A Panel For Lemons?
Positivity bias, reputation systems and data quality on MTurk

Ted Matherly[1]

12 July 2018

Forthcoming, *European Journal of Marketing*

[1] Ted Matherly is Assistant Professor of Marketing at the Spears School of Business, Oklahoma State University. Address: 438 BU, Oklahoma State University, Stillwater, OK 74078-4011. Phone: 405-744-5139 Email: ted.matherly@okstate.edu.

Abstract

A Panel For Lemons?
Positivity bias, reputation systems and data quality on MTurk

Purpose The purpose of this research is to investigate how the effectiveness of systems for ensuring cooperation in online transactions are impacted by a positivity bias in the evaluation of the work that is produced. The presence of this bias can reduce the informativeness of the reputation system, and negatively impact its ability to ensure quality.

Methodology This research combines survey and experimental methods, collecting data from 1875 MTurk workers in five studies designed to investigate the informativeness of the MTurk reputation system.

Findings The findings demonstrate the presence of a positivity bias in evaluations of workers on MTurk, which leaves them undifferentiated, except at the extremity of the reputation system and by status markers.

Research limitations Because MTurk workers self-select tasks, the findings are limited in that they may only be generalizable to those who are interested in research-related work. Further, the tasks employed in this research are largely subjective in nature, which may decrease their sensitivity to differences in quality.

Practical implications For researchers, the results suggest that requiring 99% approval rates (rather than the previously advised 95%) should be used to identify high quality workers on MTurk.

Value The research provides insights into the design and use of reputation systems, and demonstrates how design decisions can exacerbate the effect of naturally occurring biases in evaluations to reduce the utility of these systems.

Word count: 9,811
Keywords: reputation systems, positivity bias, online research, methodology, data quality

A Panel For Lemons

Positivity bias, reputation systems and data quality on MTurk

Since its introduction in 2004, Amazon's Mechanical Turk (MTurk) has become an essential tool for marketing researchers, enabling the rapid collection of data from thousands of participants worldwide. With 15,000 papers referencing MTurk published between 2006 and 2014 (Chandler & Shapiro, 2016), the importance of this tool, along with other markets for crowdsourced work such as CrowdFlower and Prolific Academic (Peer, Brandimarte, Samat & Acquisti, 2017; Vakharia & Lease, 2015), is difficult to overstate. As the popularity of these crowdsourcing panel systems has grown, a considerable body of work has emerged characterizing the nature of respondents and codifying best practices in the panels' use. A cornerstone of these recommendations is an emphasis on methods of ensuring high quality in the data that workers produce. Though some recent research has investigated the prevalence of various forms of worker deception on MTurk (Chandler & Paolacci, 2017; Sharpe Wessling, Huber & Netzer, 2017; Suri, Goldstein & Mason, 2011), for many academic research paradigms, the quality of work is more subjective.

When this is the case, data quality may be less a question of objective truth, and is more likely subject to "careless" or "inattentive" responses (Fleischer, Mead & Huang, 2015; Meade & Craig, 2012), where participants respond to questions without regard for their content. A variety of work has employed metrics based on the psychometric properties of scale responses (Buhrmester, Kwang & Gosling, 2011; Litman, Robinson & Rosenzweig, 2015; Rouse, 2015) and evidence of patterned and random responding (Chandler, Mueller & Paolacci, 2014; Peer, Vosgerau & Acquisti, 2014) to capture this aspect of data quality. This work has generally shown that results from MTurk are of reasonable quality (Behrend, Sharek, Meade & Wiebe, 2011; Kees, Berry, Burton & Sheehan, 2017; Mason & Suri, 2012) and can replicate findings from research using other participant pools (Buhrmester et al., 2011; Crump, McDonnell & Gureckis, 2013; Peer et al., 2017).

In spite of these findings, beliefs persist among researchers regarding quality problems with MTurk workers (Fleischer et al., 2015). To help assuage these concerns and ensure that workers meet the expectations of employers, the MTurk platform provides access to a reputation system, which contains measures that are intended to provide information about worker quality. These include, most notably, the number of tasks workers have completed, and the overall percentage of the tasks they have completed that have been approved. The reputation system can be used to limit those eligible to participate in a specific task to a subset of the overall workers. Most guides for researchers suggest only recruiting workers who have sufficiently high reputations (Goodman & Paolacci, 2017; MTurk Blog, 2012; Sheehan & Pittman, 2016), and authors frequently report the levels of the reputation system under which participants were recruited (e.g. Antonetti and Maklan, 2017; Goldstein, Suri, McAfee, Ekstrand-Abueg and Diaz, 2014), implying that these standards may confer information about the quality of the data these participants provided.

However, the way in which employers affect workers' reputations on MTurk – by accepting or rejecting their work – also invokes another mechanism for enforcing cooperation between parties: litigation, where damages are assessed and transferred from one party to another. On MTurk, this entails workers not receiving a wage from the requester. Considerable economic and legal work has explored how litigation and reputation mechanisms affect contract enforcement (Bakos & Dellarocas, 2011; MacLeod, 2007). These functions are usually considered to operate independently, but by combining these two functions, the effectiveness of both may be hampered by biases in requester evaluation. Of particular concern is the presence of a *positivity bias* (Bridges & Vásquez, 2016; Moe & Schweidel, 2012; Zervas, Proserpio & Byers, 2015), where requester evaluation is biased in favor of workers (Horton & Golden, 2015). The presence of such a bias could decouple the reputation system from worker quality.

While some prior research has observed higher data quality from higher reputation workers (Peer et al., 2014), other work has not observed meaningful differences (Suri et al., 2011). This mixed evidence for quality sorting, combined with the relatively low wages paid to workers (Horton & Chilton, 2010), has led some researchers to label MTurk a “market for lemons,” (Ipeirotis, 2010; Mason & Suri, 2012), recalling the classic theoretical study of used car markets (Akerlof, 1970). In this conception, when buyers cannot determine whether cars are of high or low quality before purchase, they assume all cars are of average quality and pay prices based on this assumption. However, because this price is below high quality sellers’ valuations, these sellers choose not to participate, leaving only sellers of low quality “lemons” in the market. A similar intuition could be applied to MTurk because, without an informative reputation system, requesters would not be able to distinguish high and low quality workers. Therefore, requesters expect to collect data from a mixture of high and low quality workers, and offer compensation below the rate that high quality workers expect. High quality workers thus exit the market, leaving behind only those of low quality.

The goal of this research is two-fold: first, to demonstrate the presence of the conditions necessary for and consistent with a failure of the reputation system to sort workers based on quality; and second, to test the theoretical and practical implications of this failure. The results are consistent with the presence of biased reporting of outcomes by requesters, which leads quality sorting to be unobserved at levels recommended by prior work. However, quality differences are observed based on status markers – formal recognition of a reputation for quality separately granted from reputation (e.g. an award or medal; Azoulay, Stuart and Wang, 2013; Malter, 2014), and for those workers with extremely positive reputations, whose approval rates are close to the highest levels possible within the system. In the following sections, relevant literature on methods of ensuring cooperation in online transactions is examined and then used to derive hypothetical scenarios for the platform. Five studies are presented that explore these scenarios, employing both descriptive and experimental methods. The paper closes with a discussion of these findings in the context of theory on reputation systems, along with practical suggestions for more effective use of MTurk when conducting research – specifically, it is recommended to use a 99% approval rate threshold to identify high quality workers.

ENFORCING COOPERATION IN ONLINE TRANSACTIONS

In online markets where participants interact without face-to-face contact, systems that enable cooperation by creating trust and ensuring good conduct are necessary. In these transactions, the fundamental problem is one of incomplete and asymmetric information: typically, one party (the seller) possesses private information about their type or effort level, and produces high or low quality work, while the other party (the buyer) has a preference for high quality, and wishes only to do business with those who provide high quality output (Akerlof, 1970). When the quality of the seller's product is hard to assess until some point after purchase, such as with an experience or credence good (Gao, Greenwood, Agarwal & McCullough, 2015; Kokkodis & Ipeirotis, 2015; Nelson, 1970), this information asymmetry creates the possibility for two negative outcomes for the marketplace: adverse selection and moral hazard. With adverse selection, low quality sellers attempt to conceal their type from buyers, allowing them to charge higher prices, while moral hazard occurs when there is no incentive for high quality workers to maintain quality, and they engage in opportunistic behavior by shirking (Pavlou, Liang & Xue, 2007). With both of these possibilities, buyers receive products of lower than expected quality, which undermines trust and may lead buyers to exit entirely.

To address this potential for market failure, a variety of strategies exist, with perhaps the most important being reputation. Reputations are information about prior transactions which serve as a signal of unobserved quality, and can resolve the information asymmetry (Kreps & Wilson, 1982). This information is transmitted in online transactions through reputation systems (Dellarocas, 2003; Resnick & Zeckhauser, 2002), which aggregate information about the number of transactions a seller has completed and ratings of the seller by the buyers in each transaction (Ghose, Ipeirotis & Sundararajan, 2005). As this information is publicly available to everyone in the market, an incentive is created for parties to cooperate because a negative evaluation of a transaction may limit the ability to engage in future transactions (Dellarocas, 2003). By producing quality work, as demonstrated through completion of transactions to the satisfaction of buyers over a period time, the seller is able to establish their reputation (Malter, 2014). Other parties can then use this information to make decisions about future transactions. The effectiveness of a reputation system depends upon its ability to provide information that distinguishes between high and low quality sellers, encourages participation from high quality sellers, and discourages participation from those of low quality (Resnick & Zeckhauser, 2002).

By contrast to reputation, where the effectiveness in encouraging high quality work depends on an expectation that the seller will stay in the market over time, litigation enables a buyer to more directly impact the seller's incentives in the current transaction. This is because litigation imposes a threat of monetary damages exerted on the seller through a third party, which encourages high quality workers to self-select and to exert effort in their work (Bakos & Dellarocas, 2011). While this classically takes the form of tort actions involving the legal system, information technology has enabled similar actions to occur through online dispute and buyer protection mechanisms, and these systems can also create trust between parties (Pavlou & Gefen, 2004). This is because it provides a mechanism to recover from the receipt of low quality output, resolving potential concerns about adverse selection, while also motivating workers to maintain high quality thus preventing moral hazard.

The formulation of the litigation mechanism includes the amount of damages that can be assessed, the cost of engaging in the action, as well as the probability of receiving these awards (Bakos & Dellarocas, 2011; Bebchuk, 1984). In traditional litigation, damages can be very high, but this is balanced by the cost of pursuing court action and uncertainty associated with the outcome. On the other hand, dispute resolutions systems for online transactions may involve substantially lower stakes, both in terms of costs and damages. In these contexts, the probability of a litigation effort's success also varies, from the low buyer success rates in early iterations of the eBay Buyer Protection program (Clemons, 2007), to the highly probable outcomes in buyer's favor on the MTurk platform (Silberman, Ross, Irani & Tomlinson, 2010). In the latter cases, in a one-shot game, the rational action is for the buyer to always engage in litigation, since the expected value of recovering damages is high. However, because fraudulent behavior by buyers may result in sanction by the third party, including expulsion from the platform, this type of abuse is mitigated (Morriss & Korosec, 2005).

