES). The MTurk sample demographics skewed slightly more female and slightly less educated than ANEPS, as well as significantly younger. Compared to the face-to-face samples, the MTurk sample was also more likely to be single, to rent rather than own a home, and to hold no religious affiliation. Most importantly, both MTurk and the ANES samples responded similarly to attitudinal questions. In sum, the comparisons led the researchers

to report that MTurk "does not present a wildly distorted view of the U.S. population" (Berinsky et al. 2012, p. 361).

Data validity can also be assessed by examining how samples perform on manipulation and attention checks. Kees, Berry, Burton, and Sheehan (2017) attempted a replication of a published study using five different groups: three online panels (MTurk, Qualtrics and Lightspeed) and two student samples (one group completed the study online, one in a lab). Respondents from MTurk and both student pools performed well on the manipulation check, yet Qualtrics and Lightspeed respondents showed an unacceptable reliability level. Similarly, MTurk and student samples performed acceptably on the attention checks and respondents from Qualtrics and Lightspeed performed less well. MTurk respondents completed the experiment much more quickly than respondents from the other groups, yet provided longer answers to the open-ended questions on the study. The authors concluded the MTurk is as viable platform for academic data collection as other frequently used platforms.

One key demographic difference between MTurk workers who identify as living in the U.S. and the broader U.S. population is important to note. In a range of studies, MTurk respondents tend to skew politically liberal, more so than the general population (Berinsky et al., 2012; Paolacci & Chandler, 2014). Huff and Tingley (2015), in their study comparing MTurk workers to participants in the Cooperative Congressional Election Survey (CCES), a nationally stratified sample survey, found congruence in the two samples with regard to voter registration and intentions to vote. Younger respondents in both studies were similar in their political party identification, and older individuals were more likely to skew liberal on MTurk than on the CCES. Researchers investigating political issues on MTurk have several ways to address this tendency, including weighting the sample, oversampling from states that skew conservative, or screening for political attitudes upfront and ensuring that a balance of perspectives is achieved.

MTurk has also been shown to perform well when compared to snowball samples where existing participants help recruit other participants. For instance, a study of the demographics of samples of parents collected from MTurk, Facebook, and parent-oriented listservs showed that both MTurk and Facebook were better at recruiting socioeconomically diverse parents than listservs (Dworkin, Hessel, Gliske, & Rudi, 2016). In addition, the demographics of the U.S. MTurk sample aligned with the population of the U.S. with regard to gender, race/ethnicity, and marital status more closely than samples from Facebook or parental listservs. Casler, Bickel, and Hackett (2013) found consistency in responses from snowball and MTurk samples, concluding that, "for some behavioral tests, online recruitment and testing can be a valid – and sometimes even superior – method to in-person data collection" (p. 2159).

In addition to examining consistency in samples, it is important to examine the extent to which samples replicate existing findings in the literature. Replication studies show that the MTurk samples replicate existing studies. For example, in two different studies of risk framing, Berinsky et al. (2012) used MTurk to replicate earlier studies showing that people respond differently to gain frames and loss frames of the same situation. In particular, a loss frame resulted in higher selection of a riskier policy choice. Simons, Chabris, and de Fockert (2012) also replicated findings from an earlier random digit dial telephone sample study about American's mistaken beliefs about how memory works. Moreover, Horton, Rand, and Zeckhauser (2011) replicated a lab-based priming experiment that

indicated, like the lab respondents, MTurk workers responded to priming by altering behaviors. Other studies have also demonstrated test–retest and internal validity of MTurk samples on a variety of measures, including the five factor model of trait personality (Holden, Dennie, & Hicks, 2013), body size estimation and body dissatisfaction (Gardner, Brown, & Boice, 2012), Machiavellianism and narcissism, (Jonason & Luévano, 2013; Jones & Paulhus, 2014), the Positive Affect Negative Affect Schedule, the Happiness-Depression Scale (Schütz et al., 2013), and general mental health (Schleider & Weisz, 2015).

