Social Desirability Bias and Self-Reports of Motivation: A Study of Amazon Mechanical Turk in the US and India

Judd Antin
Yahoo! Research
4401 Great America Pkwy.
Santa Clara, CA 95054
jantin@yahoo-inc.com

Aaron Shaw
Department of Sociology
University of California, Berkeley
Berkeley, CA 94720
adshaw@berkeley.edu

ABSTRACT
In this study we extend research on online collaboration by examining motivation to do work on the crowdsoucing service Amazon Mechanical Turk (MTurk). We address a significant challenge to many existing studies of motivation in online contexts: they are based on survey self-reports, which are susceptible to effects such as social desirability bias. In addition we investigate a second challenge to the extant research on motivation in the context of MTurk: a failure to examine potential differences between MTurk workers (Turkers) from different parts of the world, especially those from the US and India, MTurk's two largest worker groups. Using a survey technique called the list experiment, we observe distinct profiles of motivation and patterns of social desirability effects among Turkers in the US and India. Among US Turkers, we find that social desirability encourages over-reporting of each of four motivating factors we examined. The over-reporting was particularly large in the case of money as a motivator. In contrast, among Turkers in India we find a more complex pattern of social desirability effects, with workers under-reporting "killing time" and "fun" as motivations, and drastically over-reporting "sense of purpose." We conclude by discussing these results and proposing implications for future research and design.

Author Keywords
motivation, social desirability, user-generated content, Amazon Mechanical Turk, distributed work

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Misc.

General Terms
Human Factors, Experimentation

INTRODUCTION

As online collaboration and user-generated content have become commonplace on the web, motivations for online participation have become an important area of study. Over more than 20 years, researchers have explored motivations for participating in a wide variety of online collaborative systems, such as open-source software [19] and Wikipedia [35]. This research has produced many insights that illuminate both theory and design. However, in this paper we highlight a specific weakness in much of this literature: a reliance on potentially biased self-reports of online motivations.

Many studies on motivation rely on survey-based techniques in which participants explicitly state their motivations. Surveys, however, are vulnerable to a variety of response effects [15] – among them *social desirability bias*, or the tendency to respond in ways that participants believe would appear desirable to others. Because of the lack of attention to potential social desirability bias in self-reports of online motivation, there is very little evidence about whether such biases exist or how they may influence existing results.

In this study we examine the potential for social desirability bias in the context of Amazon's Mechanical Turk (MTurk). Part online community and part online labor market, MTurk attracts workers (Turkers) from all parts of the world. However, the vast majority of visits to MTurk come from just two nations: the United States and India [2]. Although there is likely to be significant within-group variation, in this study we acknowledge the potentially significant contrasts between the two countries and treat them as distinct sub-populations. Theories from cultural psychology and related fields suggests that motivation may vary as a result of the specific socio-cultural, economic, and political environment in which individuals live and work. As a result, we examine both motivational profiles and patterns of social desirability separately for Turkers in the US and in India.

We employ a quasi-experimental survey technique called the list experiment which mitigates social desirability bias. Comparing the results of the list experiment to traditional agreement statement-style questions, we find significant evidence of social desirability bias among both US and India Turkers. Furthermore, our results suggest different patterns of both motivation and social desirability effects between the two groups. These findings contribute to the existing literature in two ways. First, we demonstrate that self-reports of motivation *can* be subject to social desirability bias. In doing so we highlight prior studies' failure to consider the possibility for bias, the need to investigate potential biases in existing findings about online motivation, and the potential for non-optimal design decisions based on inaccurate findings. Secondly, we reveal interesting dynamics around expectations and perceptions of desirable motivations for performing crowdsourcing work in two distinct participant populations. These dynamics provide further evidence that computer-mediated crowdwork cannot wash away differences between the participants in online collaborative systems, despite the tendency of some researchers and media reports to treat them as a single population, homogenized through the abstraction of distributed work.

MOTIVATIONS FOR ONLINE PARTICIPATION

Not long after computing and networking technologies began to develop in the 1960's, the ideologies and social structures of open collaboration around computing and code began to form [38]. Especially since the advent of open-source software, collaborative content, and social media, there has been significant interest in the question of what motivates people to participate in these systems. The literature on motivations spans a wide variety of socio-technical systems and practices – open source software [19], Wikipedia [35], blogging [26], photo-sharing [29], and question-answering [32], just to name a few.

Research across online contexts has consistently found that factors such as fun, a belief in ideologies of knowledge production, the desire for social connection, and knowledge development are important motivators for online participation. Most studies of online motivation have relied on self-reports, employing surveys with variations of traditional agreement statement-style questions. However, online surveys are susceptible to response effects that can skew results [15]. In this study we focus on one such response effect – *social desirability bias* – and document its effects on survey self-reports of motivation among participants in a specific crowdsourcing system.

SURVEYS & SOCIAL DESIRABILITY

Social desirability bias refers to "the tendency of people to deny socially undesirable traits or qualities and to admit to socially desirable ones" [30]. The theory of social desirability bias suggests it is primarily the result of two underlying social psychological processes. First, providing what individuals believe to be socially desirable responses is a form of impression management [8] — an effort to mold one's public image and to construct a favorable presentation of self based on expectations, norms, and beliefs about a given context. Secondly, providing socially desirable answers can be a form of self-deception [27]. In this respect, social desirability is often an attempt to deny one's "true" attitudes, or mask an underlying belief by expressing a contradictory one. Importantly, answers influenced by social desirability bias should not be considered as merely lies. Social desirability is often subtle, unconscious, and based on implicit attitudes which individuals are not aware of or able to express [28]. Social desirability bias in survey studies is a problem primarily because it can produce inaccurate results which misrepresent the "true" prevalence of attitudes and behaviors. In addition to biasing mean values, social desirability bias can also create spurious or suppressed correlations, mediations, moderations, or other statistical relationships [7]. Because of its potentially wide-ranging influences, social desirability bias should be a real concern, especially for research that relies on survey self-reports.

