Risks and Rewards of Crowdsourcing Marketplaces

Jesse Chandler

University of Michigan/PRIME Research

Gabriele Paolacci

Erasmus University Rotterdam

Pam Mueller

Princeton University

<5298 Words>

This manuscript appears as a chapter of the Handbook of Human Computation (Springer). Please cite:
Chandler, J., Paolacci, G., & Mueller, P. (2013). Risks and rewards of crowdsourcing marketplaces. In *Handbook of Human Computation* (pp. 377-392). Springer New York.

Abstract

Crowdsourcing has become an increasingly popular means of flexibly deploying large amounts of human computational power. The present chapter investigates the role of microtask labor marketplaces in managing human and hybrid human machine computing. Labor marketplaces offer many advantages that in combination allow human intelligence to be allocated across projects rapidly and efficiently and information to be transmitted effectively between market participants. Human computation comes with a set of challenges that are distinct from machine computation, including increased unsystematic error (e.g. mistakes) and systematic error (e.g. cognitive biases), both of which can be exacerbated when motivation is low, incentives are misaligned, and task requirements are poorly communicated. We provide specific guidance to how to ameliorate these issues through task design, workforce selection, data cleaning and aggregation.

Risks and Rewards of Crowdsourcing Marketplaces

The present chapter focuses on the risks and rewards of using online marketplaces to enable crowdsourced human computation. We discuss the strengths and limitations of these marketplaces, with a particular emphasis on the quality of crowdsourced data collected from Amazon Mechanical Turk. Data quality is by far the most important consideration when designing computational tasks, and it can be influenced by many factors. We emphasize Mechanical Turk because it is currently one of the most popular and accessible crowdsourcing platforms and offers low barriers of entry to researchers interested in exploring the uses of crowdsourcing. In addition to describing the strengths and limitations of this platform, we provide general considerations and specific recommendations for measuring and improving data quality that are applicable across crowdsourcing markets.

Crowdsourcing is the distribution of tasks to a large group of individuals via a flexible open call, in which individuals work at their own pace until the task is completed (for a more detailed definition see Estelles-Arolas & Gonzalez-Ladron-de Guevera, 2012). Crowd membership is fluid, with low barriers to entry and no minimum commitment. Individuals with heterogeneous skills, motivation, and other resources contribute to tasks in parallel. Crowdsourcing leverages the unique knowledge of individual crowd members, the sheer volume of their collective time and abilities, or both to solve problems that are difficult to solve using computers, or smaller and more structured groups.

The unique strengths of groups are generally used to solve one of two basic kinds of problems. Some problems have no obvious a priori solution, but correct answers seem obvious once known (e.g. insight problems; Dominowski & Dallob, 1995) or can be verified. In these

cases, crowds can generate responses from which the "best" response can be selected according to some criteria. The volume and diversity of workers with different perspectives, strategies and knowledge can lead to quick, unorthodox, and successful solutions. The Internet has furthered this approach to problem solving by creating virtual meeting places where people can post problems for others to solve. For example, Innocentive (Allio, 2004) is a website that has helped companies find solutions to technical challenges like preventing oxygen from passing through rubber, or adding fluoride powder to toothpaste without dispersing it into the air. Often solutions to these specialized, technical problems are provided by amateurs, hobbyists, or experts in apparently unrelated fields (Larkhani, 2008).

Tasks that require resources beyond those available to a single individual or work group are also well-suited to crowdsourcing. The compilation of the Oxford English Dictionary is one early example of this approach. A unique feature of this dictionary is that it includes not only definitions, but also published examples of word use. Examples were collected on slips of paper by a large body of volunteers and then aggregated by editors (Winchester, 2004). Advances in machine computation have made it easier to manage projects of this scale. For example, The Open Science Collaboration coordinates the real time collaborative efforts of scientists and citizen-scientists to systematically code, replicate and communicate social scientific findings using freely available web-software (Open Science Collaboration, 2013).

A subset of this broad category are tasks that are easy for people to solve, but difficult for machines to solve. These assignments are particularly amenable to crowdsourcing. In many cases, a crowd's responses can be automatically aggregated, eliminating the need to comprehensively review responses. The volume of workers performing each task can allow ideosyncratic perspectives, strategies and knowledge to be homogenized removed through

aggregation, leaving consistent performance across a task even though each individual completed only a small portion of it. Consequently, advancing machine computation has increased the applications of crowdsourcing, with the development of human-machine hybrid systems that tackle ambitious projects such as describing the contents of images in near real time (e.g., VizWiz; Bigam et al., 2010), classifying millions galaxies (Galaxy Zoo; Lintott et al., 2008), or determining the shapes that proteins fold into (Foldit; Cooper et al., 2010). Each of these projects emerged as a result of the uneven ability of machine computation to handle the various necessary task elements.