While the evidence for validity based on replications is robust, researcher warn of some findings that might effect results in some situations. For example, workers on MTurk have higher needs for cognition (Berinsky et al., 2012), are more introverted (Goodman, Cryder, & Cheema, 2013), and may have slightly lower self-esteem than the general population (Arditte, Çek, Shaw, & Timpano, 2016). Goodman and Paolacci attribute these differences to the fact that MTurk workers tend to be individuals who enjoy doing solitary tasks on the Internet. Finally, a comparison of data from an MTurk sample and the General Social Survey (GSS) sample (a face-to-face probability survey) also found that MTurk respondents displayed higher scientific knowledge than GSS respondents, even after accounting for demographic differences (Cooper & Farid, 2016). The researchers suggested that this was not specifically an MTurk issue, but rather an issue with collecting data from anyone who is highly engaged with the Internet:

> It would be reasonable to assume that the difference in knowledge reported here is not isolated to only science-related knowledge and that the (MTurk) workers may be generally more literate and knowledgeable than the average population. Any studies asking new questions about population knowledge or attitudes using (MTurk) should take care to consider this potential difference. A good strategy might be to include some related standard survey questions in order to assess this knowledge difference. (p. 8)

Ford (2017) questioned how researchers could be sure that MTurk workers are representing themselves honestly – a question that applies to almost any type of online data collection. Although most studies suggest that there is minimal cheating on MTurk, the online dis-inhibition may create what could be called the "faithless respondent" – that is, respondents who are untrustworthy or deceitful when self-selecting to participate in HITs where they lie about their eligibility (such as a 50-year-old indicating that he or she is under 30 in order to participate in a well-paying or somewhat simple survey) (Springer, Martini, Lindsey, & Vezich, 2016). The dis-inhibition effect is the lowering of psychological restraints that serve to govern online behavior (Lapidot-Lefler & Barak, 2015) and when combined with anonymity, dis-inhibition may encourage persons to provide inaccurate personal information. Dis-inhibition and anonymity may factor into any type of digital data collection, although the payment aspect of MTurk work may serve to heighten the problem and researchers are concerned that individuals from outside of the U.S. present themselves as citizens in order to earn payments for HITs (Ford, 2017). Chandler, Paolacci, Peer, Mueller, and Ratliff (2015) argue that the number of people who are faithless respondents on MTurk is small, although their presence may threaten the validity of a study if those who truly possess the characteristics of interest differ significantly from those who do not. Wessling, Huber, and Netzer (2017) disagreed, identifying an imposter rate of 24–83% across a series of MTurk studies.

Siegel, Navarro, and Thomson (2015) found that overtly stating eligibility requirements can result in data that are different from data collected when eligibility requirements are not overtly stated and workers are screened out, suggesting that faithless respondents are indeed present in the research process on MTurk. Faithless respondents, however, can appear in any online research study where the researcher is not present when collecting the data.

Data collected on MTurk can be just as valid as data collected via alternative methods, and a growing body of literature provides support for this sentiment. Nevertheless, it is important that specific practices for data collection are followed in order to help prevent problems with data collection identified above. These are best practices discussed in the next section.

## What are best practices for using MTurk for research?

One concern expressed about MTurk workers is that they are "non-naïve" or experienced survey takers (Chandler et al., 2015). Some workers complete dozens of HITs every day. As a result, workers could be less thoughtful when completing a survey, particularly when answering items on a scale that they have answered several times before (Chandler et al., 2015), such as on standard uses and gratifications scales. Utilizing unique stimuli whenever possible in research can help alleviate this problem. Researchers conducting multiple studies around a similar topic that are not longitudinal studies might also avoid recruiting participants who have completed their earlier studies. Researchers can either manually assign a qualification on MTurk before data collection (Litman et al., 2015) or exclude responses from workers answering previous studies (by matching their IDs to the IDs collected on previous studies) after data collection (Goodman & Paolacci, 2017).

To date, my colleagues and I have collected data using MTurk for a variety of studies about communication, with topics including binge watching, environmental messaging, and corporate social responsibility. The best practices recommended below come not only from the literature on research methodologies, but also from our own experiences using the site.