Researchers have found social desirability biases in survey responses on many controversial topics such as immigration, affirmative action, and racial prejudice.[1] However, the effects of social desirability are not limited to hot-button, divisive issues. For example, Adams and colleagues found that social desirability influenced self-reports of physical activity [1]. Several studies have documented pervasive biases in behavioral and attitudinal scales commonly used in organizational behavior research [25, 22]. A large meta-study also found evidence of widespread social desirability effects in scales used for marketing research [16].

Self-Reports of Online Motivation

Few studies have directly examined social desirability effects in self-reports of motivation, and to our knowledge none have done so in online contexts. As a result, little is known about whether social desirability can influence reports of motivations to participate online, and if it can how prevalent or strong the bias may be. Many online systems are designed on the basis of survey studies documenting users goals and motivations. If these studies misrepresent user motivations because of social desirability bias, the result may be non-optimal design decisions which reduce user satisfaction, present barriers to engagement and collaboration, and potentially disenfranchise users.

Survey self-reports of motivation are likely subject to social desirability bias for at least two reasons. First, there are often socially desirable connections between motivations and activities. For example, we may expect that an individual who makes a charitable donation should be motivated by altruism and look favorably upon such behavior. On the other hand, a man who volunteers primarily to meet women may be looked upon unfavorably. Social norms and stigmas can provide powerful cues about the motivations that should normatively be attached to a given activity. Secondly, one's motivation is often interpreted as a signal of other characteristics. For example, in the absence of other information, individuals may ascribe positive characteristics to the charitable altruist and negative characteristics to the date-seeking volunteer.

There is some evidence that, compared to face-to-face interviews, online surveys may be less susceptible to social desirability effects [15]. Computer-mediated interactions can introduce social distance that mitigates social desirability concerns in some contexts. However many users still actively engage in impression management in computer-mediated communication [14]. Furthermore, self-deception

[1] For a review of the literature on social desirability we recommend [5].

effects are likely to be relevant across interaction mediums. As a result, we contend that the presence and character of social desirability bias within online environments is long overdue for empirical investigation.

In sum, our first research question concerns a pervasive and crucial phenomenon that remains largely overlooked:

> **RQ1: Can self-reports of online motivation be subject to social desirability bias?**

MOTIVATION TO PARTICIPATE ON MTURK

The setting for our research is Amazon Mechanical Turk (MTurk).[2] MTurk is a service which allows "requesters" to distribute small chunks of work to thousands of workers around the globe. In exchange for a few minutes or seconds of a worker's effort on what Amazon calls Human Intelligence Tasks (HITs), Turkers earn small payments which typically range from a few cents to a few dollars.

Dedicated Turkers can earn a significant amount of money by completing HITs. Indeed, MTurk itself highlights the money-making possibilities by prominently inviting visitors to "Make money by working on HITs" on the service's homepage. According to an informal study, however, most Turkers earn less than $10 per week [10]. So, monetary incentives remain an essential but complex part of the motivational landscape of MTurk [21]. Especially for long-term Turkers there are also likely to be important non-monetary motivations.

Several studies have examined non-monetary motivations for doing work on MTurk, as well as the potential interactions between monetary and non-monetary motivations. Synthesizing a variety of previous studies on crowdsourcing markets, Kaufmann and colleagues [12] suggest a motivational model that includes "enjoyment-based motivations" (e.g. fun, passing time, interest), "community-based motivations" (e.g. social interaction, community identification), and "social motivations" (e.g. ideology, social approval). In the same study cited above, Ipeirotis [10] finds that sizable proportions of Turkers reported that fun, killing time, and the feeling that MTurk is a fruitful way to spend time were important motivators. Similarly, Buhrmester and colleagues suggest that intrinsic task enjoyment forms a substantive part of many Turkers' motivation [3]. Several recent studies have confirmed this suggestion, finding evidence that framing an MTurk task as benefiting a non-profit group [33] or as advancing scientific progress [4] interacts with monetary reward to motivate Turkers.

MTurk in the US and India

Our primary purpose in this study is to examine the potential for social desirability effects in self-reports of motivation on MTurk. Any serious attempt to do so must begin with an understanding of who Turkers are, and a rejection of the unsupported assumption that Turkers should be considered as a single, homogeneous group. MTurk attracts workers from all over the world, but the vast majority of visits to the site come from just two countries: the United States and India. As of December 2011, 40.8% of MTurk traffic was traced to IP addresses in the United States, while 35.7% of traffic was traced to addresses in India. [2]. In a study conducted in 2010, Ipeirotis [10] estimated that 46.8% of Turkers reside in the US, while 34% reside in India. A meta study by Ross and colleagues also documents what they describe as a trend moving from "a primarily moderate-income, US-based workforce towards an increasingly international group with a significant population of young, well-educated Indian workers" [34]. These results do not directly indicate the amount of work done by Turkers who hail from each country. However, they are a strong indicator that Turkers from the US and India make up the vast majority of the worker population.

The diversity of MTurk workers contrasts with popular images of crowd-sourcing. According to some media and academic portrayals, phenomena such as online labor markets, crowd-sourcing, and social search represent – at least theoretically – the fulfilment of techno-futuristic fantasies in which globalized, distributed human intelligence is rendered quickly and cheaply available through the click of a few buttons [39]. Reality, however, has not borne out these fantasies of commoditized brainpower. Rather than the seamless integration of person and machine, the global spread of computer-mediated communication has consistently revealed the difficulties of designing socio-technical systems in the face of the incredible diversity of human experience.[3] Empirical comparisons between different practices of human computation and the motivations of sub-populations of "human computers" can complement studies focused primarily on system design or evaluation [31].