While some platforms for marshaling crowds have been developed to solve specific large problems, "crowdsourcing marketplaces" have also emerged to match workers and requesters with more modest needs. The most prominent example is Mechanical Turk (MTurk), a crowdsourcing website launched by Amazon in 2005 to assist with the maintenance of its own websites (e.g. identifying duplicate products; Potin, 2007). Corporations and individuals alike use crowds recruited from MTurk to conduct human computation operations. Twitter, for instance, relies on MTurk workers to categorize search queries to make them more meaningful to other users. Machine computing can easily identify a spike in the popularity of a query (e.g., "Big Bird" in Fall 2012), but not its semantic properties. Trending queries are passed on to MTurk workers, who can easily determine that this is a result of political events (Mitt Romney's comments in the US Presidential Debate) rather than *Sesame Street.*

Scientists have also been quick to harness crowd computing for academic research, relying on crowds to complete a variety of time-consuming tasks including generating corpora of stimuli for machine learning experiments (Lane, Waibel, Eck, & Rottmann, 2010; Lau, Drew, & Nichols, 2009); rating and classifying words according to meaning (e.g., Li, Liu, & Agichtein,

2008); transcribing speech (Gruenstein, McGraw & Sutherland, 2009; Marge et al., 2010); proofreading text for errors (Tetreault, Filatova & Chodorow 2010); verifying citations (Molla & Santiago-Martinez, 2011) and coding observational data (e.g. Hsieh, Kraut, & Hudson, 2010). Others are experimenting with building more complex workflows, where workers collaborate on complex multi-stage projects, or in which workers are treated as agents with a plurality of diverse responses, rather than a means of measuring the average beliefs of a population (Nickerson, Sakomoto & Yu, 2011; Yu & Nickerson, 2011)

Strengths of Crowdsourcing Marketplaces

Transaction Cost Effectiveness. The major advantage of marketplaces is that they make crowdsourcing accessible to requesters with limited financial and technical resources. The fixed costs of crowdsourcing (servers, record keeping, technical support, etc.) can be shared by many requesters and the technical challenges can be handled by dedicated specialists. Other less tangible efficiencies are also realized through sharing a common platform. Workers only need to be recruited into the market once, reducing marketing costs. Moreover, they only need to learn how to use a single standardized interface and can share their experiences with others, making it easier for them to find, understand, and successfully complete work (Ipeirotis & Horton, 2011).

Crowd Accessibility. Crowds require a certain critical mass to function. Potential workers are unlikely to invest time visiting websites unless they have a reasonable chance of finding work (a special case of a two-sided market, see Rochet & Triole, 2003). Some crowdsourcing projects, like digitizing every book in the world, or identifying all the stars in the sky, are large enough to warrant their own dedicated framework (e.g., reCaptcha; von Ahn et al., 2008). However, the majority of human computation problems are quick to complete, intermittent, or frequently change in content or required knowledge. A common market ensures a

steady enough supply of tasks to help maintain a persistent crowd, even while individual requesters recruit and dismiss workers on demand. MTurk was able to achieve this scale initially by serving as a labor market for Amazon's own in-house human computation needs.

Efficient Matching and Task Completion. Microtask sites pay workers according to the tasks they complete, rather than an hourly wage. Piece rates ensure that workers are paid according to their productivity, and even assuming minimal variation in worker ability and task demands, workers should be able to sort themselves into assignments they do best (Becker & Murphy, 1992). Piece rates also benefit requesters. Since each worker proceeds at their own pace, receiving new work only when old work is completed, the completion time for a project will be driven by the average pace at which tasks are completed, as opposed to traditional methods of dividing labor that are often constrained by the pace of the slowest worker (Davis, 1965).

Low Market Prices. Aside from a minimal payment to the web service (MTurk charges 10% of worker payments to cover overhead and financial transaction fees), the only cost faced by requesters to crowdsource their tasks is worker compensation. In 2010, Horton and Chilton estimated the median reservation wage of MTurk workers to be less than $2 per hour, i.e., less than 20% of the wage of the average general secretary in the United States (Bureau of Labor Statistics, 2010). Current rates are likely higher, but even a rate of $6 per hour is sufficient for a task to be posted to one of the the various forums where workers share well-paying HITS (e.g. http://www.reddit.com/r/HITsWorthTurkingFor, www.turkernation.com).

There are a number of reasons that workers within certain crowds accept wages which traditional workers would not: they can select tasks that are relatively interesting or meaningful (Kauffman, Schulze & Veit, 2011), they can work from any location, and they can use time that

has little other economic value (e.g., completing work between or even in parallel with other tasks). MTurk also allows requesters to use workers from regions or countries with lower costs of living and lower minimum wages. However, we should also note that the US workers are often comprised of people with limited traditional sources of income (Shapiro, Chandler & Mueller, 2013) and that researchers may want to consider the ethical implications of the wages they offer workers when making payment decisions (for discussions see Horton, 2011; Kittur et al., 2013; Silberman, Irani & Ross, 2008).

Trust and Reputation Transparency. Exchanging goods or labor requires a certain amount of trust. In offline communities, reputational information is spread informally through a community. Online, requesters and workers must interact anonymously with each other, making them vulnerable to fraud or exploitation. The division of work into smaller tasks paid as piecework prevents the need to engage in long-term commitments between workers and recruiters. Workers can try working with a requester once with minimal risk and increase their commitment if the first transaction proceeds smoothly.

Centralizing work within an online marketplace makes it possible to share information about potential exchange partners so participants can identify and avoid or sanction untrustworthy partners, even when they are effectively anonymous (Resnick, Kuwabara, Zeckhauser & Friedman, 2000). MTurk, for example, tracks the proportion of tasks that workers successfully complete, and requesters can use this information as a recruitment criterion. Particularly unscrupulous workers can be blocked by individual requesters, and multiple blocks can result in workers being banned from the marketplace. Similarly, workers maintain ratings of requesters (e.g. www.turkopticon.com) that can guide other workers' decisions about who they work for.