- Check your institution's Human Subjects Guidelines. Prior to posting their first HIT, researchers should check their own institution's IRB guidelines for any specific guidance about using MTurk. As the use of the service increases in popularity, IRBs are starting to update guidelines for informed consent and payment.
- Experience MTurk as a worker first. Researchers should join MTurk as a worker and complete a few academic HITs before collecting data themselves. Signup requires providing your name, email address, a password, and country of residence. In addition, you will be asked to sign the MTurk Participation Agreement. You will be notified via email when you are registered. Participating in a few academic HITs will allow you to experience the service through the eyes of a worker, to see the types of information that researchers use to describe their HITs to attract workers, and to examine HIT instructions to see if language is clear, the framing of the HIT is persuasive and not misleading, and if the wage seems fair. Workers value clear and succinct writing (Sheehan & Pittman, 2016) and examining others' HITs is a good way to learn how to position your HIT. Additionally, Goodman and Paolacci (2017) recommend you participate

as a worker in one of the several forums run by current workers (see Appendix) to learn how these sites talk about HITs and the researchers that post them.

- Know the rules. Before you set up your HIT, examine the types of HITs that are not allowed by MTurk (see Appendix) or that compromise respondent privacy: Amazon's terms of service for Requesters outline one set of responsibilities, but certainly not all. Never require workers to provide personally identifying information (email addresses, birthdates, real names) to complete HITs. Never require workers to access their Facebook accounts, which are also considered a personally identifiable site by Workers. Goodman and Paolacci (2017) recommend accessing the "We Are Dynamo" wiki (see Appendix) for additional recommendations for researchers from workers themselves, such as providing your full name and institutional affiliation and indicating a reasonable time estimate for completion of work.
- Design your study to make sure that you get qualified respondents. Researchers have two ways to address this issue. First, they can request (and pay for) specific characteristics for respondents as discussed earlier in this manuscript. Alternatively, researchers can implement a short screening survey that does not identify the characteristic of interest in the description of the MTurk HIT (Wessling et al., 2017). For instance, Springer et al. (2016) wanted to study Muslims and used a screening question about religion rather than stating that the research purpose was to study Muslims in order to minimize the likelihood of selection bias based on the use of enticing (or unenticing) language in the description of the task. Workers who do not answer correctly on the screening test are then "screened out" and appropriate workers are able to continue to the survey or experiment.
- Implement attention checks. Use one or more types of attention check questions to ensure first, that respondents are real people and not robots automatically completing surveys and second, that respondents are paying attention. These include reverse worded questions or statements, trap questions (such as having one statement in a matrix of statements stating "click 'always agree'"), and attention filter questions (where a large block of text describes the study, and often ends with directions to (for example) click on all sports you enjoy watching on television. This is followed by a long list of sports; the center of the large block of text contains directions telling the respondent to either click on one specific sport only, or to click on the box labeled "other" and write in a phrase such as "I'm paying attention"). If you utilize attention check questions, make this clear in the description of the HIT and in the informed consent form. Also indicate that you will withhold payment of people who do not pass attention checks. This could discourage workers who speed through studies and cheaters from participating in your HIT.
- Optimize the survey to slow down respondents. Experienced workers answer surveys much more quickly than students, creating concerns about possible satisficing behavior (Smith, Roster, Golden, & Albaum, 2016). To minimize respondents speeding through the study, consider setting up your survey to "freeze" the survey page for a specific period of time (e.g., the respondent will not be able to click to the next page for 30 seconds). This is particularly useful to make sure that respondents are exposed to a stimulus for a specific period of time, and to generate good responses to open-ended questions. Researchers can also include open-ended items that require a written response to ensure that the workers are paying attention.

- Pretest the survey to examine functionality and expected completion times. Pretests allow for confirmation that all parts of a survey are working as they should. They also provide some benchmark information on how long it takes a person to complete the survey. This information can be used to assess whether an MTurk worker is speeding through the survey too quickly; combined with information on attention checks, this can help the researcher decide whether to include the worker's responses or not. Keep in mind, though, that Kees et al. (2017) found that MTurk workers completed a study in about two-thirds of the time it took for student and Qualtrics samples and in about half of the time for the Lightspeed sample. At the same time, MTurk workers did better on both attention and manipulation checks. This suggests that MTurk workers might be more used to completing surveys than other pools.
- Limit the amount of time a worker has to complete your study. Amazon allows the researcher to indicate how much time a worker has to complete the study once he or she accepts the HIT. A good rule of thumb is to allot two to three times the length of time it should take to complete the survey for the total time allotment. This allows for variation among workers and ensures that workers complete the HIT at one sitting. HITs with longer time allotments encourage workers to "accept" a HIT and then not complete it right away as they shop around for HITs with shorter time allotments to complete. Although it is unclear how this practice might affect data quality, minimizing the time allotment seems like a beneficial practice.
- If contacted, respond to and engage with workers. Once your HIT is available to workers, it is possible that workers will contact you, particularly if something is missing (such as a completion code), if instructions are unclear, or if they experienced an issue with the instrument that might cause their work to be rejected (e.g., if the instructions mentioned an attention check question but the worker did not see the question and was concerned it was not included in the final instrument). Answering workers promptly will build trust between you and the workers, and will start to address some other power issues inherent in the site, which are discussed in the upcoming section.