Cross-cultural research in psychology[4] underscores the importance of cultural differences for conceptions of self, society, motivation, action, and emotion [20]. For our purposes, the most salient aspect of this research is the suggestion that individuals *both* understand the same action differently *and* explain their reasons for pursuing a particular action differently. Based on the extensive laboratory and field evidence supporting cross-cultural psychological differences, comparable patterns likely exist among distinct sub-populations of participants in online communities, crowd-sourcing platforms, and related phenomena.

Findings from studies comparing the experience of workers across national and cultural settings suggest that crowd-sourcing workers from different national groups might face unique pressures and employ unique self-presentation strategies in the context of their particularly atomized and alien-

[2] See `http://www.mturk.com.`

[3] A vast body of research has documented the challenges of designing socio-technical systems for users with diverse socio-cultural backgrounds. One useful high-level entry point is James Scott's work on "high modernist" utopias [36].

[4] We adopt the narrow usage of the term "cross cultural" to refer to nationality which is common in the field of cultural psychology. In the sociological or anthropological sense, nationality is considered to capture only a small part of an individual's socio-cultural identity.

ated workplace interactions. Ethnographies examining the motivations of IT sector workers in India, for example, depict a distinct environment in which such workers enact a complex set of identities. Business Process Outsourcing (BPO) workers in India encounter culturally specific forms of prejudice and respond in culturally specific ways [24]. We anticipate that crowd-sourcing workers in India may engage in similarly distinct behaviors when compared with workers in the US.

The potential for distinct patterns of motivation and social desirability bias among Turkers is even more likely in light of the fact that so much crowd-sourcing work – and especially the paid variety – is perceived in the United States as undesirable, undignified, or both [23]. Furthermore, the power dynamics of a paid labor market introduce an added dimension of inequality between researchers and their subjects: researchers surveying workers in the context of a paid crowd-sourcing platform are, if only briefly, their subjects' employers, creating an environment in which seemingly unobtrusive questions may prove sensitive. Given these circumstances, we anticipate finding social-desirability bias among both workers from the US and India, but that the character of the bias will vary between the two groups.

In addition to the theoretical arguments presented above, support for considering US and India Turkers as distinct sub-populations can be drawn from existing studies of what Turkers from the two countries actually do on the site. For example, research has illustrated that Turkers from different countries tend to produce different quality results participate at different rates and times, as well as with different attitudes and abilities that reflect their social and cultural context [37, 13, 34, 10]. Furthermore, the two populations provide distinct responses to survey questions about their motivations for working on MTurk, confront distinct work and employment conditions, and bring very different sets of skills and experiences to bear on their MTurk tasks [10, 13, 34, 37]. So, while dividing our analysis on the basis of country-of-origin certainly obscures within-group variation, and captures differences in socio-cultural, political, and economic factors only at a high level of abstraction, even this abstract classification is associated with significantly different patterns of participation.

In sum, there is ample support for considering the two sub-populations of Turkers separately could yield distinct patterns. Our second research question addresses this issue:

> **RQ2: Is there evidence of similar or distinct profiles of social desirability effects in reports of motivation across MTurk's two primary sub-populations?**

METHODS

Social desirability is a problem in survey self reports because participants must select specific answers to explicit questions. Researchers have primarily examined two general techniques for mitigating social desirability bias in experiments and surveys: (1) the use of additional survey instruments which measure individual participants' susceptibility to social desirability, and (2) survey and interview techniques which mitigate potential bias through indirect questioning, the use of so-called "proxy" subjects, and other creative techniques [27]. In this study we employ a type of indirect questioning called *the list experiment* (See, e.g. [18]). The list experiment asks each participant to report only *how many* he selects from a list of possible choices. Some participants will un-problematically select all of the potential choices (e.g. 4 out of 4) or none of them, but others will select 1, 2, or 3 out of 4. In the latter case, the list experiment provides an opportunity for the respondent to express support for his "true" attitudes without explicitly naming them: there is no way for anyone to determine exactly which items were chosen (if any). As a result of this freedom from the actual or implied judgment of others, responses to the list experiment are comparatively free of the pressure of others' actual or imagined judgment. A variety of studies comparing traditional survey questions to the list experiment have demonstrated the effects of social desirability bias and the list experiment's ability to mitigate it. For example, such comparisons revealed social desirability bias in responses to questions about racial attitudes in the American South [18], attitudes about immigrants and immigration [11], and the prevalence of vote buying in Nicaragua [9].

Condition Name	**N**	
	US	**India**
Agreement Statements	194	156
List Experiment Control	184	159
List Experiment "Sense of Purpose"	183	152
List Experiment "to Kill Time"	181	142
List Experiment "Fun"	179	152
List Experiment "Make Extra Money"	189	137

Table 1. Condition Descriptions and Sample Sizes

To assess motivations for doing HITs on MTurk, we asked survey questions using traditional agreement statements and the list experiment. First, we asked randomly selected participants to make a binary choice to agree or disagree with four motivation statements. We selected statements on the basis of a variety of prior survey research on motivations for doing work on MTurk. For a summary of this research, we suggest [12]. The four statements were: "I am motivated to do HITs on Mechanical Turk. . . " (1) "to kill time," (2) "to make extra money," (3) "for fun," and (4) "because it gives me a sense of purpose." List experiment participants were randomly assigned to one of five groups. In a control group participants were shown a list of all 4 motivations described above and given the following instructions: "How many of the following would you say is a motivation for you when you do HITs on Mechanical Turk? Please respond only with a number between 0 and 4." In the four treatment groups, participants were shown a list of only *three* motivations – each of the four motivations was absent in one group – and asked to respond with a number between 0 and 3. In each treatment condition, the mean response tells us how many of the 3 motivations participants selected. The mean in the control condition describes selection of the same three motivations plus a single additional motivation. Therefore, the difference in means between each treatment group and the control can be attributed to the addition of that single item to the list. For example, if participants in the control condition selected an average of 2.8 out of 4 items, while participants in a treatment condition selected 2.4 out of 3 items, all things being equal we can estimate that, in the absence of social desirability bias, 40% (i.e. 2.8 – 2.4) of participants would select the 4th item. Because of the indirect manner of calculating this value, the final result of the list experiment should be interpreted as an estimate or expected value for the proportion of people who would select a motivation in the absence of social desirability bias.