Data Quality. The low cost of labor, combined with the conventional wisdom that "you get what you pay for," can lead to skepticism about the true value of work performed by crowds of strangers working for below minimum wage. Empirical examinations have found that data quality is not something that can be solved through wages: poorly paid crowds produce data of the nearly the same quality as well paid crowds (albeit slowly; Buhrmester, Kwang & Gosling, 2011; Mason & Watts, 2009), community volunteers (Goodman et al., 2012), or undergraduate students (Paolacci et al., 2010). There are forces that ensure quality even when payment is low: many tasks that are difficult for machines are trivially easy for people to do, and for more difficult tasks, reputational concerns may dissuade workers from submitting poor quality work. Further, since most crowdsourcing tasks recruit workers using an open call, high wages attract more workers of all skill levels to the task equally. Consequently, features of task design, instruction clarity and worker selection drive work quality in crowds, just as they do in more traditional workplaces.

Recruitment Flexibility. Crowdsourcing marketplaces allow requesters to specify that workers possess certain attributes in order to complete a task. Worker recruitment on MTurk can be restricted to residents of a specific country, or to workers who have completed more than a certain number of tasks with a specified rate of accuracy. Moreover, as discussed below, with minimal coding knowledge requesters can create and assign ad hoc "qualifications" to workers based on nearly any measurable attribute that grant specific workers access to tasks. Thus, smaller bespoke crowds can be constructed out of the workforce to complete highly specialized tasks.

Crowds are easy to program. For those with little experience programming machines, a major advantage of crowds is that they are comparatively easy to instruct. People are

experienced at communicating with each other, and actively work to make sense of their environment. People also interpret the pragmatic meaning of a request in far more detail than a literal reading would suggest, drawing upon contextual details and assumptions based on their own experience as communicators (e.g., that all relevant information is provided, and all provided information is relevant; for a discussion see Grice, 1989). As a result, crowds are tolerant to errors and ambiguity, and can easily go beyond the information provided to complete a task as the requester intended. In contrast, even when completing a task as simple as rating the positivity of words, a machine requires numerous variables to be defined including the universe of words to be rated, the context in which they might be used and the purpose the requester will use them to ensure an appropriate range and distribution of responses.

Limitations of Crowdsourcing Marketplaces

Although crowdsourcing marketplaces offer a number of compelling opportunities, there are also some potential challenges that may interfere with the accuracy of human computation. Speed and cost are inversely related to each other, and both are constrained by marketplace features beyond the control of individual requesters. Data quality may vary by marketplace, but also varies highly across tasks and workers and is thus under the direct control of requesters. We review several issues that pertain specifically to data quality.

Lack of motivation. While workers are to some extent intrinsically motivated to participate in crowdsourcing tasks (e.g., von Ahn, 2006), motivation is fickle and workers are inclined to avoid the most difficult elements of a task (Mason & Watts, 2009, Study 2). In this sense they can be regarded as "satisficers" who are likely to do only the minimal amount required to ensure payment (Simon, 1972). For example, if workers are asked to search for information on the Internet and are paid a reward even if they indicate that the requested

information is not available, they may be inclined to report that the information does not exist without a thorough search.

Cognitive limitations. Workers are people, and consequently suffer from a long but predictable set of cognitive and perceptual biases. This has led behavioral experimentalists within diverse disciplines to use workers as a subject pool for research (Goodman, et al., 2012; Paolacci et al. 2010; Rand, 2012). However, for the same reason, human computation researchers need to acknowledge that crowdsourced workers are not infallible computational agents, but rather are boundedly rational individuals that selectively allocate limited and depletable cognitive resources (for a general overview see Kahneman, 2011). These biases efficiently lead to perceptions beliefs and decisions that are "good enough" under most circumstances. These features may make crowdsourcing less suitable for some tasks where the requester seeks objectively correct answers through the aggregation of worker responses because aggregation cannot remove systematic bias.

Instruction ambiguity. The same cognitive abilities that make it possible for people to "program" a crowd with minimal instructions can pose problems for requesters because these processes will draw upon all information – both intentionally and unintentionally communicated – to understand a task. There are numerous examples of how design features such as response formats, question order and the affiliation of a communication partner guide inferences about the interviewer's intent and thus influence the responses provided (e.g., Bao, Sakamoto & Nickerson, 2011; for a review see Schwarz, 1999). Unfortunately, these features may be selected or communicated arbitrarily by requesters, without considering the effects they can have on worker's responses.

Workers may also make inferences about what a requester wants by drawing on their prior experiences with other requesters. For example, Goodman and colleagues (2012) conducted a decision making study in which they asked workers to guess the number of countries in Africa (adapted from Tversky & Kahneman, 1974). Although the authors did not explicitly ask workers to look this information up, an unusually large proportion of them gave answers that matched information available on the Internet. One explanation for this is that it is normative for MTurk workers to provide factually correct information, which led workers to believe that the requesters desired a factually correct answer rather than a subjective impression. Although little research has directly investigated this issue on Mechanical Turk, the importance of tacit norms in other workplaces has been extensively documented (Wegner, 1998).