REPUTATION AND LITIGATION ON MTURK

Because of the effectiveness of both reputation and litigation in ensuring cooperation, they have become indispensable tools for online transactions, and both feature prominently on MTurk. In the research market on MTurk, researchers, as MTurk requesters, act as "buyers" of labor (an experience good, Kokkodis and Ipeirotis, 2015) from "sellers," i.e. MTurk workers. To facilitate these transactions, MTurk provides a reputation system and a system akin to litigation, in the ability to accept or reject work. The reputation system automatically tracks several measures that are referred to as "system qualifications." The two primary qualifications are productivity, measured as the number of tasks a worker has completed, and the approval rate, measured as the percentage of the overall number of tasks completed that have been approved by requesters (Sheehan & Pittman, 2016). MTurk's reputation system also measures workers' "Masters" status, who are identified by MTurk as "elite groups of Workers who have demonstrated accuracy on specific types of [tasks]" (Amazon, 2018). However, the criteria under which workers receive this status is unclear (Yin, Gray, Suri & Vaughan, 2016), beyond having completed 1,000 HITs and maintaining a 99% approval rating (Clickhappier, 2016). Thus, being granted the Masters qualification confers a status on workers that distinguishes them and their quality in a manner that is separate from their reputation (Azoulay et al., 2013; Malter, 2014; Simcoe & Waguespack, 2011).

In addition to the system qualifications that are managed by MTurk, requesters are also able to create their own qualifications, and these requester-generated qualifications enable the implementation of panel designs, where workers complete a task piecemeal and requesters can decide if individual workers may participate in subsequent stages. A number of recent researcher-focused methodological guides have advocated the use of panel designs (Chandler et al., 2014; Goodman, Cryder & Cheema, 2013; Goodman & Paolacci, 2017; Sharpe Wessling et al., 2017) because they provide researchers with the ability to identify quality sorting in workers independent of system qualifications. These recommendations come with at least a tacit suggestion that the system qualifications may not effectively distinguish workers in terms of quality.

When collecting data on MTurk, researchers create new “Human Intelligence Tasks" (HITs) by first designing the task itself, typically done by linking from MTurk to an external host for a survey instrument, and then choosing the parameters under which workers are compensated and deemed eligible to complete the task. The worker’s eligibility for a task is determined by satisfying the qualifications that have been chosen. Most guides for academic researchers suggest that thresholds for the system qualifications be used, with values of 95% to 97% frequently identified for approval rates (Berinsky, Huber & Lenz, 2012; Hauser & Schwarz, 2016; Peer et al., 2014; Sharpe Wessling et al., 2017; Sheehan & Pittman, 2016; Staffelbach et al., 2015), and the use of values in this range as a threshold is regularly observed in marketing research employing MTurk (e.g. Antonetti and Maklan (2017), Goldstein et al. (2014), Goodman and Paolacci (2017)). MTurk itself had previously set a default value of 95% for approval rates, which may also contribute to the frequency of its use by researchers. However, MTurk now only recommends requesters require that workers have 5,000 HITs approved and have an approval rate of 95% (MTurk Blog, 2012).

As with any reputation system, the effectiveness of these recommendations depends upon its ability to sort workers by quality – that is, it differentiates worker quality through levels observable to requesters. But this is not a foregone conclusion, because online reputation systems often suffer from systematic biases (Chevalier & Mayzlin, 2006; Fradkin, Grewal, Holtz & Pearson, 2015; Godes & Silva, 2012; Hu, Pavlou & Zhang, 2006; Hu, Zhang & Pavlou, 2009; Zervas et al., 2015). These issues are frequently observed as a reporting bias, where only a subset of individuals provide feedback, typically those who are satisfied with a transaction (Dellarocas & Wood, 2008; East, Hammond & Wright, 2007; Hu et al., 2009; Li & Hitt, 2008; Reichling, 2004).

These biases are problematic when considering the litigation method that is used on MTurk, which is contained in a requester's decision to accept (and pay) a worker, or to reject their work without pay. This serves as litigation because it enables requesters to recover damages (the entirety of the wage they would otherwise pay to the worker) after the task is delivered, and the probability of recovering the payment to the worker is all but a certainty. Making this decision is not voluntary, as all work is automatically approved in, at most, 30 days, thus all workers eventually receive some type of feedback. Therefore, while the decision can not be characterized as a reporting bias where only some users provide feedback and many are "silent" (Dellarocas & Wood, 2008; Ghose, Ipeirotis & Sundararajan, 2009), a systematic positivity bias may instead occur (Horton & Golden, 2015). This would lead requesters to be unlikely to reject low quality work, and could be driven by multiple reasons, some of which are well known such as leniency, reciprocity and fear of retaliation (Dellarocas & Wood, 2008; Fradkin et al., 2015), while others are related to the fusing of the litigation and reputation mechanisms.

Because the tasks that are used on MTurk in conducting research are subjective, there may be uncertainty in assessing their quality, which may lead to leniency in evaluation (Kusterer, Bolton & Mans, 2016). Further, though direct reciprocity in evaluations can not occur within the MTurk platform because there is no internal reputation system for requesters, third party reputation systems for requesters exist, such as TurkOpticon (Irani & Silberman, 2013; Sheehan & Pittman, 2016; Silberman & Irani, 2016). Workers can leave feedback about their experiences with requesters, and requesters trying to establish a reputation may be less likely to reject low quality work. Finally, the fear of retaliation for rejecting work may also play a role in these decisions for the same reasons, because rejections may negatively impact a requester's reputation on these information sources and limit their ability to recruit future workers (Clemons, 2007; Gao et al., 2015).

In addition, with punishment through litigation and reputation combined into a single mechanism, there are several further reasons to expect a positivity bias to emerge. First, despite that there is an economic incentive to do so by recovering the payment to a worker (Zhou, Dresner & Windle, 2008), the process of identifying and rejecting low quality work may take a non-trivial amount of time and effort. Thus, the small amount of payment recovered may not be balanced against the ease with which a set of responses can be marked "Approve All" in the MTurk interface. Second, because they impact both reputation and income, rejections will frequently lead workers to contact the requesters and ask to have these rejections reversed (Brawley & Pury, 2016), creating an additional cost. Third, rejecting a worker may have implications related to Institutional Review Board rules for academic requesters regarding payments and the ability of participants to withdrawal from studies without penalty (Paolacci, Chandler & Ipeirotis, 2010), thereby preventing the use of the mechanism by academic researchers.

Taken together, these factors impose substantial costs of employing the litigation mechanism to reject work, and because these costs can outweigh its ability to recover damages, it may not be used by requesters. However, because it is tied to the reputation mechanism, these positively biased actions also impact worker reputations, partially disconnecting their relationship to the true quality of the workers (Gao et al., 2015; Hu et al., 2006). If there is in fact such a positivity bias, this would likely be observed in a highly skewed distribution of reputations for sellers (Zervas et al., 2015), and findings from worker self-reports have observed 99% average approval rates for the MTurk population (Schulze, Nordheimer & Schader, 2013; Sharpe Wessling et al., 2017; Yin et al., 2016). However, the presence of a skew alone is not sufficient evidence to suggest that a positivity bias is present. Further, it also does not suggest that the reputation system is not informative, though evidence for this is decidedly mixed. Some studies showing little or no observable differences (Downs, Holbrook & Peel, 2012; Eickhoff & de Vries, 2013), while others have shown differences in worker quality as a function of better reputations, as measured by approval rates (Naderi, Polzehl, Wechsung, Köster & Möller, 2015; Peer et al., 2014). In Naderi et al. (2015), a weak negative effect was observed in the inconsistency of responses at higher levels of discretized approval rates, while Peer et al. (2014) found stronger evidence of differences between workers with low approval rates and high approval rates.

Theoretically, this skew in ratings and mixed evidence for informational content can be accounted for in several ways. Assuming there is no positivity bias, then the reputation system accurately reflects the true quality of the workers and their effort levels. This would suggest that all low quality workers have exited the market, and the remaining high quality workers are sufficiently motivated by the litigation and reputation mechanisms to maintain quality. In this case, there is little worker heterogeneity, nor adverse selection or moral hazard. However, if there is in fact a positivity bias among requesters, it presents three potential scenarios, depending upon whether there is differentiation among workers, and whether the reputation system provides information about these differences (summarized in table 1):

Scenario 1. *The distribution of reputation represents the true quality of the workers, and the low variance and lack of information in the reputation system is because all workers are effectively the same. As all workers are of equal quality, adverse selection is not relevant, but moral hazard may be present because the positivity bias reduces the expected penalty from litigation.*

The intuition in this scenario is that there is no differentiation between workers, because all low quality workers have been driven out of the market, leaving only high quality workers. Thus, the primary concern is that while all workers are capable of producing high quality work, there may be little incentive to do so if requesters have a positivity bias and are unlikely to reject work on the margin. This gives rise to moral hazard, where high quality workers shirk, choosing not to exert effort without a credible threat of response by requesters. Consistent with this scenario, prior research (Chandler & Paolacci, 2017; Sharpe Wessling et al., 2017) has observed workers exhibiting shirking behaviors, in the form of falsely claiming traits to gain access to work, because these behaviors are difficult to identify and unlikely to be punished by requesters.

Scenario 2. *The distribution of reputation is skewed, making it uninformative, despite the presence of actual differences in quality. Because of this uncertainty in identifying low and high quality workers, there is the potential for adverse selection.*

Here, despite there being differences between workers, the presence of a positivity bias leaves the reputation system unable to distinguish between high and low quality workers, as both see their work approved. Because there is uncertainty in the quality of workers based on reputation, other tools for identifying and sorting workers may be useful. For example, the Masters qualification could potentially provide information about quality when reputation is uninformative because it serves as a status marker. Therefore, it contains information about worker quality over and above an undifferentiated reputation (Washington & Zajac, 2005), because it is granted separately from the intertwined litigation and reputation mechanisms.

Scenario 3. *The distribution of reputation is skewed, but reputation system is not entirely uninformative about the differences in worker quality. However, the skew implies that information is contained only in the extremity of the scale, with differences in quality manifesting between the levels directly below the ceiling. Because there is information in the reputation system, adverse selection can be prevented.*

In this scenario, as in scenario 2, there are differences in worker quality that are masked to some extent by a positivity bias. However, there is information in the reputation system, but only at the extremity due to low incidences of rejections and negative feedback. Thus, negative reviews – even if only a handful – are more informative relative to positive reviews (Chevalier & Mayzlin, 2006; Khopkar, Li & Resnick, 2005; Standifird, 2001; Zervas et al., 2015). In this case, small differences in the ratings may be informative about worker quality, as is sometimes observed in reputation systems (Bridges & Vásquez, 2016; Dellarocas & Wood, 2008; Luca, 2011).

[Table 1 about here.]

Five studies are presented to demonstrate the necessary conditions for these scenarios. and to test them empirically. In studies 1A and 1B, the distribution of workers is characterized based on observable reputation, and the findings replicate those of prior work by demonstrating the highly skewed nature of the metric, while extending them by revealing its shape. In studies 2-4, each of the scenarios is tested, with evidence supporting scenarios 2 and 3, suggesting that there is sorting based on status markers and at the extremity of the reputation system, and that adverse selection can be an issue.

STUDY 1A

The purpose of the study 1A was to provide evidence for the potential skew in worker reputations on MTurk, by parameterizing this distribution. Though prior work has provided estimates of average approval rates on MTurk of approximately 99% (Schulze et al., 2013; Yin et al., 2016), these results are limited for two reasons: first, they rely on self-reports and thus may be subject to potential misrepresentation (Chandler & Paolacci, 2017), and second, they do not provide insight into the values that are actually observed by requesters on the MTurk platform. These values are the most relevant for researchers, since they are observable prior to data collection, and can be used to limit participation to workers who meet the requester’s criteria.