## What are the ethical issues surrounding MTurk?

The purpose of academic research is to create and disseminate new knowledge, and in order for research to be feasible and meaningful, the relationship between researcher and respondent is crucial. Institutional processes help to ensure that respondents understand their rights in academic studies, have the ability to provide informed consent, and know whether their responses will be anonymous and/or confidential. All these mechanisms create trust between the researcher and the respondent. Because MTurk is essentially an unregulated workforce, however, the researcher–respondent relationship also becomes of one of employer–employee. The researcher is arguably hiring an MTurk worker to complete a task within a prescribed period of time in exchange for payment. This can create a power imbalance. Workers have few legal protections "as the cyberspace in which they work remains essentially unregulated for employment and labor law purposes" (Schmidt, 2013, p. 532).

Institutional Review Board concerns generally do not include low payments. Most IRBs instead worry about too high of a payment that might coerce someone to participate or

continue in a study that they do not want to do (Sheehan & Pittman, 2016). Thus, one source of ethical tension comes with the payment aspect of MTurk: how much should respondents be paid? Cushing (2013) referred to sites like MTurk as digital sweatshops where tensions arise when sites operate simultaneously as a utility (to provide a service) and a business (to keep costs down and profits up). Cash-strapped academic researchers become attracted to the "fast and cheap" association of MTurk data collection online. Researchers rationalize payment decisions by suggesting that paying low wages to many people is the only way to establish the statistical power necessary to create publishable research (Sheehan & Pittman, 2016). They also argue that lower payments reflect lower respondent opportunity costs (such as travel costs) for MTurk workers than traditional lab respondents (Mason & Suri, 2012). These rationales, however, can diminish trust between researchers and respondents because respondents feel they are exploited when they are paid rates far below a living wage (Sheehan & Pittman, 2016). This can lead workers to feel that they are a part of a precariat or a group of insecure workers who teeter on the edge of survival (Dobson, 2013). Some MTurk workers will also voice their concerns if the researcher is not paying them enough for their service. Researchers should consider how much they are paying the MTurk workers and pay them a fair amount for the time they are spending on the study. They should not necessarily look at what MTurk workers are paid for other work, but what is considered a fair and ethical amount for someone in the U.S. who is completing their study. Paying about 15 cents per minute (based on pretesting times) is likely to generate good-quality responses in a fairly short period of time. You may wish to consider the minimum wage in your state as well in order to calculate an appropriate cost. Researchers should also include information in their manuscript about what they paid their participants. One way to ensure that the research community recognizes the importance of treating MTurk workers fairly is to raise awareness of reasonable practices.

Payment issues are just the tip of the "power dynamic iceberg". Researchers have the ability to charge what they wish: Amazon will not set payment floors. Researchers can also to refuse payment to any worker and can block any worker from participating in their studies if they believe that the worker is providing substandard work. Given their hands-off policy, Amazon will not address any type of worker complaints, and will suspend workers who are researchers frequently block. Workers are not even protected by anonymity. If a worker ever left a product review or even assigned a "star" rating to a product anywhere in the Amazon universe, their user information could be linked to their MTurk account (Lease et al., 2013). This is particularly problematic if an MTurk worker uses his or her real name to rate products on Amazon because researchers could learn an individual's gender and possibly other information. Indicating that responses may not be anonymous on the consent agreement allows workers to choose whether they wish to continue with the HIT or not.

Spending some time and energy considering the payment is important because it effects the reputation of a researcher on MTurk and also has important ethical implications for the researcher and the larger research community. The unbalanced power dynamics on MTurk have created several different online communities (for examples, see Appendix) where workers discuss individual HITs, including whether the task was well defined and straightforward to accomplish, whether the wage is fair given the task requirements,

how well the researcher communicated with workers, and how quickly workers are paid. Amazon itself has not provided a rating system for workers to rate MTurk Requesters.