Worker (non-)naivety. Workers may complete the same task several times or share information about tasks with each other. Prior knowledge about the contents or objectives of a task may benefit some crowdsourcing tasks. However, it is possible for workers to have *too much* information. At the most basic level, if the requester is interested in measuring the average rating of a target to smooth out the idiosyncratic beliefs of workers, it is obviously preferable to ensure that several different individuals rate it, rather than the same individual several times. Indeed, all "wisdom of crowds" tasks (Lyon & Pacuit, this volume) that aggregate worker responses *require* that judgments are made independently; when worker responses are not independent, errors will be correlated with each other and cannot be canceled out through aggregation (e.g., Anderson and Holt, 1997; Hullman, Adar & Shah, 2011). Independence across different tasks may also matter in more complex workflows. For example, if workers are required to complete several related tasks in stages, such as transcribing text and then rating

other workers' transcriptions for accuracy, requesters would want to avoid situations in which the same worker translates and evaluates the accuracy of their own translation.

The sheer size and anonymity of crowds makes it easy to underestimate the likelihood of duplicate workers. After all, with thousands of tasks and thousands of workers, what is the probability that the same worker would end up processing the same information twice? Two factors make this more likely than it might otherwise seem. First, workers tend to follow favorite requesters by subscribing to websites that alert them whenever favored requesters make work available for completion (e.g., www.turkalert.com). Second, workers complete varying numbers of tasks, with most of the work completed by a small group of extremely prolific workers. For example, we found that in a sample of sixteen thousand completed task submissions, the most prolific 1% of workers was responsible for completing 10% of the work, and the most prolific 10% were responsible for providing 41% of the observations (Chandler et al., in press, see also Berinsky, Huber & Lenz, 2012; Grady & Lease, 2010).

While Amazon by default prevents workers from completing the same task twice as a part of a single batch of tasks, additional measures (such as the use of Qualifications or third party software; Chandler et al., in press; Goldin & Darlow, 2013; Pe'er, Paolacci, Chandler & Mueller, 2012) must be used to ensure that workers across different tasks or across different HITs within the same batch of tasks are kept unique.

Workers may also share information with each other about the nature of a task, or collude in the responses they provide (Kazai & Milic-Frayling, 2009). Workers gather in forums (e.g., http://www.reddit.com/r/HITsWorthTurkingFor, mturkforum.com) to share information and opinions about tasks (e.g., particularly interesting and lucrative HITs), which could potentially lead them to have foreknowledge of certain task details. Thus, tasks that rely heavily on initial

impressions of a target of judgment, or tasks that screen out workers based on specific responses, should be designed with care to minimize worker foreknowledge.

Worker Honesty. Some tasks may require that people post information that is not directly verifiable or that has no factually correct response. For example, a requester may want to solicit opinions about a particular image or idea, or may want to know a worker's geographical location to assess their knowledge about local businesses. In general, workers provide factually accurate information (Shapiro et al., 2013) but deception can increase substantially is workers benefit from lying (Suri, Goldstein & Mason, 2011). In particular, on MTurk, large numbers of non-US workers claim to be US residents in order to receive cash payments (perhaps because workers in most other countries are paid with Amazon credit rather than cash).

Ensuring Data Quality in Crowdsourcing Marketplaces

Data quality is determined by numerous factors, some of which are under the control of requesters. Obtaining quality data is most straightforward for tasks that can be divided into many smaller components. This makes it easier for workers to select elements of the task that they enjoy or are good at while minimizing the learning curve. Further, smaller tasks are often completed more efficiently because minimally motivated workers can still provide useful data (Mason & Watts, 2009). Additional steps can be added to ensure quality control. For example, Mechanical Turk workers can successfully proofread and condense complex text, when a task is broken into smaller subtasks of finding problems, fixing problems and verifying proposed fixes (Bernstein et al., 2010).

For complex tasks, it may also be necessary to test worker ability before hand, and restrict access to workers who possess the necessary skills, or to consider other online labor markets (e.g. oDesk) that match requesters with more specialized workers. Regardless of the

software platform requesters use to recruit workers, they should also consider what software is best suited to the collection of work. Even sites like MTurk that allow tasks to be created using their own website also allow tasks to be created on a separate webpage or software program that is linked to or embedded within the web interface (Mason & Suri, 2012). Thus, requesters should not feel constrained by the platform used to distribute the work.

Task Design. There are many potential uses of crowdsourcing websites, and there is no one-size-fits-all solution to task design. In general, the approach requesters take when designing a task is more important than the specific design choices they make. Tasks should always be pilot tested, first by the requester and then by a small pool of workers, before being fully distributed to workers. Crowd interest is greatest when a HIT is first posted (Chilton, Horton, Miller & Azenkot, 2010), and minor mistakes can quickly become expensive. MTurk provides a "requester sandbox" in which the technical details of tasks can be tested by a requester. For pilot testing on workers, requesters should provide both the task of interest, and questions about the task of interest, to identify potential improvements in design (Collins, Joseph, & Bielaczyc, 2004). They should also have a clear benchmark against which the quality of work can be evaluated.

Although comparatively little research has been done on task design itself (for exceptions see Hosseini et al, 2012; Khanna, Ratan, Davis & Thies, 2010), there is a large literature on survey design that is relevant to requesters, which may be useful when considering data quality issues identified in pilot testing. Surveys are similar to crowdsourcing tasks in that instructions are communicated to workers rather than jointly discussed, and responses are collected through similar standardized methods. Consequently, it may be useful to requesters to consult a general

overview of web survey construction when designing tasks (e.g. Couper, 2008) in addition to more general resources on web design (e.g., Krug, 2009).