Method

Groups of 12 HITs were posted approximately simultaneously (within five seconds), with each group having the same task. The task was trivial, asking workers to respond to a single factual question whose answer could be searched for online, such as “What country won the 1990 FIFA World Cup?” Each HIT within the group was posted with different approval rate qualifications. One HIT was posted requiring rates below 90%, while the remaining 11 HITs required rates exactly 90% up to exactly 100%. The other qualifications were the same, requiring 100 completed HITs (ensuring that the approval rate value was valid) and that all participants were located in the United States. Participants were paid $.51 to complete the task.

One hundred assignments were available for each approval rate, and data collection for the group of HITs ceased as soon as any individual HIT finished. By examining the number of HITs completed for each approval level, estimates of the probability mass function approximating the distribution of approval rates among workers could be computed.

To ensure the reliability of the estimates, the procedure was repeated (with replacement – workers could complete a task in each group) six times using different questions. The first three groups were posted at the same time (1:00PM UTC) on consecutive weekdays. An issue when trying to estimate the distribution using a single time of day is the potential for differences in the worker cohorts, as the characteristics of MTurk workers vary by time of day (Casey, Chandler, Levine, Proctor & Strolovitch, 2017; Komarov, Reinecke & Gajos, 2013). Therefore, the last three groups were posted at six-hour intervals (7:00PM, 1:00AM, 7:00AM UTC) on the final day.

Results

Across all six samples, the first individual HIT to collect 100 responses was that requiring a 99% approval rate. The 99% approval rate condition served as the baseline, and responses in all groups that were submitted before the final submission in the 99% group were counted to construct the estimate of the distribution. The average time to completion took 37.7 minutes, and the average number of participants recruited in each sample was 125.2.

To test the reliability of the estimates derived from multiple independent samples, comparisons were made using the k-sample generalization of the Anderson-Darling rank test (Scholz & Stephens, 1987). This was appropriate because it makes no assumptions about the underlying distributions and can address the discrete nature of the observations. Comparisons of the three groups posted at the same time on consecutive days ($p = .354$), the four groups posted over 24 hours ($p = .062$), and the six groups overall ($p = .121$) all suggested that the data were drawn from an identical distribution. Therefore, an estimate of the probability mass function was created from the average proportions observed across all six samples (reported in table 2).

[Table 2 about here.]

[Figure 1 about here.]

The mean approval rate was 98.75% (SD = .96, skewness = -17.58, kurtosis = 25.03), and the overwhelming majority (80.6%) of MTurk workers had approval rates of 99%. Workers with 100% (8.5%) and 98% (7.2%) approval rates make up the next two largest groups. This is consistent with prior research employing worker self-reports that has observed 99% average approval rates (Schulze et al., 2013; Sharpe Wessling et al., 2017; Yin et al., 2016), but extends these findings by showing how these approval rates are distributed as observed by requester. Most importantly, this demonstrates that the distribution of reputation on MTurk is heavily skewed, with all but the entirety having an approval rate equal to or greater than 98%, similar to what would be expected due to a positivity bias.

STUDY 1B

To extend these findings by showing the relationship between what is observable to researchers and true worker characteristics, a survey was conducted. Four separate HITs containing three questions were posted simultaneously on a weekday at 1:00PM UTC, each recruiting 40 U.S. participants who had completed at least 100 HITs. Similarly to Study 1A, each HIT had a different approval rate requirement, of exactly 97%, 98%, 99% or 100%. In the task, participants were asked to enter the total number of HITs they had submitted, and the number of HITs they had worked on that had been approved and rejected as reported in their worker dashboard.

Results

The results are reported in table 2. To test for differences between the observed approval rates, the self-reported number of approved HITs was log transformed to address the skewed distribution. An ANOVA revealed significant differences between the observed approval rates and the transformed number of self-reported approved HITs ($F(3, 156) =$ 33.00, $p <$ .001). The total number of approved HITs among 99% workers ($M_{99\%} =$ $101,776$) was more than an order of magnitude higher than any other group ($M_{97\%} =$ 6890, $M_{98\%} =$ 5942, $M_{100\%} =$ 1041) . This broadens the picture of the earlier analysis, revealing that 99% workers are not only the most prevalent workers, but also the most productive.

In addition, these results reveal that approval rates as captured by the MTurk qualification (and therefore as observed by requesters) are calculated as the floor of the true approval rate. This means that a worker with a near-perfect approval rate – for example, one worker reported completing 207,683 HITs with only three rejections (for an actual approval rate greater than 99.998%) – is still given an approval rate qualification of 99%. Because even the most careful workers can incur rejections for capricious or duplicitous reasons (Horton, Rand & Zeckhauser, 2011), given a sufficient amount of experience all workers likely incur a rejection. This explains why workers with 100% approval rate qualifications were overall less experienced.

Discussion

The results indicate that most MTurk workers have approval rates of 99%, which is consistent with self-reports noted in prior work, as well as the distributions observed in other online reputation systems (Dellarocas & Wood, 2008). Furthermore, the considerable skew in the reputation scores is consistent with the theorized conditions under which a positivity bias would manifest.

Because of the way that the observed reputation score is calculated, as the number of tasks completed increases, the probability of transitioning from a 100% to a 99% approval rate approaches 1 because of the likelihood of incurring an inaccurate rejection. Combined with the finding that 99% approval rate workers are significantly more productive than at other observed approval rates, these results are consistent with the idea that the reputation system is effective in encouraging participation from presumed high quality workers, while discouraging those of lower quality (Resnick & Zeckhauser, 2002), which conforms to the hypothesized conditions of scenario 1. In study 2, the presence of moral hazard is tested to provide a stronger assessment of this possibility.

STUDY 2

The purpose of study 2 was to test conditions consistent with moral hazard. The assumption is that if all workers are of high quality, then in the presence of a positivity bias, workers may engage in shirking unless they are incentivized not to do so. A potential way to demonstrate this is to use repeated interactions in a panel design, because the awareness of future opportunities reduces incentives to defect (J.-Y. Kim, 1996). Thus, if workers believe their performance may affect their ability to work on a future task, they should be less likely to engage in behaviors that produce low quality responses. However, a key consideration of this approach is establishing awareness of this performance contingency, but not of the precise criteria (Sharpe Wessling et al., 2017). Therefore, two different recruitment procedures were used for the panel conditions. In the first, at the beginning of the initial interaction, participants were informed about a subsequent study and that invitations were contingent upon performance. In the second, this was not mentioned. Responses in these two conditions were compared against each other, as well as against a one-shot condition where the entire interaction would be completed in a single HIT. If shirking is present, this should manifest in higher quality responses when participants are aware of the possibility of completing future jobs.

Method

The HITs contained instructions and a link to an external survey hosted on Qualtrics. The survey was the same in all conditions, and incorporated five tasks. The first was the 18-item NFC scale (Cacioppo, Petty & Kao, 1984), which has been used in prior work investigating MTurk quality (Chandler et al., 2014) and is a scale workers would likely find familiar. The second was a scrambled sentence task with six scrambled sentences, each having between eight and 14 words. This task was used because it was relatively difficult, fatiguing, and could be scored objectively. Next, they completed a transcription for a 90-second video, which was approximately 215 words in length. This task was selected because transcriptions are common tasks for MTurk workers (Marge, Banerjee & Rudnicky, 2010), and the task is tedious and depleting. After completing the transcriptions, participants rated how interesting the video was, how easy the speaker was to understand,

and how enjoyable the task was, all using seven-point scales. Participants then completed a second set of six scrambled sentences, allowing for comparisons between the first and second sets. The final task was a 40-item version of the IPIP scale (Goldberg, 1992), selected in part because of its length and complexity, which would allow for analysis of its psychometric properties along multiple dimensions. The scales' length would also be likely to create a perception of “bubble hell" (Brawley & Pury, 2016; Sharpe Wessling et al., 2017). By placing it at the end of the survey, following several fatiguing tasks, participants may be more likely to engage in behaviors such as systematic or random responding, which would be observed in the quality of the measure.

Participants were recruited in three conditions. All workers were recruited through MTurk with approval ratings over 95% (thereby including more than 98% of all workers based on the results of study 1) and more than 1,000 HITs approved, and were paid a total of $1 for completing the entire survey. All of the conditions were launched at the same time (2:00PM UTC, corresponding to 10:00AM in the Eastern time zone of the United States) on weekdays. There was never overlap in the posting of the HITs, and participants were excluded from participating in more than one task by employing a JavaScript-based tool that prevented workers from viewing the link to the external study if they had already participated, with similar procedures used for the subsequent studies. The task was described as a “Consumer Psychology Survey,” and the description contained an estimate of the amount of time to complete the study.

The first group of MTurk workers ($N = 50$) completed the survey in one HIT (single HIT), which estimated the duration at 20 minutes, and workers were paid the full rate upon completion. The panel conditions were completed in two phases: a Time 1 survey including the NFC and first sentence-unscrambling task (with an estimated time of eight minutes), and a Time 2 survey including the transcription, second sentence-unscrambling task, and IPIP (estimated time of 12 minutes). Two panels (each $N = 100$) were recruited from MTurk to complete the T1 survey. In eligibility mentioned condition, participants were told before starting the study that “based on your performance in this task, you may be invited to complete a similar follow-up study," while no information was provided in the no eligibility condition.

In both versions, participants were asked at the end of the T1 survey to indicate their interest in a follow up study, and were paid $.25. One week later, the T2 survey was posted and participants who indicated interest at T1 were invited to complete it. A reminder notice was sent to participants who had not yet completed the T2 survey five days later. Participants who completed the T2 survey were paid $.75, making their compensation equal to the single HIT condition.

The panel design of the study also allowed testing of fatigue effects. Longer studies require higher levels of focus from respondents to complete the task (Cannell & Kahn, 1968), and MTurk workers are well known for engaging in other tasks while using MTurk, including cell phone use, TV watching (Clifford & Jerit, 2014) and other forms of multitasking (Chandler et al., 2014). A benefit of panel designs is that they can address this decline in quality over time in longer tasks (Goodman et al., 2013). Thus, participants who complete a long task in a single duration are likely to experience greater fatigue, leading to lower quality responses.

Results

Plan of Analysis. To test performance on the NFC scales, the quality of the data was examined on four dimensions: reliability, consistency and evidence of both systematic and random responding. Reliability was assessed using Cronbach’s alpha, and comparisons between these reliability coefficients were made using Fisher-Bonett tests (Bonett, 2003; S. Kim & Feldt, 2008). Consistency was assessed using a psychometric synonyms and antonyms approach (Johnson, 2005; Meade & Craig, 2012). Pairwise correlations across all participants for each of the individual items on the NFC scale were examined, and those pairs of items whose correlations exceeded the .60/-.60 threshold used in prior work (Meade & Craig, 2012) served as the focal items. Six pairs of items exhibited positive correlations above this threshold, and no pairs of items exhibited negative correlations below this threshold. The six pairs of items identified were used to form the psychometric synonym index, and similar procedures were used to construct these indexes in the subsequent studies. These indexes were formed by calculating the correlations among these pairs within individuals, with the intuition that consistent responders should respond in the same direction for psychometric synonyms and in opposite directions for psychometric antonyms. Thus, correlations for these items within a single participant's responses should be positive for psychometric synonyms and negative for psychometric antonyms. Inconsistent responding is reflected in values closer to zero.