The general lack of social presence in the MTurk user interface makes it a conservative test for social desirability bias effects in online collaborative settings. Social desirability bias is magnified by the actual or perceived presence of others when answering questions. In the context of MTurk, however, there are few user-to-user interactions or signs of social presence. This reduced social presence in a computer-mediated context should make social desirability effects less likely [15]. If we observe social desirability effects among MTurk workers nonetheless, it will increase our confidence about observing them in other contexts with greater social presence.

Data Collection & Sample

A secondary benefit of our focus on MTurk is that it is possible to collect survey responses from Turkers using the service itself [17]. In July 2010 we recruited 1200 US-Based Turkers into our study, and in November 2010 we recruited 1216 India-based Turkers. Recruitment of specific or random users is not possible on MTurk, so our sample is composed of individuals who chose to respond to our request for survey responses. Participants were paid 5 cents for completing the survey, which took approximately 30 seconds to complete. Each user within each of the two sub-population samples was randomly assigned to one of six experimental conditions and restricted to participating only once.[5] In addition to questions about motivation, in each condition we asked users about their gender, age, education level, and frequency of completing HITs. Although random assignment should eliminate the potential influence of other individual differences between participants, we nonetheless followed the method used in [11], and included these co-variates in a series of OLS regression models to assess potential influences of demographic differences between conditions. No significant differences were found across conditions which suggests that, within the US and India samples, random assignment successfully mitigated the potential counfounding effects of demographic factors. Our analysis directly compares results across treatment (list experiment) and control (agreement-statement) groups *within* the samples of US and India Turkers.

We do not conduct hypothesis tests for significant differences between US and India Turkers because the list experiment provides only group-level (rather than individual-level) data. As a result it is not possible to compare the two samples. However, we do present the results from the two groups side-by-side and make descriptive observations between them.

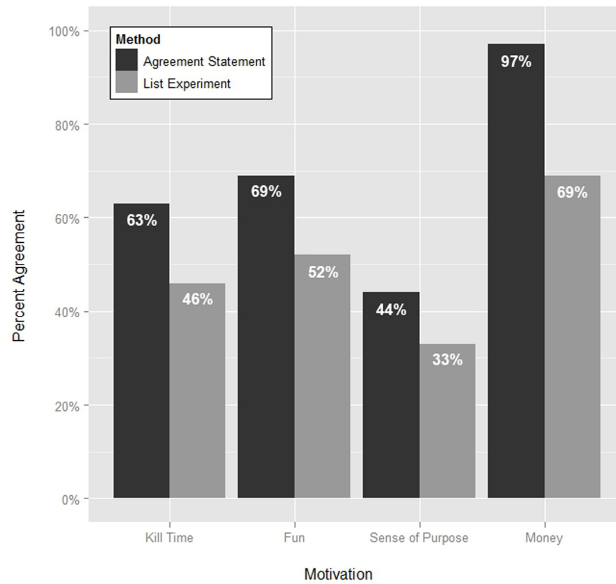

Figure 1. The proportion of US participants who selected each motivation using agreement statement-style questions or the list experiment. N = 1132.

RESULTS

Out of 1200 US responses, 68 (5.6%) were removed because they were invalid or incomplete, leaving a final valid sample of 1132. A higher proportion of responses in the India sample were invalid. Out of 1216 India responses, 318 (26%) were removed, leaving a final valid sample of 898. The higher rate of low quality work provided by India Turkers is consistent with prior research [13, 37]. Rates of invalid responses between conditions were not significantly different from random in either of the two samples (For the US sample, $\chi^2 = 5.84, p = 0.32$; For the India sample, $\chi^2 = 6.68, p = 0.25$). Invalid-response rates were also consistent across conditions for the combined samples $\chi^2 = 8.32, p = 0.14$). Table 1 illustrates sample sizes in each of the six conditions for both US and India sub-populations. Examining first the US sample, a comparison of the two survey methods showed that a smaller percentage

		Gender		Education	
	Age	Female	Male	College Degree	Graduate Degree
US	32.7(11.6)	60.7%	39.3%	51.6%	14.2%
India	29.3(8.9)	36.2%	63.8%	84.1%	53.6%

Table 2. Age, gender, and educational level among respondents in the US and India samples. Age is reported as mean with s.d. in parenthesis.

[5]To enforce condition assignment, country location restrictions, and the single-use restriction we employed a third-party service called CrowdFlower (`http://crowdflower.com`).