Screening Workers. As discussed earlier, MTurk allows requesters to select workers for inclusion in tasks based on whether or not they possess specific attributes. In general, workers with more experience and a higher reputation should be less likely to provide poor quality work. There is also evidence of differences in the quality of work provided by workers from different geographical locations, perhaps reflecting language difficulties or differences in education (Goodman et al., 2012; Kazai, Kamps, & Millec-Frayling, 2012). Alternatively, or additionally, requesters can create their own qualifications to screen workers according to more specific criteria such as their competence on particular tasks (e.g., Chua, Milosavlijevic, & Curran, 2009; Zhou, Resnick & Mei, 2011; for details on how to implement these procedures in Mechanical Turk see Chandler et al., in press).

Preventing Satisficing. Since many workers are motivated by money to complete tasks as efficiently as possible, satisficing (providing minimally adequate responses; Krosnick, 2006) is a major concern. Instructions or task elements can be presented sequentially with delays between each new piece of information to slow workers down (Kapelner & Chandler, 2010). Satisficing can be further reduced by introducing features that require workers to think about the "correct" response rather than simply providing their first impressions. One study asked workers and experts to evaluate the quality of Wikipedia pages. Worker ratings and expert ratings were uncorrelated, except when workers were also required to include answers to objectively verifiable questions (Kittur et al., 2008). Similarly, other researchers found that accuracy improved when workers were asked to predict how other workers would respond to a question

rather than simply offer their own opinion ("Bayesian truth serum"; Shaw, Horton & Chen, 2011; for a discussion see Prelec, 2004).

Worker motivation can also be increased. Crowds perform better on meaningful tasks (Chandler & Kapelner, 2013, see also Reed et al., this volume). Another alternative is to simply pay workers to pay attention. MTurk allows requesters to award bonuses to workers above and beyond the initial rate paid for completing work. Thus, requesters can structure a task to make it monetarily rewarding for workers to pay attention. To illustrate, in a pair of virtually identical studies (conducted by the third author of the present chapter), MTurk workers were paid either a total sum for participating ($1) or a smaller initial sum ($.30) with the remainder ($.70) paid as a bonus for successfully recalling details about the experimental manipulation. Although both sets of workers had the same potential earnings, those paid a smaller sum plus a performance bonus were more likely to correctly answer the factual multiple choice questions (98.2%) than participants who were paid a lump sum (87.0%), $\chi^2(1, N = 494) = 23.03, p < .001$ (see also Shaw, Horton & Chen, 2011). Interestingly, the success of bonuses in promoting attention seems to be independent of the bonus amount (Chandler & Horton, 2011).

Identifying Poor Quality Workers. There are a number of strategies that can be used to identify poor quality workers. Responses by workers who frequently disagree with their peers can be excluded (Elson & McKeown, 2010; Sheerman-Chase, Own & Bowden, 2011). Alternatively, "gold-standard" questions with factually correct answers, or "catch-trials" with obviously correct responses can be included along with the task of interest to measure worker ability and attentiveness (e.g., Sayeedet al., 2011). Tasks submitted along with incorrect responses to these questions can be excluded from analysis under the assumption that other components of the task are likely to also be incorrect. Additionally, or alternatively, all of the

responses provided by workers who fail a predetermined number of such checks can be excluded.

Multiple choice questions are frequently used to measure data quality because they are easily scored. The assumption is that workers who do not take the task seriously, or who do not understand the instructions, will likely respond at random, and are thus likely to select incorrect responses. In general, the sensitivity of gold-standard multiple choice questions to detect quality responses increases asymptotically: All else being equal, a single, four-item multiple choice question will only identify the 75% of random responders who select one of the three incorrect answers, while two four-item multiple choice questions will identify the 96% of random responders who select an incorrect answer on either or both questions. The actual ability of multiple choice questions to detect random responding is also dependent on the quality of the response alternatives (cf., Case & Swanson, 2001).

Measuring Data Quality. Data quality is often quantifiable and measurable. Reliability of categorical or continuous ratings can be evaluated based on its agreement with ground-truth, expert ratings or worker consensus. The critical question is whether agreement is sufficiently better than chance, although the level of agreement necessary is highly task dependent. Crowdsourced data is unusual in that not all workers complete all elements of a task (Krippendorff, 2004). Reliability of data with this property can be measured using Krippendorff's alpha (for SAS and SPSS macros see Hayes & Krippendorff, 2007). High reliability scores between workers is a function of both task difficulty and the number of raters and is a necessary precondition for valid responses. If reliability is low, it could suggest poorly communicated instructions or a plurality of acceptable answers. Reliability can be increased by refining worker instructions and increasing the number of workers who perform each task.

Cleaning and aggregating responses. Responses by different workers can also be combined. In general, aggregating the ratings of many independent judgments, even through averaging or a simple majority, will increase their accuracy, as idiosyncratic errors cancel each other out (Galton, 1907). More complex methods of aggregating responses can improve data quality yet further. Some approaches use quantitative methods to improve quality, trimming responses that are likely to be outliers (Jung & Lease, 2011) or estimating worker quality and then weighting their responses on specific tasks accordingly (Hosseini et al., 2012; Tang & Lease, 2011). Other approaches use workers themselves to review and combine responses in an interactive, iterative process (Nickerson et al., 2011).