The presence of systematic responding was examined using two procedures. The first was to test for straight line responding (Herzog & Bachman, 1981), where participants choose a single response category for all items using the same scale. This "long string" index was calculated as the total number of consecutive responses using the same category (Johnson, 2005; Meade & Craig, 2012). An index capturing central-tendency bias (Peer et al., 2014) was determined by the participant's frequency of choosing the mid-point of scales. As both of these indexes were count data, Poisson regressions with robust standard errors were used to model them (Cameron & Trivedi, 2013). Finally, to test for random responding, a procedure was adapted from Chandler et al. (2014), which consists of testing the relationship between the positive and reverse scored items in the NFC scale interacted with condition indicators. If participants' responding behavior differed in one condition, this would appear as an interaction effect in these regressions.

Performance on the scrambled sentence tasks was assessed by comparing responses to a key, excluding punctuation and capitalization, and the number of correct responses was analyzed using Poisson regression with robust standard errors. Two independent research assistants, blind to the hypotheses of the study, indicator coded the set of responses as invalid if any of the responses were left blank or were otherwise inappropriate (e.g. "i dont know"). The reliability of the coders for this measure was high (Set 1: Krippendorff's $\alpha = .808$, Set 2: Krippendorff's $\alpha = .977$), and disputes were resolved by the researcher. The invalid response measures were analyzed using logit models.

For the transcription task, due to the possibility for minor spelling and grammatical differences that would not materially affect quality, the same research assistants also coded the transcription as valid if it was a reasonable approximation (of similar length and containing the main elements) of the true script. Reliability for this measure was also high (Krippendorff's $\alpha = .945$). This measure was analyzed using logit models.

The analysis of the five IPIP subscales, each composed of eight items, was conducted using the same approach as for the NFC scale. There were 10 pairs of psychometric synonyms and four pairs of antonyms. To account for within-subject variance across the five IPIP subscales, regressions including subject fixed effects with robust standard errors were employed to analyze the relationship between positive and reverse scored items, again interacted with condition indicators.

Throughout the analysis, the nature of the comparisons necessitated a large number of tests. However, given the expectation based on prior work that most results would be null, a conservative approach of employing the standard alpha threshold of significance ($p < .05$) was employed in all studies. The full results are presented in Tables 3 and 4. For the sake of brevity, only significant differences will be discussed.

[Table 3 about here.]

Eligibility. To test for the potential effect of mentioning eligibility, the three conditions were compared on the T1 tasks. There were no differences in performance (all $ps > .388$), suggesting that mentioning eligibility did not affect T1 performance and the differences in motivation proposed by moral hazard are unlikely to be present. However, because this conclusion is based on a null result, additional investigation of fatigue was conducted to demonstrate the sensitivity of the design.

Comparing panel to non-panel conditions. For the eligibility-mentioned and no eligibility panel groups, invitations to complete the T2 survey were distributed to 91 and 88 participants who indicated a desire to complete the follow-up task, and of these, 42% and 48% completed the T2 task, respectively. There were little differences between the completers and attritors in terms of demographics (Completers: 51.3% female, Age = 35.9, English fluency = 6.93; attritors: 55.0% female, Age = 35.0, English fluency = 6.83) or performance on T1 tasks (all $ps > .202$). This suggests that panel attrition was not a factor.

Completers among the eligibility mentioned and no eligibility conditions were also compared, and there was a slight difference in reliability for the emotional stability subscale within the IPIP ($\alpha_{Eligibility} = .861$, $\alpha_{NoEligibility} = .931$; $\chi^2(1) = 4.12, p = .042$), but there were otherwise no differences in the 27 other tests (all $ps > .125$). Therefore, to maximize power, the two panel conditions were combined ($N = 80$) and compared to the single HIT condition ($N = 50$).

The total time spent for the panel conditions (sum of the time spent on the T1 and T2 tasks) was compared against that of the single HIT conditions. An ANOVA revealed no differences in time spent ($F(1, 128) = .00, p > .99$), suggesting the panel design did not alter the speed of task completion. The average time taken was 1225 seconds, well beyond the threshold where MTurk participants have exhibited lowered performance (Goodman et al., 2013).

[Table 4 about here.]

Task performance. Comparing the combined, completed panel conditions to the single HIT condition, there were no differences on any of the tasks prior to the discontinuity in the panel (all $ps > .259$). For the transcription task, there was a significant difference in interest in the content of the task, with panel participants finding the task content more interesting ($M = 4.74$) compared to single HIT participants ($M =$

$3.94, F(1, 128) = 5.69, p = .019$), suggestive of potential fatigue. However, none of the other performance or experience evaluation measures exhibited significant differences (all $ps > .109$). Analysis of the final tasks revealed some additional support for the presence of fatigue. In the second scrambled sentence task, a Poisson regression indicated that, relative to the single HIT condition ($M = 1.86$), the panels performed better on the objectively scored task ($M = 2.39, \beta = .25, z = 2.02, p = .043$). Further, a regression with the difference in correct sentences between the first and second tasks as the dependent measure indicated that panelists' performance over the two tasks improved relative to the single HIT condition ($\beta = .65, t(128) = 2.21, p = .029$). For the IPIP scale, there were no differences on any of the measures (all $ps > .085$).

Discussion

The results provide evidence that moral hazard was not present among respondents in the task, as indicated by the lack of differences between those who were informed of the performance contingency for the current tasks and those who were not aware. While this conclusion depends upon a null finding, the identification of fatigue effects suggests that this task was sensitive enough to demonstrate differences if present. This finding implies that the situation proposed in scenario 1, where a positivity bias leads to shirking by the cohort of high quality workers that is otherwise undifferentiated, is unlikely.

STUDY 3

The purpose of the third study was to examine the second scenario, where worker differentiation is present, but the reputation system is uninformative about these differences. The focus was on the two primary measures of the reputation system – HIT approval rates and productivity, along with a status marker in the Masters qualification. In addition, conditions were included to compare the effects of payment on performance.

The survey itself was a shortened version of that used in study 2, incorporating the NFC scale, the first six-item scrambled sentence task, and the IPIP scale. As before, all of the 10 conditions were launched at the same time of day (2:00PM UTC, corresponding to 10:00AM in the Eastern time zone of the United States) on weekdays, and every condition except one was completed in less than 24 hours. There was never overlap in the posting of the HITs, and the description contained an estimate of 8 minutes for the amount of time to complete the study. A summary of the conditions is provided in Table 5. In each condition, 50 participants were recruited (though in some conditions, an additional participant completed the task due to technical limitations in MTurk, De Langhe and Puntoni, 2016), with a total of 504 completed responses.

The first set of conditions tested the effects of HIT approval rates. Participants were recruited at approval rates greater than or equal to 75%, 95%, and 99% (comparing a threshold including all of the observed population to two of the typically suggested thresholds). All workers required to have at least 1,000 approved HITs, and were paid $.25.

The second set of conditions tested the effects of productivity, measured as the minimum total number of HITs approved, with the total number of approved HITs at 100 and 10,000 ($.25 payment, $\geq$ 95% approval), which were tested against the comparable 95% HIT from the first set of conditions.

The third set of conditions examined “Masters" workers. These participants were recruited with $.25 payments and 1,000 approved HITs, and tested against the equivalent non-Masters condition from the first set ($.25 payment, $\geq$ 99% approval, 1,000 approved HITs)

The final set of conditions tested the effects of payments on performance and completion of the task. Participants were recruited with payments of $.05, $.10, $.25, $.50 and $1.00 (100 approved HITs, 95% approval rates). The $.25 payment condition was the same as 100 total approved HIT used in the productivity set.

Results

Plan of Analysis. The analysis was conducted similarly to that of study 2. For the NFC scale, 30 psychometric synonym pairs and three antonym pairs were identified, while for the IPIP scale, there were 18 synonym pairs and three antonym pairs. The same two research assistants again coded for invalid responses (Krippendorff's $\alpha = .907$).

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

Approval Rate. Among the NFC psychometric antonyms ($F(2, 147) = 4.59$, $p = .012$), there were weaker correlations in the 99%+ condition ($M = -.62$) compared to the 75%+ ($M = -.83$, $F(1, 147) = 6.96, p = .009$), and 95%+ ($M = -.82$, $F(1, 147) = 6.81, p = .010$) conditions. For the NFC long string index, there was a higher level of systematic responding in the 95%+ condition ($M = 3.92$) compared to the 75%+ condition ($M = 3.06$, $z = 2.22$, $p = .026$). No other comparisons were significant (all $ps > .101$).

Productivity. No differences were observed between the productivity conditions on any of the measures (all $ps > .056$).

Masters Status. There were differences in NFC consistency, with higher correlations between psychometric synonyms ($M_{Masters} = .58$, $M_{Non-Masters} = .42$; $F(1, 98) = 4.51, p = .036$). There were also differences in systematic responding based on the long string ($\chi^2(1) = 4.45, p = .035$) and central tendency indexes ($\chi^2(1) = 8.19, p = .004$), with lower values for the Masters (long string, $M = 2.90$; central tendency, $M = 2.60$) compared to the non-Masters (long string, $M = 3.70$; central tendency, $M = 4.52$). There were also differences in consistency for the IPIP scale based on psychometric synonyms

($F(1, 98) = 7.03, p = .009$), with higher correlations for Masters ($M = .703$) compared to non-Masters ($M = .525$), and evidence for systematic responding on the IPIP scales, with higher values for the central tendency index for non-Masters ($M = 8.56$) compared to Masters ($M = 5.64$; $\chi^2(1) = 6.47, p = .011$).

Payment. For the payment conditions, differences were observed in the NFC central tendency index ($\chi^2(4) = 15.40, p = .004$), with higher rates of center scale item usage by the \$.10 condition ($M = 4.71$) compared to the \$.05 ($M = 2.84, z = -2.51, p = .012$) and \$1.00 ($M = 2.30, z = -3.68, p < .001$) conditions. On the IPIP scales, the psychometric synonyms measure exhibited some differences ($F(4, 248) = 2.83, p = .025$), with a higher average correlation for the \$1.00 condition ($M = .74$) compared to the \$.10 ($M = .58$, $F(1, 248) = 8.48, p = .004$) and \$.05 conditions ($M = .61, F(1, 248) = 6.37, p = .012$). There was also evidence for differences in systematic responding on the IPIP central tendency index ($\chi^2(4) = 24.24, p < .001$), with lower frequencies chosen by participants in the \$1.00 condition ($M = 3.16$) compared to the \$.25 ($M = 4.19, z = 2.09, p = .036$) and \$.10 conditions ($M = 4.82, z = 2.14, p = .032$).

Discussion

The results suggest that MTurk workers are largely undifferentiated by the two primary measures of the reputation system. For both productivity and approval rates, the few observed differences seem counter to the presumption of improved performance with higher reputation. This was also true for the different pay rates. While participants in the highest payment condition (\$1.00) did exhibit somewhat less systematic responding behavior, participants in the *lowest* payment condition (\$.05) also exhibited less systematic responding behavior, suggesting these differences may not be that meaningful.

The exception to this was in comparisons between Masters to non-Masters workers, where a consistent pattern emerged with the Masters producing higher quality work. However, excluding these comparisons, there were 97 planned statistical tests conducted. In those tests, five results were observed that were suggestive of differences between conditions, which is almost exactly what would be expected due to random chance (4.9). None of these results seemed indicative of a broader trend in quality differences based on reputation or pay rates. Because differences were observed when comparing Masters and non-Masters workers, the design was likely sensitive enough to detect the presence of quality differences. The fact that they were not provides compelling evidence in favor of scenario 2, suggesting that the reputation system provides little differentiation at levels recommended by previous research, and that there are differences in quality between workers that are encoded in status markers.