One independent site, the Turkopticon, is dedicated solely to worker ratings of Requesters and their HITs, and includes numeric ratings as well as worker comments (similar to Amazon's service where purchasers can rate sellers and products). The Turkopticon allows workers to rate Requesters on four attributes (Irani & Silberman, 2013):

- Communication: the responsiveness of the requester to emails expressing concerns. Workers can email Researchers with questions or comments directly from MTurk if they are having a problem with the HIT itself.
- Generosity: how well the HIT is paid for the amount of time it takes to complete.
- Fairness: the degree to which the Requester is fair in approving or rejecting work and
- Promptness: how quickly work was approved and paid.

Each attribute is rated on a scale from 1 to 5, with 1 being low and 5 being high. Workers are allowed to write comments to provide context for their ratings, and many workers regularly check the Turkopticon before agreeing to do a HIT.

Researchers can positively influence their reputations by being transparent on MTurk. For example, they can use their real names and indicate their institutional affiliation in the HIT instructions. Providing informed consent is essential as well, given that IRB oversight is one of the few means of recourse workers have if researchers are not acting ethically. In the past, workers have contacted IRBs in situations where there is Requester misrepresentation, or when tasks do not provide informed consent (Sheehan & Pittman, 2016).

This section outlined a number of best practices for conducting research on MTurk, but it is important to note that as more researchers utilize the platform, more issues arise. For example, does working for several hours at a time on MTurk result in fatigue and decreased attention, influencing how people respond? Does requiring Workers to have completed a certain number of studies, or have a high HIT acceptance rate, influence validity? Does paying extra to Amazon for certain qualifications improve data validity or merely line Amazon's pockets?

In addition, MTurk is not the only crowdsourcing platform that researchers can access: other options include Prolific Academic and Crowdflower. These services are somewhat comparable to MTurk, in that they can be used for data collection but have much smaller respondent pools. Peer, Brandimarti, Samat, and Acquisti (2017) found that participants on Prolific Academic and Crowdflower were less experienced, less dishonest, and more diverse compared to MTurk participants in two different studies, but Crowdflower participants failed more attention checks. As collecting data via crowdsourcing continues to attract new researchers, further examination of different platforms will be warranted.

## Conclusion

In the age of ubiquitous technology, it is difficult for researchers to conduct a truly random sample for research purposes. Researchers have increasingly relied on different types of online samples to collect data, and all such samples suffer from issues of self-selection. This essay has examined the benefits and challenges of using Amazon's Mechanical Turk online workforce for data collection. Data can be collected quickly and at a lower

cost than other online resources, and respondents are demographically more diverse than traditional student samples. A range of studies show that the samples are rather congruent with other samples in terms of demographics and responses to established surveys. In order to ensure that data are valid and reliable, researchers should take steps to avoid the "faithless respondent" who provides false information to be able to participate in studies, utilize attention checks to encourage attentive responses, and understand the unique responsibilities of being both a researcher and an employer when using MTurk.

## References


Abshire, M., Dinglas, V. D., Cajita, M. I. A., Eakin, M. N., Needham, D. M., & Himmelfarb, C. D. (2017). Participant retention practices in longitudinal clinical research studies with high retention rates. *BMC Medical Research Methodology*, *17*(1), 458. doi:10.1186/s12874-017-0310-z

Alonso, O., & Lease, M. (2011). *Crowdsourcing for information retrieval: Principles, methods, and applications*. Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval, pp. 1299–1300. ACM. doi:10.1145/2009916.2010170

Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in mechanical Turk research. *Psychological Assessment*, *28*(6), 684–691.