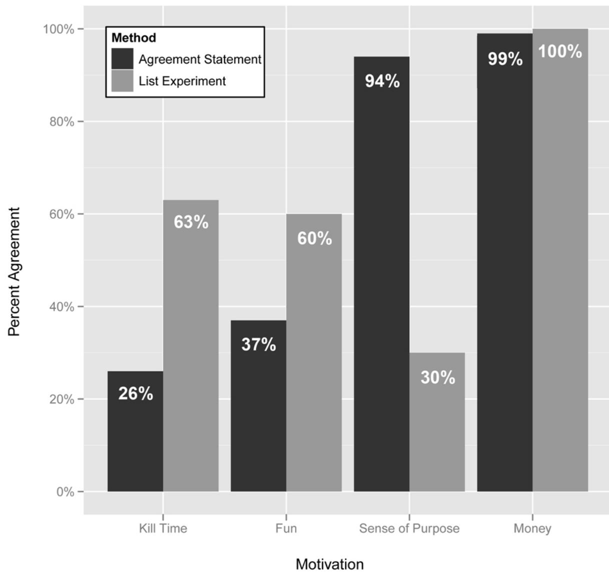

Figure 2. The proportion of India participants who selected each motivation using agreement statement-style questions or the list experiment. N = 898.

of respondents said they were motivated by each factor in the list experiment compared to the agreement statements. Figure 1 provides a visual representation of these results. Pearson's chi-square tests showed significantly *smaller* proportions among US participants in the list experiment for "killing time" ($\chi^2 = 10.19, p < .01$), "fun" ($\chi^2 = 10.89, p < .001$), and "money" ($\chi^2 = 53.82, p < .001$), and a moderately significant difference for "sense of purpose" ($\chi^2 = 3.47, p = .06$). The differences in proportions between the two methods were also quite large: 17% for both "killing time" and "fun," 11% for "sense of purpose," and 28% for "money."

The India sample showed a substantially different pattern. Figure 2 provides a visual representation of these results. Pearson's chi-square tests showed significantly *larger* proportions of agreement among India participants in the list experiment for "killing time" ($\chi^2 = 40.71, p < .001$) and "fun" ($\chi^2 = 16.04, p < .001$) and significantly *smaller* proportions in the list experiment for "sense of purpose" ($\chi^2 = 132.83, p < .001$). There was no significant difference between the two techniques for "money" ($\chi^2 = .38, p = .53$). Again, the differences in observed agreement between the two methods was large: 37% for "killing time," 23% for "fun," and a remarkable 64% for "sense of purpose."

Several aspects of our findings are notable. First, social desirability bias appears to encourage US Turkers to over-report the number of factors that motivate them. Participants in the agreement-statement condition from the US sample reported an average of 2.7 out of 4 motivations, while participants in the list experiment conditions reported only 2.1 out of 4 motivations on average. So, across conditions most US Turkers still reported being motivated by at least two factors, and each of four the motivations remained salient for a sizable proportion of our sample. However, accounting for the effects of social desirability, in each case that proportion was reduced.

We also found evidence of social desirability bias in our sample of Turkers in India, but the results do not suggest a uniform pattern. Participants in the agreement-statement condition over-reported "sense of purpose" and under-reported "killing time" and "fun." Results suggest no social desirability effects in reporting of money as a motivator. In contrast with the US sample, workers in the India sample reported approximately the same number of average motivations across the agreement-statement and list experiment conditions – 2.6 out of 4 compared with 2.5 out of 4 respectively. As with the US sample, all motivations were salient for at least some India participants, regardless of the measurement method used.

Examining MTurk's two key sub-populations separately allows us to observe potential differences in motivational profiles and social desirability biases between the two. The differences we observe are likely the result of a number of factors, including (but not limited to) gender, age, and class. However, building on the research discussed above, we argue it is likely that macro-scale socio-cultural, political, and economic factors, as captured broadly by nationality (India or US) also help to account for observed differences between the two groups. The specific causes, mechanisms, and implications of these patterns merit further investigation and comparison.

Figure 3 illustrates the degree to which the agreement statements over or under-reported agreement with each motivation compared to the theoreticaly more accurate value obtained by accounting for social desirability with the list experiment. Within the US sample, agreement statements modestly over-reported three of the four motivations compared to the list experiment: "killing time" (36% over-reported), "fun" (32% over-reported), and "sense of purpose" (33% over-reported). The magnitude of over-reporting was largest in the case of "money" as a motivation (40% over-reported). So, more than for other types of motivations, social desirability appears to encourage individuals to say that money is a motivator even though that may not be the case.

Turkers in India showed larger and more variable social desirability effects compared to US Turkers. Within the India sample, agreement statements under-reported 2 motivations — "killing time" (–142%) and "fun" (–62%), and dramatically over-reported "sense of purpose" by almost 200% compared to the list experiment. Results showed no social desirability bias related to money as a motivator, as effectively all participants in both conditions agreed that it was an important motivator.[6] So, in the case of Turkers in India, social desirability appears to encourage individuals

[6] Values for the list experiment conditions are estimates based on the tabulation of the responses across each permutation of answers (see the Methods section, above). The point estimate in this case was actually 103%, but since this is an artifact of the mathematical procedure, we report the maximum possible value of 100%.

to over-state the importance of "sense of purpose" as a motivator, and under-state the importance of both "killing time" and "fun."

DISCUSSION

The recent rise of computing metaphors such as the "cloud" and the "crowd" have, perhaps, encouraged a stereotype of the faceless and undifferentiated Internet user. Similar views are particularly prominent in the case of crowdsourcing work which itself tends to be comparatively featureless, repetitive, and unskilled. However, the robust evidence of diversity in opinions, perspectives, and backgrounds among users testifies to the failure of large-scale socio-technical systems to abstract out the foundations of online and offline life. In this study we turn a particular spotlight on the diversity of workers on MTurk, investigating whether Turkers from different countries might be influenced by distinct economic, socio-cultural, and political environments, and thus express different motivations and social desirability effects for doing work on the site.