STUDY 4

The goal of study 4 was to examine the final scenario, which predicted that the skewed distribution was the result of a positivity bias, but that information was present in the extreme levels of the reputation system. Based on the results in study 1, the recommended reputation level threshold of 95% effectively includes more than 98% of the workers on MTurk, and this is only 9.4% more of the total worker population compared to the group of workers with approval rates of 99% or higher. Thus, it is reasonable to think heterogeneity in quality exists between the discrete levels of the reputation system above 95% approval rates that was unobserved in study 3.

An additional goal of study 4 was to address the issue of power. Post-hoc power calculations based on prior results examining data quality on MTurk (Peer et al., 2014) suggest that the samples used in study 3 were sufficiently powered to demonstrate any effects that were present. However, some of the tests used in this analysis have not been tested before, and these effect size estimates may not be accurate. Therefore, power calculations based on effect sizes in the comparisons of the Masters conditions were conducted, where significant differences were consistently observed. These calculations suggested total sample sizes of 207 would have power at least equal to .80 (Cohen, 1992) across all conditions.

Method

Two hundred ten workers were recruited in three HITs that were posted simultaneously, requiring exact observed approval rates of 98%, 99% and 100%, which served as the conditions in the study. The qualifications for each HIT were otherwise similar to those used in study 3, requiring US locations and at least 1000 HITs approved, and workers were paid $.50. The task used was same as in study 3, with three additional questions at the end asking participants to self-report their number of HITs submitted, approved and rejected.

Results

The analysis was conducted in the same manner as study 3. For the NFC scale, there were 31 psychometric synonym pairs and 14 antonym pairs, and for the IPIP scale, there were 22 synonym and 11 antonym pairs, and the same research assistants coded invalid responses (Krippendorff's $\alpha = .796$).

[Table 8 about here.]

For the NFC scale, the psychometric synonyms and antonyms indexes exhibited differences in consistency (Positive: $F(2, 207) = 3.91, p = .022$, negative: $F(2, 207) = 6.38, p = .002$). Contrasts showed that the correlations were lower for the 98% group ($M_{Positive} = .43$, $M_{Negative} = -.56$) compared to the 99% ($M_{Positive} = .60$, $F(1, 207) = 7.06, p = .009$; $M_{Negative} = -.75$, $F(1, 207) = 11.46, p < .001$) and 100% conditions ($M_{Positive} = .57$, $F(1, 207) = 4.33, p = .039$; $M_{Negative} = -.71$, $F(1, 207) = 7.17, p = .008$), while the 99% and 100% conditions did not differ (all $Fs < .50$, all $ps > .48$). There was also evidence for systematic responding, with higher values of the long string index in the 98% ($M = 3.77$) condition compared to the 100% condition ($M = 2.90$,

$z = -2.55, p = .011$), as well as higher values of the central tendency index in the 98% condition ($M = 4.40$) compared to the 99% ($M = 2.91$, $z = -2.20, p = .028$) and 100% ($M = 2.91$, $z = -2.42, p = .015$) conditions. On the scrambled sentence task, workers in the 98% condition solved fewer of the sentences ($M = 1.49$) compared to those in the 99% ($M = 2.23$, $z = 3.61, p < .001$) and 100% ($M = 2.19$, $z = 3.48, p = .001$) conditions.

For the IPIP scale, a significant difference was observed in the reliability of the agreeableness subscale ($\chi^2(2) = 6.45, p = .040$), with higher reliability in the 99% condition ($\alpha = .907$) compared to the 100% condition ($\alpha = .821$). The psychometric synonym and antonym measures exhibited differences in consistency (Positive: $F(2, 207) = 3.34, p = .037$, negative: $F(2, 207) = 6.47, p = .002$), with lower correlations in the 98% conditions $M_{Positive} = .53$, $M_{Negative} = -.52$) compared to the 99% ($M_{Positive} = .63$, $F(1, 207) = 4.73, p = .031$; $M_{Negative} = -.68$, $F(1, 207) = 10.22, p = .002$) and 100% conditions ($M_{Positive} = .64$, $F(1, 207) = 5.27, p = .023$; $M_{Negative} = -.67$, $F(1, 207) = 9.16, p = .003$). Higher values for the long string index for the 98% condition ($M = 5.27$) compared to the 100% condition ($M = 3.43$, $z = -2.50, p = .012$), and for the central tendency index for the 98% ($M = 9.20$) compared to the 99% condition ($M = 6.57$, $z = -2.03, p = .042$) suggested there was higher levels of systematic responding. There were no differences on any of the other measures (all $ps > .085$).

Analysis based on self-reported approval rates (calculated by dividing the reported number of approved HITs by the sum of approved and rejected HITs) largely replicated the findings based on observed approval rates, with higher self-reported approval rates associated with higher consistency and lower systematic responding on both of the scales[2], and higher numbers of correct responses to the scrambled sentence task. However, the only relationships observed for self-reported productivity (the sum of reported approved and rejected HITs, log-transformed) were for the psychometric antonyms meas-

[2] Reliability measures were excluded from these comparisons.

ures, with higher levels of productivity associated with stronger (i.e., more negative) correlations.

Discussion

Across all three of the tasks, a consistent pattern emerged to suggest that workers with observed approval rates of 98% exhibited lower quality compared to those with 99% and 100%. The stability of the results across multiple metrics makes this conclusion robust. This provides evidence supporting scenario 3, where a positivity bias leads to a skew in evaluations, but there is still information contained at high reputation levels, immediately below the ceiling. It is also notable that, while there were significant differences observed on dimensions such as systematic responding and the number of correct answers to the scrambled sentence task, there were no differences in the number of invalid responses. This suggests that, while the 98% approval rate workers may have been less engaged or less motivated to complete the tasks at a quality level comparable to the other groups, their work was not so flawed that it would be likely to be identified in the absence of a higher quality comparison group.

GENERAL DISCUSSION

The purpose of this research was to investigate how the merging of reputation and litigation mechanisms into a single tool, along with the presence of a positivity bias, can create conditions for a reputation system to fail to provide quality sorting. In five studies exploring this phenomenon on Amazon's Mechanical Turk, evidence was presented that the distribution of reputation is highly skewed in a manner consistent with a positivity bias, and that there is heterogeneity in quality among workers that is captured by status markers and at extreme values of reputation. While the findings did support the prediction that there can be adverse selection unless these are accounted for, no evidence was found for moral hazard, suggesting that workers on the platform were sufficiently motivated to limit shirking.

Theoretically, the results provide insights into the design of systems for ensuring cooperation. In the MTurk panel, the reputation and litigation functions are combined, and this alters the decision making calculus of buyers when evaluating work. To address the known biases in providing feedback in online reputation systems, prior work has suggested furnishing financial incentives (Ba, Whinston & Zhang, 2003; Zhou et al., 2008). Ostensibly, this appears to be what MTurk has done by making payments contingent on accepting work, which should incentivize requesters to evaluate it carefully. However, by tying this incentive directly to both an immediate financial cost for a worker as well as a long lasting reputational cost, the requester's decision is now subject to many other pressures that appear to systematically skew their evaluations positively. This appears to reduce the effectiveness of both functions as tools to ensure cooperation.

Practical implications

For researchers using MTurk to collect data, these results provide several insights. First, the recommendations of many methodological guides (Goodman & Paolacci, 2017; Peer et al., 2014; Sheehan & Pittman, 2016) and by MTurk itself for using the reputation system to ensure worker quality may not be reliable for this purpose. A 95% threshold includes 98% of all MTurk workers, and the results indicate that there are differences in quality among these workers. Employing a higher threshold, such as 99%, is more appropriate for this purpose. There is some evidence to suggest that ratings on online work platforms have inflated over time (Horton & Golden, 2015), and this may explain why these prior recommendations are no longer useful.

Based on the findings of study 3, it could also be argued that limiting participation to workers who have Masters status may also be an effective strategy. However, large scale use of Masters workers is hampered by several significant problems. First, because they are highly experienced, they are more likely to present issues of non-naïvetè about research paradigms (Chandler et al., 2014). Second, the number of Masters workers is quite small, which means that data collection is extremely slow: in study 3, the collection of only 50 participants took almost one hundred times longer than for most other conditions. Third, there is a significant price premium to employing Masters workers, and these latter two factors together hinder recruitment of a sufficient number of Masters workers for most studies.

The findings are further elucidated by considering them alongside theoretical work comparing the effectiveness of litigation and reputation mechanisms (Bakos & Dellarocas, 2011). When there is some degree of certainty about worker types, this work shows that depending upon the level of damages that can be assessed, it is possible that the presence of a reputation system may actually lower overall efficiency. If damages are set below an economically optimal point, then a reputation system puts additional pressure on sellers to produce high quality work. However, at or above a certain level of damages, when the marginal cost of effort from the seller is equal to the marginal benefit they receive from that effort, the reputation system has no impact on worker's decision making and thus it would be more efficient to rely solely on litigation.

Critically, this optimal level depends upon both the likelihood of identifying low quality work and the costs associated with pursuing the litigation action. In the context of MTurk and other online research pools, these factors are presumably related, while damages are fixed to the wage paid to workers. This implies that when the costs of investigating, rejecting and responding to worker retribution are relatively high, the reputation system may ensure the collection of high quality data. However, if these costs can be reduced – for example, through automated identification of low quality responses, using the litigation system alone without the reputation system may be successful. This also highlights the importance of requesters' role in maintaining the quality of the panel, as actions taken to discourage poor quality work provide benefits to the entire research community (Sharpe Wessling et al., 2017).

Finally, it is worth noting that, in spite of the issues identified in the reputation and litigation systems, the overall quality of the work provided was quite high, with low incidence rates for invalid responses and high reliabilities for the scale measures which were comparable to, if not better, than those obtained in other work investigating reliability on crowdsourced work platforms (Peer et al., 2017; Peer et al., 2014; Rouse, 2015). Thus, this research does generally support the findings of existing work validating the use of MTurk as a subject pool (Buhrmester et al., 2011; Chandler & Shapiro, 2016; Crump et al., 2013; Goodman & Paolacci, 2017; Mason & Watts, 2010).

Limitations

These findings should be considered in light of several significant limitations. First, the types of tasks employed in this research were such that there were few opportunities for deceptive responding. The responses to the scales do not have objectively correct answers, and could only be evaluated psychometrically and for evidence of responding patterns. Additionally, the scrambled sentence task items, while having objectively correct answers, did not have answers that could be searched for, and questions with easily accessible answers may lead workers to cheat (Goodman et al., 2013). Further, the questions were not factual characteristics about the participants, which might have suggested that future work depended upon a particular answer (Sharpe Wessling et al., 2017). Therefore, the tasks present a higher hurdle for observing quality differences, though potentially one closer to typical tasks employed in many types of research.

However, more objective criteria may be useful if researchers are interested in automating the identification of high quality workers and employing a pure litigation strategy. These measures can include time filters (Bardos, Friedenthal, Spiegelman & Williams, 2016) and attention check questions (ACQs) (Hauser & Schwarz, 2016). These criteria can easily be assessed automatically, and research has shown that ACQs can sort workers based on quality (Peer et al., 2014). However, such an approach can also systematically affect participants' responses (Hauser & Schwarz, 2016) by creating feelings of distrust (Mayo, Alfasi & Schwarz, 2014), and ACQs are commonly disclosed in worker discussion forums (Sharpe Wessling et al., 2017), which suggests they may also have other unintended outcomes. Further research comparing these approaches and these potential second-order effects would be useful in clarifying the strengths and weaknesses of pure reputation and pure litigation strategies.