Barr, J., & Cabrera, L. F. (2006). AI gets a brain. *Queue*, *4*(4), 24. doi:10.1145/1142055.1142067

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

Brabham, D. C. (2010). Moving the crowd at threadless: Motivations for participation in a crowdsourcing application. *Information, Communication & Society*, *13*(8), 1122–1145. doi:10.1080/13691181003624090

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi:10.1037/e527772014-223

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156–2160. doi:10.1016/j.chb.2013.05.009

Chambers, S., Nimon, K., & Anthony-McMann, P. (2016). A primer for conducting survey research using MTurk: Tips for the field. *International Journal of Adult Vocational Education and Technology (IJAVET)*, *7*(2), 54–73. doi:10.4018/IJAVET.2016040105

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, *26*(7), 1131–1139. doi:10.3758/s13428-013-0365-7

Cooper, E. A., & Farid, H. (2016). Does the sun revolve around the earth? A comparison between the general public and online survey respondents in basic scientific knowledge. *Public Understanding of Science*, *25*(2), 146–153. doi:10.1177/0963662514554354

Cushing, E. (2013). Amazon mechanical Turk: The digital sweatshop. *Utne Reader*. Retrieved from http://www.utne.com/science-and-technology/amazon-mechanical-turk-zm0z13jfzlin

Deng, X., & Joshi, K. D. (2013). *Understanding crowd workers' perceptions of crowdsourcing career*. Thirty Fourth International Conference on Information Systems, Milan. Retrieved from https://pdfs.semanticscholar.org/73ef/ab88621309fdf3d39ac2aff8c70b193c0606.pdf

Dobson, J. (2013). Mechanical Turk: Amazon's new underclass. *Huffington Post*, p. 19. Retrieved from http://www.huffingtonpost.com/julian-dobson/mechanical-turk-amazons-underclass_b_2687431.html

Dworkin, J., Hessel, H., Gliske, K., & Rudi, J. H. (2016). A comparison of three online recruitment strategies for engaging parents. *Family Relations*, *65*(4), 550–561. doi:10.1111/fare.12206

Ford, J. B. (2017). Amazon's mechanical Turk: A comment. *Journal of Advertising*, *46*(1), 156–158. doi:10.1080/00913367.2016.1277380

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, *37*(2), 413–420. doi:10.1162/coli_a_00057

Gardner, R. M., Brown, D. L., & Boice, R. (2012). Using Amazon's mechanical Turk website to measure accuracy of body size estimation and body dissatisfaction. *Body Image*, *9*(4), 532–534. doi:10.1016/j.bodyim.2012.06.006

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. doi:10.1002/bdm.1753

Goodman, J., & Paolacci, G. (2017). *Crowdsumers take over*. Retrieved from https://u.osu.edu/goodman/files/2016/12/JCR-MTurk-Tutorial-1dnc063.pdf

Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *BehaviorRresearch Methods*, *48*(1), 400–407. doi:10.3758/s13428-015-0578-z

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. doi:10.1017/s0140525x0999152x

Hitlin, P. (2016). *Research in the crowdsourcing age, a case study*. Retrieved from http://assets.pewresearch.org/wpcontent/uploads/sites/14/2016/07/PI_2016.07.11_Mechanical-Turk_FINAL.pdf

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's mechanical Turk. *Computers in Human Behavior*, *29*(4), 1749–1754. doi:10.1016/j.chb.2013.02.020

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425. doi:10.1007/s10683-011-9273-9

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6), 1–4.

Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, *2*(3), 205316801560464. doi:2053168015604648

Ipeirotis, P. G. (2010). Analyzing the Amazon mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, *17*(2), 16–21. doi:10.1145/1869086.1869094

Irani, L. C., & Silberman, M. (2013). *Turkopticon: Interrupting worker invisibility in Amazon mechanical Turk*. Proceedings of the SIGCHI conference on human factors in computing systems, pp. 611–620, ACM. doi:10.1145/2470654.2470742

Jonason, P. K., & Luévano, V. X. (2013). Walking the thin line between efficiency and accuracy: Validity and structural properties of the dirty dozen. *Personality and Individual Differences*, *55*(1), 76–81. doi:10.1016/j.paid.2013.02.010

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment*, *21*(1), 28–41. doi:10.1177/1073191113514105

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's mechanical Turk. *Journal of Advertising*, *46*(1), 141–155. doi:10.1080/00913367.2016.1269304

Kuek, S. C., Paradi-Guilford, C., Fayomi, T., Imaizumi, S., Ipeirotis, P., Pina, P., & Singh, M. (2015). *The global opportunity in online outsourcing* (No. 22284). The World Bank. Retrieved from http://documents.worldbank.org/curated/en/138371468000900555/The-global-opportunity-in-online-outsourcing

Lapidot-Lefler, N., & Barak, A. (2015). The benign online disinhibition effect: Could situational factors induce self-disclosure and prosocial behaviors? *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *9*(2), article 3. doi:10.5817/CP2015-2-3

Lease, M., Hullman, J., Bigham, J. P., Bernstein, M. S., Kim, J., Lasecki, W., Bakhashi, S., Mitra, T., & Miller, R. C. (2013). Mechanical Turk is not anonymous. *SSRN*. doi:/10.2139/ssrn.2228728

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on mechanical Turk. *Behavior Research Methods*, *47*(2), 519–528.