Demographic differences between the samples of US and India Turkers likely contribute to the variations in reported motivation and social desirability bias we observe across the two groups. As we discussed above, while random assignment *within* each national group assures that demographic attributes were also distributed randomly across the list experiment and survey conditions, the nature of the MTurk platform and our study design made it impossible to ensure random distribution *across* the two groups. As a result, we cannot identify the precise cause(s) of the differential presence of social desirability bias across MTurk workers in the US versus those in India. However, while the nature of our experiment precludes explicit statistical comparison between national groups, treating them as distinct sub-populations does allow us to observe the different contours and characteristics of the data.

While social desirability did not alter the rank-order of motivations between the agreement-statement and list experiment conditions, the consistent over-reporting of motivation in the US sample is an important finding in itself.[7] This finding is suggestive of our participants' desire to appear highly motivated to others, as well as their belief that clearly articulating those motivations is socially desirable. The finding that at least some US Turkers over-state their motivation is particularly interesting given (largely unsubstantiated) media reports suggesting that many Turkers are individuals with lots of discretionary time (e.g. stay-at-home parents, security guards) For these Turkers and others, social desirability pressure may come from an expectation that behavior should be well-reasoned, from the tendency to ascribe positive characteristics to highly motivated individuals, and from the desire to express intrinsic interest in work (even rote side-work). Economic theories of rational choice as well as social psychological theories such as the theory of planned behavior rely on similar assumptions that human action is (or normatively should be) intentional and thoroughly considered. An individual is assumed to have established preferences, as well as known desired goals or outcomes for her behavior. While many scholars have argued that these normative models are incomplete and inaccurate, their prevalence in education and popular culture may have created social pressure among US Turkers to appear well-reasoned and highly motivated.

[7]The apparent proportional "consistency" of the bias does not diminish its importance. In the context of a multivariate analysis of these results, the inaccuracy of survey responses means of each motivation could produce spurious or suppressed correlations, mediations, moderations, and other statistical relationships. See [7].

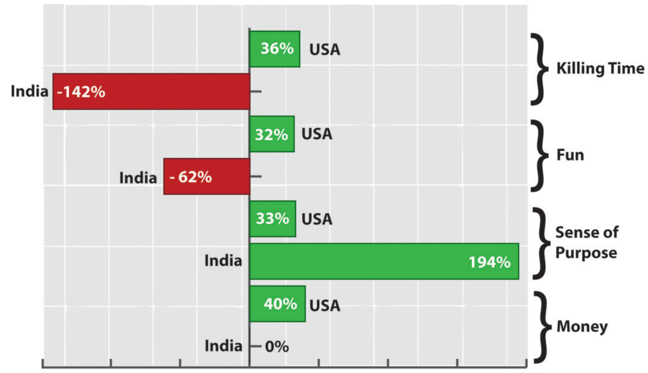

Figure 3. The degree of over or under-reporting of each motivation using traditional agreement statements, compared to the list experiment conditions. For example, compared to using the list experiment estimate as the "expected value," US participants over-reported that "killing time" was a motivation by 36%.

Again, our findings provide no direct information about why we should observe such distinct patterns of over- and under-reporting in the India sample, but previous research suggests potential explanations. Transnational IT work in India involves complex negotiations of identity as well as strategies of self-presentation that attempt to disguise experiences and values perceived to be inconsistent with foreign employers' or clients' desires [6, 24]. This may explain respondents' perception that "sense of purpose" is more desirable than "killing time" or "fun," even though the latter two appear to be important to many. Here again, our results may signify a form of self-deception in which some India participants attempt to mask their genuine motivations for doing work with the motivation they believe is socially "correct." The relatively high status sometimes accorded to BPO and call-center work in India may likewise contribute to the perception that workers who want to kill time or have fun are insufficiently serious and dedicated about their (desirable) jobs.[8] Requesters in the U.S. or elsewhere may ultimately find these sorts of motives inoffensive.

The Almighty Dollar (or Rupee)?

The over-reporting of monetary motivation among the US workers in contrast to workers in India raises particularly interesting questions in the context of MTurk. Upon its release in 2004, some reports painted MTurk as a "virtual sweatshop" [23], assuming that most HITs would be

[8]Note also the higher average education level of the workers in our India sample versus the US sample.

completed by workers in the developing world for whom even a few dollars could be a lot of money. However, according to web-tracking firm alexa.com, as of December 2011 nearly 50% of visits to the site were from users in the United States [2]. Our research builds on popular press reports [23] and prior studies [10, 21] that have illustrated the complex relationship between monetary and non-monetary motivations in crowdsourcing work. Our contribution is to provide strong evidence not only that both types of motivation are important, but that participants' motivations are often in tension with their perceptions about the motivations that others would look favorably upon.

In addition, our results indicate that money may be a motivator for fewer US Turkers than previously thought. The presence of an especially strong social desirability effect for reports of monetary motivation also suggests that US Turkers believe that money is the reason they *should* complete HITs. In the context of an online labor market like MTurk, workers shape their responses to conform with imagined expectations they attribute to job requesters (even researchers) or to the creators of MTurk who promote money-making as the primary reason to the use the service.[9] Alternatively, they may be influenced by the popular perception that crowdsourcing is an undesirable, "low class" form of labor [23], and likewise assume the popular belief that it would be undesirable to cast such labor as a fulfilling, fun, or personally satisfying way to spend one's time. However, given prior studies showing that task framing interacts with monetary motivation [33, 4], it is likely that the specific type of task, the beneficiary of the work, and other contextual factors would moderate these issues. Further research is necessary on this issue.