It is also important to consider that the employment process on MTurk involves worker self-selection into tasks, and workers may also adjust their effort levels based on their assessment of the likelihood their work will be rejected. This potential issue could be exacerbated by the availability of information about the tasks outside of MTurk, in locations such as the aforementioned TurkOpticon, as well as online forums such as Turkernation (Goodman & Paolacci, 2017; Sheehan & Pittman, 2016), reddit's HITsWorthTurkingFor (Casey et al., 2017), and MTurk Crowd (Sharpe Wessling et al., 2017). Workers use these tools to gather information about HITs and requesters since MTurk does not provide this within the platform (Brawley & Pury, 2016; Sheehan & Pittman, 2016). These systems can impact data quality by presenting opportunities for workers to discuss tasks before they complete them (Chandler & Paolacci, 2017; Sharpe Wessling et al., 2017). Several of the tasks reported here were discussed, but these posts only shared a link and did not disclose any information about the nature of the task. The task in study 1A was discussed extensively, likely because it featured relatively high pay for low effort. Because the identification strategy depended upon constant arrival rates for workers, this could affect its validity, but all of these comments were posted after data collection was concluded. Nonetheless, the feedback loops created by the separate information systems of requesters and workers provide interesting avenues to explore the impacts of reputation changes on behavior from both parties.

CONCLUSION

Careful consideration is necessary when creating systems to ensure cooperation in marketplaces. The findings of this work suggest the design of MTurk's system introduces biases into evaluations, which affects the infromation it contains about worker quality. In practice, the presence of these biases means that most requesters should raise their approval rate thresholds to 99%, well above those recommended by most literature on MTurk and commonly employed by researchers. Further, it is important for researchers to be mindful of their obligations to help maintain the panel, by accounting for these biases.

ACKNOWLEDGMENTS

The author gratefully acknowledges the comments of Daniel Malter, Zachary G. Arens, Eunyoung Jang, Alejandra Rodriguez and Steven Shepherd, the associate editor and three anonymous reviewers on previous versions of this manuscript. The author also appreciates the feedback of April Bequette, who provided unique information and insights into the MTurk platform. However, all errors within are my own.

REFERENCES

Akerlof, G. A. (1970), "The market for 'Lemons': Quality uncertainty and the market mechanism", *The Quarterly Journal of Economics*, Vol. 84 No. 3, pp. 488–500.

Amazon. (2018), "Worker web site FAQs", available at: https://www.mturk.com/mturk/help?helpPage=worker#what_is_master_worker (accessed 30 June 2018).

Antonetti, P. & Maklan, S. (2017), "Concerned protesters: From compassion to retaliation", *European Journal of Marketing*, Vol. 51 No. 5/6, pp. 983–1010.

Azoulay, P., Stuart, T. & Wang, Y. (2013), "Matthew: Effect or fable?", *Management Science*, Vol. 60 No. 1, pp. 92–109.

Ba, S., Whinston, A. B. & Zhang, H. (2003), "Building trust in online auction markets through an economic incentive mechanism", *Decision Support Systems*, Vol. 35 No. 3, pp. 273–286.

Bakos, Y. & Dellarocas, C. (2011), "Cooperation without enforcement? A comparative analysis of litigation and online reputation as quality assurance mechanisms", *Management Science*, Vol. 57 No. 11, pp. 1944–1962.

Bardos, J., Friedenthal, J., Spiegelman, J. & Williams, Z. (2016), "Cloud based surveys to assess patient perceptions of health care: 1000 respondents in 3 days for US $300", *JMIR Research Protocols*, Vol. 5 No. 3, pp. e166.

Bebchuk, L. A. (1984), "Litigation and settlement under imperfect information", *The RAND Journal of Economics*, Vol. 15 No. 3, pp. 404–415.

Behrend, T. S., Sharek, D. J., Meade, A. W. & Wiebe, E. N. (2011), "The viability of crowdsourcing for survey research", *Behavior Research Methods*, Vol. 43 No. 3, pp. 800–813.

Berinsky, A. J., Huber, G. A. & Lenz, G. S. (2012), "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk", *Political Analysis*, Vol. 20 No. 3, pp. 351–368.

Bonett, D. G. (2003), "Sample size requirements for comparing two alpha coefficients", *Applied Psychological Measurement*, Vol. 27 No. 1, pp. 72–74.

Brawley, A. M. & Pury, C. L. S. (2016), "Work experiences on MTurk: Job satisfaction, turnover, and information sharing", *Computers in Human Behavior*, Vol. 54, pp. 531–546.

Bridges, J. & Vásquez, C. (2016), "If nearly all Airbnb reviews are positive, does that make them meaningless?", *Current Issues in Tourism*, Vol. forthcoming.

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*, Vol. 6 No. 1, pp. 3–5.

Cacioppo, J. T., Petty, R. E. & Kao, C. F. (1984), "The efficient assessment of need for cognition", *Journal of Personality Assessment*, Vol. 48 No. 3, pp. 306–307.

Cameron, A. C. & Trivedi, P. K. (2013), *Regression analysis of count data*, Cambridge, UK: Cambridge University Press,

Cannell, C. F. & Kahn, R. L. (1968), "Interviewing", In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology*, Reading, MA: Addison-Wesley.

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A. & Strolovitch, D. Z. (2017), "Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection", *SAGE Open*.

Chandler, J. J., Mueller, P. & Paolacci, G. (2014), "Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers", *Behavioral Research Methods*, Vol. 46 No. 1, pp. 112–130.

Chandler, J. J. & Paolacci, G. (2017), “Lie for a dime: When most prescreening responses are honest but most study participants are impostors”, *Social Psychological and Personality Science*, Vol. 8 No. 5, pp. 500–508.

Chandler, J. J. & Shapiro, D. (2016), “Conducting clinical research using crowdsourced convenience samples”, *Annual Review of Clinical Psychology*, Vol. 12, pp. 53–81.

Chevalier, J. A. & Mayzlin, D. (2006), “The effect of word of mouth on sales: Online book reviews”, *Journal of Marketing Research*, Vol. 43 No. 3, pp. 345–354.

Clemons, E. K. (2007), “An empirical investigation of third-party seller rating systems in e-commerce: The case of buySAFE”, *Journal of Management Information Systems*, Vol. 24 No. 2, pp. 43–71.

Clickhappier. (2016), “Masters qualification info - everything you need to know”, available at: http://www.mturkcrowd.com/threads/masters-qualification-info-everything-you-need-to-know.1453/ (accessed 25 July 2017).

Clifford, S. & Jerit, J. (2014), “Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies”, *Journal of Experimental Political Science*, Vol. 1 No. 2, pp. 120–131.

Cohen, J. (1992), “A power primer”, *Psychological Bulletin*, Vol. 112 No. 1, pp. 155.

Crump, M. J. C., McDonnell, J. V. & Gureckis, T. M. (2013), “Evaluating Amazon’s Mechanical Turk as a tool for experimental behavioral research”, *PLoS One*, Vol. 8 No. 3.

De Langhe, B. & Puntoni, S. (2016), “Productivity metrics and consumers’ misunderstanding of time savings”, *Journal of Marketing Research*, Vol. 53 No. 3, pp. 396–406.

Dellarocas, C. (2003), “The digitization of word of mouth: Promise and challenges of online feedback mechanisms”, *Management Science*, Vol. 49 No. 10, pp. 1407–1424.

Dellarocas, C. & Wood, C. A. (2008), “The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias”, *Management Science*, Vol. 54 No. 3, pp. 460–476.

Downs, J. S., Holbrook, M. B. & Peel, E. (2012), “Screening participants on Mechanical Turk: Techniques and justifications”. In Z. Gürhan-Canli, C. Otnes & R. (Zhu (Eds.), *Advances in Consumer Research* (Vol. 40, pp. 112–116). Duluth, MN: Association for Consumer Research.

East, R., Hammond, K. & Wright, M. (2007), “The relative incidence of positive and negative word of mouth: A multi-category study”, *International Journal of Research in Marketing*, Vol. 24 No. 2, pp. 175–184.

Eickhoff, C. & de Vries, A. P. (2013), “Increasing cheat robustness of crowdsourcing tasks”, *Information Retrieval*, Vol. 16 No. 2, pp. 121–137.

Fleischer, A., Mead, A. D. & Huang, J. (2015), “Inattentive responding in MTurk and other online samples”, *Industrial and Organizational Psychology*, Vol. 8 No. 2, pp. 196–202.

Fradkin, A., Grewal, E., Holtz, D. & Pearson, M. (2015), “Bias and reciprocity in online reviews: Evidence from field experiments on Airbnb”. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation* (pp. 641–641). Portland, OR: ACM.

Gao, G. G., Greenwood, B. N., Agarwal, R. & McCullough, J. S. (2015), “Vocal minority and silent majority: How do online ratings reflect population perceptions of quality”, *MIS Quarterly*, Vol. 39 No. 3, pp. 565–589.

Ghose, A., Ipeirotis, P. G. & Sundararajan, A. (2005), “Reputation premiums in electronic peer-to-peer markets: Analyzing textual feedback and network structure”. In *Proceedings of the 2005 ACM SIGCOMM workshop on Economics of peer-to-peer systems* (pp. 150–154). New York, NY: ACM.

Ghose, A., Ipeirotis, P. G. & Sundararajan, A. (2009), “The dimensions of reputation in electronic markets”, New York University, Working Paper CeDER-06-02.

Godes, D. & Silva, J. C. (2012), “Sequential and temporal dynamics of online opinion”, *Marketing Science*, Vol. 31 No. 3, pp. 448–473.

Goldberg, L. R. (1992), “The development of markers for the big-five factor structure”, *Psychological Assessment*, Vol. 4 No. 1, pp. 26–42.

Goldstein, D. G., Suri, S., McAfee, R. P., Ekstrand-Abueg, M. & Diaz, F. (2014), “The economic and cognitive costs of annoying display advertisements”, *Journal of Marketing Research*, Vol. 51 No. 6, pp. 742–752.

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*, Vol. 26 No. 3, pp. 213–224.

Goodman, J. K. & Paolacci, G. (2017), “Crowdsourcing consumer research”, *Journal of Consumer Research*, Vol. 44 No. 1, pp. 196–210.

Hauser, D. J. & Schwarz, N. (2016), “Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants”, *Behavior Research Methods*, Vol. 48 No. 1, pp. 400–407.

Herzog, A. R. & Bachman, J. G. (1981), “Effects of questionnaire length on response quality”, *Public Opinion Quarterly*, Vol. 45 No. 4, pp. 549–559.

Horton, J. J. & Chilton, L. B. (2010), “The labor economics of paid crowdsourcing”. In *Proceedings of the 11th ACM Conference on Electronic Commerce* (pp. 209–218). Cambridge, MA: ACM.

Horton, J. J. & Golden, J. (2015), “Reputation inflation: Evidence from an online labor market”, New York University, Working paper.

Horton, J. J., Rand, D. G. & Zeckhauser, R. J. (2011), “The online laboratory: Conducting experiments in a real labor market”, *Experimental Economics*, Vol. 14 No. 3, pp. 399–425.

Hu, N., Pavlou, P. A. & Zhang, J. (2006), "Can online reviews reveal a product's true quality?: Empirical findings and analytical modeling of online word-of-mouth communication". In *Proceedings of the 7th ACM conference on electronic commerce* (pp. 324–330). Ann Arbor, MI: ACM.