As a final note, aggregation does not increase the likelihood of a correct solution unless each judgment is independent. If a majority of answers are identical but agreement is not independent - either because workers have discussed their responses beforehand or because care was not taken to avoid duplicate respondents (see limitations section) - then the value of the majority's opinion may be suspect. Likewise, aggregation will not provide a correct solution for problems in which workers are systematically wrong, either because they lack the necessary information to reach a correct conclusion or because cognitive biases lead workers to draw incorrect conclusions.

Conclusions

Crowdsourcing marketplaces present an opportunity to researchers who require human computation services, especially for tasks that are small, require a variety of different skills or interests, or are intermittent in their availability. They offer a persistent workforce that is available on demand for an affordable price. However, data provided by workers is not inevitably high quality: tasks must be designed to maximize the likelihood and ease with which workers

can provide useful responses. This is fundamentally an iterative process, and worker feedback in initial stages can provide insight into improving tasks.

While specific design considerations largely depend on the researcher's goals, task design can be improved iteratively through pilot testing, and a number of principles exist that can improve the quality of data collected on crowdsourcing marketplaces. In particular, crowd members are heterogeneous and requesters can take advantage of this by preselecting workers who are most capable of performing specific tasks. Further, tasks can be optimized so that workers can understand them and feel motivated to complete them correctly. Finally, despite varying rates of participation by individual workers, quality can be measured, and to a certain extent improved, through aggregating responses. In this sense, the output of the crowd can be greater than the sum of its parts.

Online marketplaces have developed rapidly in the past few years. While it is notoriously difficult to predict what will happen in the future (e.g., Tetlock, 2005), there are a few developments that seem particularly plausible. Network effects give Mechanical Turk a large competitive moat against alternative platforms, but individuals are working to counter some of its limitations within its current framework. Requesters are beginning to use it as merely a gateway to request labor, and are directing workers to complete tasks on other software platforms that allow dynamic and real-time collaborative tasks.

Perhaps more crucially, workers and requesters alike are developing the means to increase market transparency. While Amazon has implemented minimal channels for transmitting information directly between requesters and workers, and indirectly between various requesters, much of the increased transparency discussed in this chapter is a result of requesters and workers finding their own means of communicating with each other outside of Amazon's platform.

However, information exchange is still relatively limited. There is no public register of market participants, and workers can only be recruited using a narrow range of metadata. Additionally, requesters are unable to access information about general market conditions or task completion rates that would allow them to optimize tasks and compensation rates, or to directly match tasks with workers of varying levels of skill and motivation. Often, requesters must build their own panel of workers (which takes time) based on information that was privately collected, or shared in informal, insecure ways. Perhaps worse, workers have no access to requesters' profiles, making the relationship between Requesters and Workers inherently asymmetrical. Some workers rely on independent websites that allow workers to rate and subscribe to requesters. However, in general workers are unable to determine which tasks pay fairly and which qualifications are worth the unpaid effort necessary to complete them. For requesters, completions times thus depend heavily on whether their tasks are credentialed in an external forum (Chandler et al., in press). More generally, poor quality requesters run the risk of creating something close to a "market of lemons" in which the highest quality workers refuse to participate because of these issues (Akerlof, 1970; for a discussion see Horton, 2010). All of these issues hinder the effectiveness of MTurk as a labor market, and we anticipate that workers and requesters will continue to increase information exchange and transparency.

Another interesting question is what tasks online labor markets will be used for in the future. As machine perception and language processing improve, it is likely that demand for human and human-machine hybrid computational solutions will no longer be needed for these tasks. Just as steam drills replaced railroad workers, and office productivity software has replaced middle class white collar employees, so too will software replace crowds, for some tasks. It remains to be seen whether crowdsourcing, especially microtask labor markets, are

merely a solution to temporary deficiencies in the advance of machine computing, or if, as has occurred in other labor markets, new tasks will continue to emerge as a technology advances. For instance, as workflow management platforms become more automated, iterative tasks may become possible. As research about task decomposition develops, there will be opportunities to use microtask markets for problems that require increasingly complex and creative solutions. As more data becomes digitized and interconnected, there will be more opportunity to search for interrelations between increasingly disparate topics. Finally, a larger sociological question that remains to be answered is how these changes within crowdsourcing marketplaces may impact other labor markets and society at large.

References

Abbott, R., Walker, M., Anand, P., Tree, J. E. F., Bowmani, R., & King, J. (2011). How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Language in Social Media* (pp. 2-11). Stroudsburg, PA: Association for Computational Linguistics.

Akerlof, G. A. (1970). The market for" lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, 488-500.

Von Ahn, L. (2006). Games with a purpose. *Computer*, 39, 6, 92-94.

Allio, R. J. (2004). CEO interview: the InnoCentive model of open innovation.*Strategy & Leadership*, *32*(4), 4-9.

Anderson, L. R., & Holt, C. A. (1997). Information cascades in the laboratory.*The American economic review*, 847-862.

Bao, J., Sakamoto, Y., & Nickerson, J. V. (2011). Evaluating Design Solutions Using Crowds. In *Proceedings of the 17th Americas Conference on Information Systems*.

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education, 53*, 182-200.