LIMITATIONS, IMPLICATIONS & CONCLUSION

Our study is subject to a number of important limitations. First, as we noted earlier, our data does not allow us to make structured statistical comparisons between US and India sub-populations. Our discussion is based on treating them as separate groups, and observing similarities and differences in the data. Demographic differences between the two samples probably account for at least part of the differences we observe; however, it is unlikely that gender or age alone (for example) could explain the differences. Furthermore, factors such as gender and age are inextricably tied to the socio-cultural, economic, and political foundations of crowd-work in India and the US. Macro-scale conditions also inform the types of people to whom crowd work is available and attractive.

Secondly, we note that our study is limited only to the unique socio-technical context of MTurk, and it would be inappropriate to generalize our findings beyond this context. Furthermore, our method is more descriptive than explanatory, which limits the ability of our findings to contribute to theories of social desirability in online motivation. Only through in-depth examination in mixed-methods studies will we learn how accurate our estimates are or how they may generalize to other crowd-sourcing and online collaborative systems.

We believe, however, that our findings strongly inform RQ1: reports of online motivation *can* be subject to social desirability bias. This finding demands that more attention be paid to the issue of social desirability in other contexts. Our findings also suggest that survey results in other online settings, especially those with greater social presence and a wider array of potentially sensitive topics at stake, could reflect some degree of social desirability bias in the reporting of online motivation. With respect to RQ2, the cross-national comparison indicates that the precise contours of social desirability depend on the context of interaction, the task at hand, and the populations involved. In a future study we intend to expand this methodology to search for social desirability bias in other online contexts such as open-source software and Wikipedia, and to include additional motivations that would be relevant in those contexts.

Many prior studies have used non-survey techniques such as behavioral/experimental tasks and qualitative interviewing to assess motivation. These methods, however, generally do not allow researchers to capture as large or as diverse a sample as surveys do. Despite this important benefit of surveys, our results cast a shadow of doubt on prior studies which have used direct questions to assess motivation in MTurk. For example, Ipeirotis' study [10] found that "very few Indian workers participate on MTurk for 'killing time'." Our results suggest this result is likely due to social desirability bias, and in fact that the majority of Turkers in India are at least partly motivated by "killing time." Likewise, accounting for the effects of social desirability, our results suggest that 60% of Turkers in India do HITs at least partly because they think it's fun, while only 20% reported fun as a motivation in Ipeirotis' study.

While it would be premature to suggest specific design implications on the basis of this work, dynamics such as the interaction of "fun," "killing time," and "sense of purpose" in India highlight the importance of identifying potential social desirability effects for design. For example, on the basis of a survey study Amazon might reasonably make decisions about how to message to potential Turkers and how to provide feedback about their participation. With inaccurate information similar to what was provided in our agreement statement conditions, designers might choose to highlight money-making and the meaningfulness of crowd-work at the expense of fun and killing time. In addition, believing that only a small portion of India Turkers are motivated by fun, designers might shun the integration of game-like features. However, such decisions might at best fail to capitalize on opportunities to enhance user satisfaction, and at worst alienate Turkers who perceive a motivational mismatch.

This speculative example illustrates the importance of this topic for the HCI community. Our results strongly suggest that future studies of motivation in online environments

[9] Half of the MTurk website's landing and login page is dedicated to workers, and the largest, boldest, and blue-est typeface is reserved for the phrase "**Make Money** by working on HITs" (See: http://mturk.com; original emphasis).

should take account of potential social desirability bias. The list experiment is just one method for doing so, and techniques for addressing social desirability bias and other response effects have been under development for at least 30 years [27]. In addition, the different patterns of motivation and bias across US and India Turkers underscore the importance of cross-national and cross-cultural studies of online motivation. As socio-technical systems become more global, designers must seek to facilitate social interaction and collaboration across national as well as cultural boundaries, strive to understand the differential foundations of motivation for participation, and consider how those motivations may influence divergent patterns of activity.

We also add further evidence that monetary incentives represent just one part of a complex motivational array behind paid work. Incentive systems that rely solely on paying users for content, ignoring other key motivations such as fun and providing a sense of purpose do so at their peril. Likewise. collaborative systems that seek to engage workers or collaborators from diverse backgrounds may benefit from context-specific adaptations to sub-populations with divergent attitudes about money or employment. While we used national boundaries as a proxy for macro-scale cultural differences, there are numerous sub-groups within both the US and India. As a result, studies of any popular socio-technical system such as Wikipedia, open-source software, or even Facebook would benefit from explicitly identifying and comparing important sub-populations and their perspectives on motivation and activity.

Finally, we provide another example case in which asking users to explicitly state their motivations can produce inaccurate and incomplete data [28]. Interviews, interactive games, and other behavioral data will provide complementary information that can be used to better understand the stated and unstated preferences of participants in online systems. A program of rigorous and creative research will be necessary to develop a thorough understanding of motivation in online contexts and inform the intelligent design of incentive systems.

Despite the immense popularity of crowdsourcing and other forms of online collaboration, it is likely that current systems are only beginning to realize their potential. Efforts to expand the reach of online collaboration to previously under-served and under-represented populations are already underway [13]. Understanding the path forward for designing effective and engaging systems of online collaboration — including crowd-sourcing systems — will depend on developing clear, accurate, and nuanced views of users and their motivations. As researchers and practitioners inform ongoing development with investigations of user motivation, it will be essential to consider social desirability and other response effects in order to maximize the transformative potential of online collaboration.

ACKNOWLEDGMENTS

We are grateful for the many anonymous reviewers whose commentary significantly improved this paper. An earlier version of this work also benefited from the feedback of discussants at a 2010 CSCW Horizons session.

REFERENCES

1. S. A. Adams, C. E. Matthews, C. B. Ebbeling, C. G. Moore, J. E. Cunningham, J. Fulton, and J. R. Hebert. The effect of social desirability and social approval on Self-Reports of physical activity. *American Journal of Epidemiology*, 161(4):389 –398, Feb. 2005.