Hu, N., Zhang, J. & Pavlou, P. A. (2009), "Overcoming the j-shaped distribution of product reviews", *Communications of the ACM*, Vol. 52 No. 10, pp. 144–147.

Ipeirotis, P. G. (2010), "Mechanical Turk, low wages, and the market for lemons", available at: http://www.behind-the-enemy-lines.com/2010/07/mechanical-turk-low-wages-and-market.html (accessed 25 July 2017).

Irani, L. C. & Silberman, M. S. (2013), "Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 611–620). Paris, France: ACM.

Johnson, J. A. (2005), "Ascertaining the validity of individual protocols from web-based personality inventories", *Journal of Research in Personality*, Vol. 39 No. 1, pp. 103–129.

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*, Vol. 46 No. 1, pp. 141–155.

Khopkar, T., Li, X. & Resnick, P. (2005), "Self-selection, slipping, salvaging, slacking, and stoning: The impacts of negative feedback at eBay". In *Proceedings of the 6th ACM conference on electronic commerce* (pp. 223–231). New York, NY: ACM.

Kim, J.-Y. (1996), "Cheap talk and reputation in repeated pretrial negotiation", *The RAND Journal of Economics*, Vol. 27 No. 4, pp. 787–802.

Kim, S. & Feldt, L. S. (2008), "A comparison of tests for equality of two or more independent alpha coefficients", *Journal of Educational Measurement*, Vol. 45 No. 2, pp. 179–193.

Kokkodis, M. & Ipeirotis, P. G. (2015), "Reputation transferability in online labor markets", *Management Science*, Vol. 62 No. 6, pp. 1687–1706.

Komarov, S., Reinecke, K. & Gajos, K. Z. (2013), "Crowdsourcing performance evaluations of user interfaces". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 207–216). New York, NY: ACM.

Kreps, D. M. & Wilson, R. (1982), "Reputation and imperfect information", *Journal of Economic Theory*, Vol. 27 No. 2, pp. 253–279.

Kusterer, D., Bolton, G. & Mans, J. (2016), "Inflated reputations: Uncertainty, leniency and moral wiggle room in trader feedback systems", Cologne Graduate School, Working Paper Series 06-04.

Li, X. & Hitt, L. M. (2008), "Self-selection and information role of online product reviews", *Information Systems Research*, Vol. 19 No. 4, pp. 456–474.

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*, Vol. 47 No. 2, pp. 519–528.

Luca, M. (2011), "Reviews, reputation, and revenue: The case of Yelp.com", Harvard Business School, Working Paper 12-016.

MacLeod, W. B. (2007), "Reputations, relationships, and contract enforcement", *Journal of Economic Literature*, Vol. 45 No. 3, pp. 595–628.

Malter, D. (2014), "On the causality and cause of returns to organizational status: Evidence from the Grands Crus Classés of the Médoc", *Administrative Science Quarterly*, Vol. 59 No. 2, pp. 271–300.

Marge, M., Banerjee, S. & Rudnicky, A. I. (2010), "Using the Amazon Mechanical Turk for transcription of spoken language". In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5270–5273).

Mason, W. & Suri, S. (2012), "Conducting behavioral research on Amazon's Mechanical Turk", *Behavior Research Methods*, Vol. 44 No. 1, pp. 1–23.

Mason, W. & Watts, D. J. (2010), “Financial incentives and the performance of ‘Crowds’”, *ACM SIGKDD Explorations Newsletter*, Vol. 11 No. 2, pp. 100–108.

Mayo, R., Alfasi, D. & Schwarz, N. (2014), “Distrust and the positive test heuristic: Dispositional and situated social distrust improves performance on the Wason Rule Discovery Task”, *Journal of Experimental Psychology: General*, Vol. 143 No. 3, pp. 985.

Meade, A. W. & Craig, S. B. (2012), “Identifying careless responses in survey data”, *Psychological Methods*, Vol. 17 No. 3, pp. 437–455.

Moe, W. W. & Schweidel, D. A. (2012), “Online product opinions: Incidence, evaluation, and evolution”, *Marketing Science*, Vol. 31 No. 3, pp. 372–386.

Morriss, A. P. & Korosec, J. (2005), “Private dispute resolution in the card context: Structure, reputation, and incentives”, *Journal of Law, Economics and Policy*, Vol. 1 No. 2, pp. 473–496.

MTurk Blog. (2012), “Improving Quality with Qualifications–Tips for API Requesters”, available at: https://blog.mturk.com/improving-quality-with-qualifications-tips-for-api-requesters-87eff638f1d1 (accessed 25 July 2017).

Naderi, B., Polzehl, T., Wechsung, I., Köster, F. & Möller, S. (2015), “Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm”. In *Sixteenth Annual Conference of the International Speech Communication Association*, Dresden, Germany: Interspeech.

Nelson, P. (1970), “Information and consumer behavior”, *Journal of Political Economy*, Vol. 78 No. 2, pp. 311–329.

Paolacci, G., Chandler, J. J. & Ipeirotis, P. G. (2010), “Running experiments on Amazon Mechanical Turk”, *Judgment and Decision Making*, Vol. 5 No. 5, pp. 411–419.

Pavlou, P. A. & Gefen, D. (2004), “Building effective online marketplaces with institution-based trust”, *Information Systems Research*, Vol. 15 No. 1, pp. 37–59.

Pavlou, P. A., Liang, H. & Xue, Y. (2007), “Understanding and mitigating uncertainty in online exchange relationships: A principal-agent perspective”, *MIS Quarterly*, Vol. 31 No. 1, pp. 105–136.

Peer, E., Brandimarte, L., Samat, S. & Acquisti, A. (2017), “Beyond the Turk: Alternative platforms for crowdsourcing behavioral research”, *Journal of Experimental Social Psychology*, Vol. 70, pp. 153–163.

Peer, E., Vosgerau, J. & Acquisti, A. (2014), “Reputation as a sufficient condition for data quality on Amazon Mechanical Turk”, *Behavior Research Methods*, Vol. 46 No. 4, pp. 1023–1031.

Reichling, F. (2004), “Effects of reputation mechanisms on fraud prevention in eBay auctions”, Stanford University, Working Paper.

Resnick, P. & Zeckhauser, R. (2002), “Trust among strangers in internet transactions: Empirical analysis of eBay’s reputation system”, In M. R. Bay (Ed.), *Advances in Applied Microeconomics: The Economics of the Internet and E-commerce* (11, pp. 127–157), Amsterdam: Emerald Group Publishing Limited.

Rouse, S. V. (2015), “A reliability analysis of Mechanical Turk data”, *Computers in Human Behavior*, Vol. 43, pp. 304–307.

Scholz, F. W. & Stephens, M. A. (1987), “K-sample Anderson–Darling tests”, *Journal of the American Statistical Association*, Vol. 82 No. 399, pp. 918–924.

Schulze, T., Nordheimer, D. & Schader, M. (2013), “Worker perception of quality assurance mechanisms in crowdsourcing and human computation markets”. In *Proceedings of the 19th Americas Conference on Information Systems*, Chicago, IL: Association for Information Systems.

Sharpe Wessling, K., Huber, J. & Netzer, O. (2017), “MTurk character misrepresentation: Assessment and solutions”, *Journal of Consumer Research*, Vol. 44 No. 1, pp. 211–230.

Sheehan, K. & Pittman, M. (2016), *Amazon's Mechanical Turk for Academics: the HIT handbook for social science research*, Irvine, CA: Melvin & Leigh, Publishers.

Silberman, M. S. & Irani, L. (2016), "Operating an employer reputation system: Lessons from Turkopticon, 2008-2015", *Comparative Labor Law and Policy Journal*, Vol. 37 No. 3.

Silberman, M. S., Ross, J., Irani, L. & Tomlinson, B. (2010), "Sellers' problems in human computation markets". In *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 18–21). Washington, DC: ACM.

Simcoe, T. S. & Waguespack, D. M. (2011), "Status, quality, and attention: What's in a (missing) name?", *Management Science*, Vol. 57 No. 2, pp. 274–290.

Staffelbach, M., Sempolinski, P., Kijewski-Correa, T., Thain, D., Wei, D., Kareem, A. & Madey, G. (2015), "Lessons learned from crowdsourcing complex engineering tasks", *PLoS One*, Vol. 10 No. 9.

Standifird, S. S. (2001), "Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings", *Journal of Management*, Vol. 27 No. 3, pp. 279–295.

Suri, S., Goldstein, D. G. & Mason, W. A. (2011), "Honesty in an online labor market.", *Human Computation*, Vol. 11 No. 11.

Vakharia, D. & Lease, M. (2015), "Beyond Mechanical Turk: An analysis of paid crowd work platforms". In *Proceedings of the iConference*, Newport Beach, California: iSchools.

Washington, M. & Zajac, E. J. (2005), "Status evolution and competition: Theory and evidence", *Academy of Management Journal*, Vol. 48 No. 2, pp. 282–296.

Yin, M., Gray, M. L., Suri, S. & Vaughan, J. W. (2016), "The communication network within the crowd". In *Proceedings of the 25th International Conference on World Wide Web* (pp. 1293–1303). Montreal, Canada: International World Wide Web Conferences Steering Committee.

Zervas, G., Proserpio, D. & Byers, J. (2015), "A first look at online reputation on Airbnb, where every stay is above average", Boston University School of Management, Working Paper.

Zhou, M., Dresner, M. & Windle, R. J. (2008), "Online reputation systems: Design and strategic practices", *Decision Support Systems*, Vol. 44 No. 4, pp. 785–797.

Figure 1: Study 1A - Proportions of MTurk workers by researcher-observed approval rate

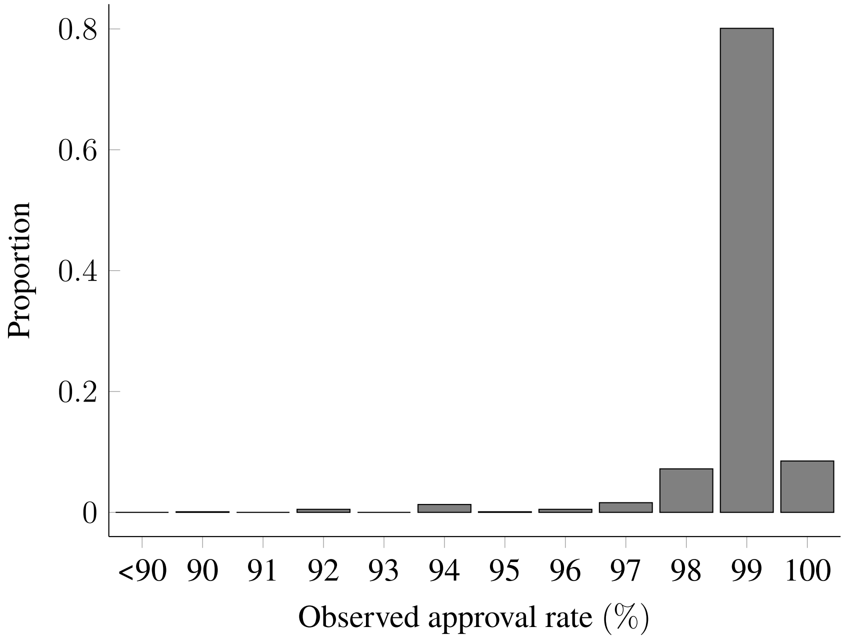

Table 1: Summary of scenarios, studies and findings

Scenario	Worker Differentiation	Reputation System Information	Study	Finding
1	No	No	2	Unsupported
2	Yes	No	3	Supported
3	Yes	Yes	4	Supported