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

Morton, E. (2015). *The mechanical chess player that unsettled the world*. Retrieved from http://www.slate.com/blogs/atlas_obscura/2015/08/20/the_turk_an_supposed_chess_playing_robot_was_a_hoax_that_started_an_early.html

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184–188. doi:10.1177/0963721414531598

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. doi:10.1016/j.jesp.2017.01.006

Peer, E., Paolacci, G., Chandler, J., & Mueller, P. (2012). *Screening participants from previous studies on Amazon mechanical turk and qualtrics* (Unpublished manuscript).

Schleider, J. L., & Weisz, J. R. (2015). Using mechanical Turk to study family processes and youth mental health: A test of feasibility. *Journal of Child and Family Studies*, *24*(11), 3235–3246. doi:10.1007/s10826-015-0126-6

Schmidt, F. A. (2013, September). *The good, the bad and the ugly: Why crowdsourcing needs ethics*. Cloud and Green Computing (CGC), 2013 Third International Conference, pp. 131–535. doi:10.1109/cgc.2013.89

Schütz, E., Sailer, U., Al Nima, A., Rosenberg, P., Arntén, A. C. A., Archer, T., & Garcia, D. (2013). The affective profiles in the USA: Happiness, depression, life satisfaction, and happiness-increasing strategies. *PeerJ*, *1*, e156. doi:10.7717/peerj.156

Sheehan, K. B., & Pittman, M. (2016). *Amazon's mechanical Turk for academics: The HIT handbook for social science research*. Irvine, CA: Melvin & Leigh.

Siegel, J. T., Navarro, M. A., & Thomson, A. L. (2015). The impact of overtly listing eligibility requirements on MTurk: An investigation involving organ donation, recruitment scripts, and feelings of elevation. *Social Science & Medicine*, *142*, 256–260. doi:10.1016/j.socscimed.2015.08.020

Simons, D. J., Chabris, C. F., & de Fockert J. (2012). Common (mis) beliefs about memory: A replication and comparison of telephone and mechanical Turk survey methods. *PloS One*, *7*(12), e51876.

Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, *69*(8), 3139–3148. doi:10.1016/j.jbusres.2015.12.002

Springer, V. A., Martini, P. J., Lindsey, S. C., & Vezich, I. S. (2016). Practice-based considerations for using multi-stage survey design to reach special populations on Amazon's mechanical Turk. *Survey Practice*, *9*(5). Retrieved from http://www.surveypractice.org/index.php/SurveyPractice/article/view/305/html_75

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

Wessling, K. S., Huber, J., & Netzer, O. (2017). MTurk Character Misrepresentation: Assessment and Solutions. *Journal of Consumer Research*, *44*(1), 211–230. doi:10.1093/jcr/ucx053

# Appendix

## *Resources to learn more about MTurk*

- How MTurk Works: http://www.pewinternet.org/2016/07/11/what-is-mechanical-turk/).
- Resources for global researchers
  - Turk Prime: a resource that allows global researchers to utilize the service he enhances researchers' use of MTurk by providing mechanisms for longitudinal studies. https://www.turkprime.com/
  - MTurk Data Consultants: http://mturkdata.com/index.html
- MTurk Fees: https://requester.mturk.com/pricing
- HITS that are not allowed on MTurk: https://requester.mturk.com/help/faq#restrictions_use_mturk
- MTurk Worker Communities
  - Turker Nation http://turkernation.com/
  - MTurk Crowd: http://www.mturkcrowd.com/
  - MTurk forum: http://www.mturkforum.com/
  - mTurk Grind forum: http://www.mturkgrind.com/forum.php
  - WeAreDynamo http://www.wearedynamo.org/