2. Alexa.com. mturk.com - traffic details. http://alexa.com/siteinfo/mturk.com, Dec. 2011.

3. M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk. *Perspectives on Psychological Science*, 6(1):3 –5, Jan. 2011.

4. D. Chandler and A. Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *University of Chicago mimeo*, 2010.

5. T. J. DeMaio. Social desirability and survey measurement: A review. In C. F. Turner, E. Martin, and N. R. C. U. P. on Survey Measurement of Subjective Phenomena, editors, *Surveying subjective Phenomena*, volume 2, pages 257–282. Russell Sage Foundation, June 1984.

6. S. Derné. Cultural conceptions of human motivation and their significance for culture theory. In D. Crane, editor, *The Sociology of Culture: Emerging Theoretical Perspectives*, pages 267–287. Blackwell, Cambridge, MA, 1994.

7. D. C. Ganster, H. W. Hennessey, and F. Luthans. Social desirability response effects: Three alternative models. *The Academy of Management Journal*, 26(2):321–331, June 1983. ArticleType: research-article / Full publication date: Jun., 1983 / Copyright 1983 Academy of Management.

8. E. Goffman. *The presentation of self in everyday life*. Doubleday, NY, NY, 1973.

9. E. Gonzalez-Ocantos, C. K. de Jonge, C. Melndez, J. Osorio, and D. W. Nickerson. Vote buying and social desirability bias: Experimental evidence from nicaragua. *American Journal of Political Science*, Forthcoming.

10. P. Ipeirotis. The new demographics of mechanical turk. `http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html`, Mar. 2010.

11. A. L. Janus. The influence of social desirability pressures on expressed immigration attitudes*. *Social Science Quarterly*, 91(4):928–946, 2010.

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

13. S. Khanna, A. Ratan, J. Davis, and W. Thies. Evaluating and improving the usability of mechanical turk for low-income workers in india. In *Proceedings of the First ACM Symposium on Computing for Development*, ACM DEV '10, pages 12:1–12:10, New York, NY, USA, 2010. ACM.
14. S. Kiesler, J. Siegel, and T. W. McGuire. Social psychological aspects of computer-mediated communication. *American Psychologist*, 39(10):1123–1134, 1984.
15. S. Kiesler and L. S. Sproull. Response effects in the electronic survey. *Public Opinion Quarterly*, 50(3):402–413, Sept. 1986.
16. M. F. King and G. C. Bruner. Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing*, 17(2):79–103, Feb. 2000.
17. A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456, Florence, Italy, 2008. ACM.
18. J. H. Kuklinski, M. D. Cobb, and M. Gilens. Racial attitudes and the "new south". *The Journal of Politics*, 59(2):323–349, May 1997.
19. K. R. Lakhani, R. G. Wolf, J. Feller, B. Fitzgerald, S. Hissam, and K. R. Lakhani. Why hackers do what they do: Understanding motivation and effort in Free/Open source software projects. In *Perspectives on Free and open Source Software*. MIT Press, Cambridge, Mass, 2005.
20. H. R. Markus and S. Kitayama. Culture, self, and the reality of the social. *Psychological Inquiry*, 14:277–283, Oct. 2003.
21. W. Mason and D. J. Watts. Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.*, 11(2):100–108, 2009.
22. A. Meade, A. Watson, and C. Kroustalis. Assessing common methods bias in organizational research. In *22nd Annual Meeting of the Society for Industrial and Organizational Psychology, New York*, 2007.
23. K. Mieszkowski. "I make $1.45 a week and i love it". `http://www.salon.com/tech/feature/2006/07/24/turks/index_np.html`, Sept. 2007.
24. K. Mirchandani. Practices of global capital: Gaps, cracks and ironies in transnational call centres in india. *Global Networks*, 4(4):355–373, 2004.
25. R. H. Moorman and P. M. Podsakoff. A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behavior research. *Journal of Occupational and Organizational Psychology*, 65(2):131–149, 1992.
26. B. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz. Why we blog. *Communications of the ACM*, 47(12):41–46, 2004.
27. A. J. Nederhof. Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, 15(3):263–280, 1985.
28. R. Nisbett and T. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(1):231–259, 1977.
29. O. Nov, M. Naaman, and C. Ye. Motivational, structural and tenure factors that impact online community photo sharing. In *Third International Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA, 2009.
30. D. L. Phillips and K. J. Clancy. Some effects of "social desirability" in survey studies. *The American Journal of Sociology*, 77(5):921–940, Mar. 1972.
31. A. J. Quinn and B. B. Bederson. Human computation: A survey and taxonomy of a growing field. In *Proceedings of ACM CHI 2011 Conference on Human Factors in Computing Systems*, 2011.
32. S. Rafaeli, D. R. Raban, and G. Ravid. How social motivation enhances economic activity and incentives in the google answers knowledge sharing market. *International Journal of Knowledge and Learning*, 3(1):1 – 11, 2007.
33. J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets, 2011.
34. J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, pages 2863–2872, Atlanta, Georgia, USA, 2010. ACM.
35. J. Schroer and G. Hertel. Voluntary engagement in an open Web-Based encyclopedia: Wikipedians and why they do it. *Media Psychology*, 12(1):96, 2009.
36. J. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, New Haven, CT, 1998.
37. A. Shaw, J. Horton, and D. Chen. Designing incentives for inexpert raters. In *The 2011 ACM Conference on Computer Supported Cooperative Work (CSCW, 2011)*, Hangzhou, China, 2011.
38. S. Weber. *The Success of Open Source*. Harvard University Press, Apr. 2004.
39. J. Zittrain. Ubiquitous human computing. *Philosophical Transactions of the Royal Society A*, 366:3813–3821, 2008.