Table 2: Study 1 - Proportions of MTurk workers and self-reported approval rates by researcher-observed approval rate

	Study 1A	Study 1B		
Observed Approval Rate	Proportion	Approved	Rejected	True Approval Rate
<90%	0.000			
90%	0.001			
91%	0.000			
92%	0.005			
93%	0.000			
94%	0.013			
95%	0.001			
96%	0.005			
97%	0.016	6890.0	169.4	0.9760
98%	0.072	5942.4	73.5	0.9878
99%	0.801	101,775.6	48.4	0.9995
100%	0.085	1040.8	0.0	1.0000

Table 3: Study 2 - Comparison of single and panel conditions on time 1 tasks

	T1		Need for Cognition Scale					Scrambled Sentence	
Condition	N	AvgT (Min)	α	SynR	LString	CenTend	+/- int.β	Correct	Invalid
Eligibility	99	6.10	0.956	0.559	4.59	4.07	Exc.	2.44	2
No Eligibility	101	5.92	0.955	0.569	4.85	4.11	0.040	2.26	4
Single HIT	50	20.39[†]	0.953	0.661	4.40	3.34	-0.002	2.38	0

AvgT = average time spent on task, SynR = psychometric synonyms correlation, AntR = psychometric antonyms correlation, LString = long string index, CenTend = central tendency index, +/- int. = interaction of condition and positive scaled items on negative scaled items, Correct = number of scrambled sentences correctly solved, Invalid = number of invalid responses to scrambled sentences, Exc.= parameter excluded, condition serves as baseline

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

[†] Average time for entire task

Table 4: Study 2 - Comparison of combined panel conditions to single HIT

	T1	T2			Need for Cognition Scale					Scrambled Sentence 1	
Condition	N	Invited	Completed (%)	AvgT (Min)	α	SynR	LString	CenTend	+/- int.β	Correct	Invalid
Single HIT	50			20.0	0.951	0.662	3.40	3.34	Exc.	2.38	0
Panel	200	179	80 (44.6%)	20.4	0.956	0.611	3.54	3.64	0.002	2.26	6

AvgT = average total time spent on task, SynR = psychometric synonyms correlation, LString = long string index, CenTend = central tendency index, +/- int. = interaction of condition and positive scaled items on negative scaled items, Correct = number of scrambled sentences correctly solved, Invalid = number of invalid responses to scrambled sentences, Exc.= parameter excluded, condition serves as baseline.

	Transcription					Scrambled Sentence 2	
Condition	Comp (%)	Avg Time (Min)	Interest	Understandable	Enjoy	Correct	Invalid
Single HIT	38 (76.0%)	7.50	3.94*	5.36	2.96	1.86*	4
Panel	61 (76.3%)	7.38	4.74*	5.56	3.53	2.39*	3

Comp (%) = percentage of completed transcriptions, Avg Time = average time spent on transcription task, Interest = interest in transcription task, Understandable: how understandable dialogue was in transcription task, Enjoy = enjoyment of transcription task, Correct = number of scrambled sentences correctly solved, Invalid = number of invalid responses to scrambled sentences.

	International Personality Item Pool Scales									
Condition	Extravaα	Agreeα	Conscα	Stabα	Imagα	SynR	AntR	LString	CenTend	+/- int.β
Single HIT	0.924	0.809	0.818	0.897	0.853	0.545	-0.652	4.86	8.84	Exc.
Panel	0.895	0.834	0.872	0.901	0.763	0.605	-0.707	4.44	8.43	-0.077

Subscales: Extra = extraversion, Agree = agreeableness, Consc = conscientiousness, Stab = stability, Imag = imagination; SynR = psychometric synonyms correlation, AntR = psychometric antonyms correlation, LString = long string index, CenTend = central tendency index, +/- int. = interaction of condition and positive scaled items on negative scaled items, Exc.= parameter excluded, condition serves as baseline.
$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Table 5: Study 3 - Summary of conditions

Pay ($)	Approval Rate	Number Approved	Masters	N	Total Time (Hours)
HIT approval rate conditions					
0.25	75%	1000	N	50	2.18
0.25[a]	95%	1000	N	50	1.63
0.25[b]	99%	1000	N	50	13.02
Total HITs approved conditions					
0.25[c]	95%	100	N	51	2.83
0.25[a]	95%	1000	N	50	1.63
0.25	95%	10,000	N	51	2.00
Masters conditions					
0.25[b]	99%	1000	N	50	13.02
0.25	99%	1000	Y	50	222.78
Payment conditions					
0.05	95%	100	N	50	53.73
0.10	95%	100	N	51	5.53
0.25[c]	95%	100	N	51	2.83
0.50	95%	100	N	50	1.32
1.00	95%	100	N	50	1.95

[a, b, c] Denotes conditions used to make comparisons among two sets of HITs.

Table 6: Study 3 - Comparison of approval rate thresholds, number approved, Masters status and payment - NFC and scrambled sentence tasks

		Need for Cognition Scale						Scrambled Sentence	
Condition	AvgT (Min)	α	SynR	AntR	LString	CenTend	+/- int.β	Correct	Invalid
HIT approval rate conditions									
75%	6.95	0.957	0.534	-0.826^{*}	3.06^{*}	3.40	Exc.	2.12	3
95%	8.10	0.949	0.521	-0.823^{*}	3.92^{*}	3.54	0.137	1.90	1
99%	7.16	0.943	0.415	-0.623^{*}	3.70	4.52	−0.128	2.52	2
Total HITs approved conditions									
100	9.03	0.949	0.514	−0.737	3.51	3.69^{**}	−0.120	2.51	1
1,000	8.10	0.949	0.521	−0.823	3.92	3.54^{**}	Exc.	1.90	1
10,000	6.62	0.964	0.559	−0.756	3.33	3.76	−0.130	2.18	1
Masters conditions									
Non-Masters	7.16	0.943	0.415^{*}	−0.623	3.70^{*}	4.52^{*}	Exc.	2.52	1
Masters	7.81	0.925	0.577^{*}	−0.794	2.90^{*}	2.60^{*}	−0.136	2.46	0
Payment conditions									
.05	8.92	0.948	0.516	−0.794	3.20	2.84^{**}	−0.012	2.55	3
.10	7.74	0.944	0.464	−0.704	3.75	$4.71^{**,***}$	0.023	1.96	0
.25	9.03	0.932	0.514	−0.737	3.51	3.69	Exc.	2.51	1
.50	7.46	0.959	0.547	−0.724	3.16	3.56	−0.047	2.06	2
1.00	7.54	0.951	0.596	−0.843	3.20	2.30^{***}	−0.027	2.48	0

AvgT = average time spent on task, SynR = psychometric synonyms correlation, AntR = psychometric antonyms correlation, LString = long string index, CenTend = central tendency index, +/- int. = interaction of condition and positive scaled items on negative scaled items, Correct = number of scrambled sentences correctly solved, Invalid = number of invalid responses to scrambled sentences, Exc.= parameter excluded, condition serves as baseline.

Table 7: Study 3 - Comparison of approval rate thresholds, number approved, Masters status and payment - IPIP results

	International Personality Item Pool Scales									
	Extra α	Agree α	Consc α	Stab α	Imag α	SynR	AntR	LString	CenTend	+/- int.β
HIT approval rate conditions										
75%	0.834	0.867	0.825	0.892	0.823	0.473	−0.495	3.86	8.20	Exc.
95%	0.853	0.828	0.808	0.919	0.817	0.617	−0.461	4.28	7.48	0.072
99%	0.900	0.835	0.813	0.902	0.804	0.525	−0.519	4.24	8.56	−0.021
Total HITs approved conditions										
100	0.896	0.847	0.847	0.877	0.786	0.652	−0.630	4.20	7.29	−0.081
1,000	0.853	0.828	0.808	0.919	0.817	0.617	−0.461	4.28	7.48	Exc.
10,000	0.994	0.890	0.838	0.891	0.826	0.596	−0.559	4.57	7.43	−0.079
Masters conditions										
Non-Masters	0.900	0.835	0.813	0.902	0.804	0.525^{**}	−0.519	4.24^{*}	8.56^{*}	Exc.
Masters	0.945	0.858	0.848	0.902	0.771	0.703^{**}	−0.577	3.24^{*}	5.64^{*}	−0.080
Payment conditions										
.05	0.888	0.827	0.866	0.904	0.838	0.606^{**}	−0.526	3.22	6.33	−0.048
.10	0.880	0.838	0.826	0.887	0.827	0.584^{***}	−0.534	4.82^{*}	10.76	−0.086
.25	0.896	0.847	0.847	0.877	0.786	0.652	−0.630	4.20^{*}	7.29	Exc.
.50	0.924	0.898	0.850	0.930	0.851	0.696	−0.606	3.40	7.92	0.001
1.00	0.925	0.918	0.864	0.923	0.791	$0.744^{**,***}$	−0.686	3.16^{*}	5.26	−0.002

Subscales: Extra = extraversion, Agree = agreeableness, Consc = conscientiousness, Stab = stability, Imag = imagination; SynR = psychometric synonyms correlation, AntR = psychometric antonyms correlation, LString = long string index, CenTend = central tendency index, +/- int. = interaction of condition and positive scaled items on negative scaled items, Exc.= parameter excluded, condition serves as baseline.

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

Table 8: Study 4 - Comparison of high approval rates

		Need for Cognition Scale						Scrambled Sentence	
Condition	AvgT (Min)	α	SynR	AntR	LString	CenTend	+/- int.β	Correct	Invalid
98%	9.77	0.945	$0.430^{*,**}$	$-0.560^{**,***}$	3.77^{**}	4.40^{*}	Exc.	$1.49^{**,***}$	4
99%	10.43	0.953	0.603^{**}	-0.749^{***}	3.11	2.91^{*}	0.025	2.23^{***}	1
100%	10.91	0.959	0.565^{*}	-0.709^{**}	2.90^{**}	2.91^{*}	-0.038	2.19^{**}	0

AvgT = average time spent on task, SynR = psychometric synonyms correlation, AntR = psychometric antonyms correlation, LString = long string index, CenTend = central tendency index, +/- int. = interaction of condition and positive scaled items on negative scaled items, Correct = number of scrambled sentences correctly solved, Invalid = number of invalid responses to scrambled sentences, Exc.= parameter excluded, condition serves as baseline.

	International Personality Item Pool Scales									
Condition	Extra α	Agree α	Consc α	Stab α	Imag α	SynR	AntR	LString	CenTend	+/- int.β
98%	0.884	0.869	0.875	0.871	0.849	0.534^{*}	-0.516^{**}	5.27^{*}	9.20^{*}	Exc.
99%	0.921	0.907^{*}	0.892	0.916	0.803	0.635^{*}	-0.682^{**}	3.91	6.57^{*}	-0.125
100%	0.921	0.821^{*}	0.868	0.915	0.889	0.640^{*}	-0.673^{**}	3.43^{*}	7.26	-0.118

Subscales: Extra = extraversion, Agree = agreeableness, Consc = conscientiousness, Stab = stability, Imag = imagination; SynR = psychometric synonyms correlation, AntR = psychometric antonyms correlation, LString = long string index, CenTend = central tendency index, +/- int. = interaction of condition and positive scaled items on negative scaled items, Exc.= parameter excluded, condition serves as baseline.

	Self-reported Number of HITs			
Condition	Submitted	Approved	Rejected	Approval Rate
98%	4396.3	4183.5	54.0	0.987
99%	32,003.2	31,910.4	55.3	0.998
100%	3580.9	3559.5	0	1.000

$^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$