Becker, G. S., & Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *The Quarterly Journal of Economics*, *107*(4), 1137-1160.

Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D., & Panovich, K. Soylent: A Word Processor with a Crowd Inside. In *Proc. UIST 2010*, ACM Press (2010), 313-322.

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*, 351–368. doi:10.1093/pan/mpr057

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., ... & Yeh, T. (2010, October). VizWiz: nearly real-time answers to visual questions. In*Proceedings of the 23nd annual ACM symposium on User interface software and technology* (pp. 333-342). ACM.

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

Case, S.M. & Swanson, D.B. (2001) Constructing Written Test Questions for the Basic and Clinical Sciences, 3rd edn (Philadelphia, National Boardof Medical Examiners).

Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*.

Chandler, D., & Horton, J. (2011, August). Labor Allocation in Paid Crowdsourcing: Experimental Evidence on Positioning, Nudges and Prices. In*Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Chandler, J., Mueller, P., & Paolacci, G. (in press). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*.

Chilton, L. B., Horton, J. J., Miller, R. C., & Azenkot, S. (2010, July). Task search in a human computation market. In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 1-9). ACM.

Chua, C. C., Milosavljevic, M., & Curran, J. R. (2009). A sentiment detection engine for internet stock message boards. In L. A. Pizzato & R. Schwitter (Eds.), *Proceedings of the Australasian Language Technology Association Workshop 2009* (pp. 89-93). Sydney: Australia.

Collins, A., Joseph, D., & Bielaczyc. K. (2004). Design research: Theoretical and methodological issues." *The Journal of the learning sciences* 13(1), 15-42.

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

Couper, M. (2008). *Designing effective Web surveys*. New York, NY: Cambridge University Press.

Davis, L. E. (1965). Pacing effects on manned assembly lines. *International Journal of Production Research*, *4*(3), 171-184.

Dominowski, R. L., & Dallob, P. I. (1995). Insight and problem solving. In R. J. Sternberg & J. E. Davidson (Eds.), The nature of insight (pp. 33-62). Cambridge, MA: MIT Press

Elson, D. K., & McKeown, K. R. (2010). Automatic attribution of quoted speech in literary narrative. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (pp. 1013–1019). Menlo Park, CA: The AAAI Press.

Estellés-Arolas, E., & González-Ladrón-de-Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, *38*(2), 189-200.

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

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

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives*, 191-209.

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. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. Journal of Behavioral Decision Making, *26*(3), 213-224.

Grady, C., & Lease, M. (2010). Crowdsourcing document relevance assessment with Mechanical Turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk* (pp. 172-179). Association for Computational Linguistics.

Grice, H.P. (1989). *Studies in the Way of Words*. Harvard University Press.

Gruenstein, A., McGraw, I., & Sutherland, A. (2009). A self-transcribing speech corpus: Collecting continuous speech with an online educational game. In *Proceedings of the Speech and Language Technology in Education (SLaTE) Workshop.* Warwickshire, England.

Horton, J. J. (2011). The condition of the Turking class: Are online employers fair and honest?. *Economics Letters*, *111*(1), 10-12.

Horton, J. J. (2010). *Online labor markets* (pp. 515-522). Springer Berlin Heidelberg.

Hsieh, G., Kraut, R. E., & Hudson, S. E. (2010). Why pay?: Exploring how financial incentives are used for question & answer. *Proceedings of the 28th International Conference on Human Factors in Computing Systems,* 305–314. doi: 10.1145/1753326.1753373

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*, 77–89. doi: 10.1080/19312450709336664

Horton, J. J., & Chilton, L. B. (2010, June). The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce* (pp. 209-218). ACM.

Hosseini, M., Cox, I., Milić-Frayling, N., Kazai, G., & Vinay, V. (2012). On aggregating labels from multiple crowd workers to infer relevance of documents.*Advances in Information Retrieval*, 182-194.Ipeirotis, P. (2010). Demographics of Mechanical Turk. *CeDER-10–01 working paper*, New York University

Hullman, J., Adar, E., & Shah, P. (2011, May). The impact of social information on visual judgments. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (pp. 1461-1470). ACM.

Ipeirotis, P. G., & Horton, J. J. (2011). The Need for Standardization in Crowdsourcing. CHI.

Jung, H. J., & Lease, M. (2011, August). Improving Consensus Accuracy via Z-score and Weighted Voting. In *Proceedings of the 3rd Human Computation Workshop (HCOMP) at AAAI*.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kaufmann, N., Schulze, T., & Veit, D. (2011). More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk. In *Proceedings of the Seventeenth Americas Conference on Information Systems, Detroit, MI*.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2012). The Face of Quality in Crowdsourcing Relevance Labels: Demographics, Personality and Labeling Accuracy.

Kazai, G., & Milic-Frayling, N. (2009, July). On the evaluation of the quality of relevance assessments collected through crowdsourcing. In *SIGIR 2009 Workshop on the Future of IR Evaluation* (p. 21).

Kapelner, A., & Chandler, D (2010). Preventing Satisficing in Online Surveys: A 'kapcha' to ensure higher quality data. In The World's First Conference on the Future of Distributed Work (CrowdConf2010).

Khanna, S., Ratan, A., Davis, J., & Thies, W. (2010, December). Evaluating and improving the usability of mechanical turk for low-income workers in india. In *Proceedings of the first ACM symposium on computing for development* (p. 12). ACM.

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

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013). The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 1301-1318). ACM.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research*, *30*(3), 411-433.

Krosnick, J.A. (2006). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5(3)*, 213-236.

Krug, S. (2009). *Don't make me think: A common sense approach to web usability*. New Riders.

Lane, I., Weibel, A., Eck, M., & Rottmann, K. (2010) Tools for collecting speech corpora via Mechanical-Turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (184-187). Stroudsbug, PA: Association for Computational Linguistics.

Lakhani, K. R. (2008). InnoCentive. com (A). Harvard Business School Case, (608-170).

Lau, T., Drews, C., & Nichols, J. (2009). Interpreting written how-to instructions. In H. Kitano (Ed.), *Proceedings of the 21st International Joint Conference on Artificial Intelligence* (pp. 1433-1438). San Francisco, CA: Morgan Kaufmann Publishers Inc.

Li, B., Liu, Y., & Agichtein, E. (2008). *CoCQA*: Co-training over questions and answers with an application to predicting question subjectivity orientation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 937-946). Stroudsburg, PA: Association for Computational Linguistics. doi: 10.3115/1613715.1613836

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

Marge, M., Banerjee, S., Rudnicky, A.I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* (5270–5273). Washington, DC: Institute of Electronics and Electrical Engineers. doi: 10.1109/ICASSP.2010.5494979

Mason, W., & Watts, D. J. (2009, June). Financial incentives and the performance of crowds. In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 77-85). ACM.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods*, *44*(1), 1-23.

Molla, D., & Santiago-Martinez, M. E. (2011). Development of a corpus for evidence based medicine summarisation. In *Proceedings of Australasian Language Technology Association Workshop* (pp. 86-94). Melbourne, Australia: Australasian Language Technology Association.

Nelson, L., Held, C., Pirolli, P., Hong, L., Schiano, D., & Chi, E. H. (2009). With a little help from my friends: Examining the impact of social annotations in sensemaking tasks. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (pp. 1795-1798). New York, NY: ACM. doi: 10.1145/1518701.1518977

Nickerson, J. V., Sakamoto, Y., & Yu, L. (2011). Structures for creativity: The crowdsourcing of design. In *CHI Workshop on Crowdsourcing and Human Computation* (pp. 1-4).

Open Science Collaboration. (2013). The Reproducibility Project: A model of large-scale collaboration for empirical research on reproducibility. In V. Stodden, F. Leisch, & R. Peng (Eds.), *Implementing Reproducible Computational Research (A Volume in The R Series)*. New York, NY: Taylor & Francis.

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

Pe'er, E., Paolacci, G., Chandler, J., & Mueller, P. (2012). Screening participants from previous studies on Amazon Mechanical Turk and Qualtrics. *Available at SSRN 2100631*.

Pontin, J. (2007, March 25). Artificial intelligence, with help from the humans. *New York Times*, *25*. Retrieved from http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*(5695), 462-466.

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.

Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45-48.

Rochet, J. C., & Tirole, J. (2003). Platform competition in two‐sided markets.*Journal of the European Economic Association*, *1*(4), 990-1029.

Sayeed, A. B., Rusk, B., Petrov, M., Nguyen, H. C., Meyer, T. J., & Weinberg, A. (2011). Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *Proceedings of the 5th ACL-HLT Workshop on*

Language Technology for Cultural Heritage, Social Sciences, and Humanities (pp. 69-77). Stroudsburg, PA: Association for Computational Linguistics.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American psychologist*, *54*(2), 93.

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

Shaw, A. D., Horton, J. J., & Chen, D. L. (2011, March). Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 275-284). ACM.

Sheerman-Chase, T., Ong, E. J., & Bowden, R. (2011). Cultural factors in the regression of non-verbal communication perception. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 1242-1249).

Silberman, M., Irani, L., & Ross, J. (2010). Ethics and tactics of professional crowdwork. *XRDS: Crossroads, The ACM Magazine for Students*, *17*(2), 39-43.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*,*1*, 161-176.

Suri, S., Goldstein, D. G., & Mason, W. A. (2011). Honesty in an online labor market. In L. von Ahn, & P. G. Ipeirotis (Eds.), *Papers from the 2011 AAAI Workshop.* Menlo Park, CA: AAAI Press.

Tang, W., & Lease, M. (2011). Semi-supervised consensus labeling for crowdsourcing. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.

Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know?*. Princeton University Press.

Tetreault, J. R., Filatova, E., & Chodorow, M. (2010, June). Rethinking grammatical error annotation and evaluation with the Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 45-48). Association for Computational Linguistics.

Tversky, A., Kahneman, D. (1974), Judgment under Uncertainty: Heuristics and Biases, *Science*, 211 (January), 453-458.

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. *Science*, *321*(5895), 1465-1468.

Wenger, Etienne (1998). Communities of Practice: Learning, Meaning, and Identity. Cambridge: Cambridge University Press

Wichester, S. (2003). *The meaning of everything: The story of the Oxford English Dictionary*. Oxford University Press.

Yu, L., & Nickerson, J. V. (2011, May). Cooks or cobblers?: crowd creativity through combination. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (pp. 1393-1402). ACM.

Zhou, D. X., Resnick, P., & Mei, Q. (2011). Classifying the political leaning of news articles and users from user votes. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 417-424). Menlo Park, CA: The AAAI Press.