Experiences Surveying the Crowd: Reflections on Methods, Participation, and Reliability

Catherine C. Marshall
Microsoft Research, Silicon Valley
1065 La Avenida
Mountain View, CA 94043
cathymar@microsoft.com

Frank M. Shipman
Department of Computer Science
Texas A&M University
College Station, TX 77843-3112
shipman@cs.tamu.edu

ABSTRACT
Crowdsourcing services such as Amazon's Mechanical Turk (MTurk) provide new venues for recruiting participants and conducting studies; hundreds of surveys may be offered to workers at any given time. We reflect on the results of six related studies we performed on MTurk over a two year period. The studies used a combination of open-ended questions and structured hypothetical statements about story-like scenarios to engage the efforts of 1252 participants. We describe the method used in the studies and reflect on what we have learned about identified best practices. We analyze the aggregated data to profile the types of Turkers who take surveys and examine how the characteristics of the surveys may influence data reliability. The results point to the value of participant engagement, identify potential changes in MTurk as a study venue, and highlight how communication among Turkers influences the data that researchers collect.

Author Keywords
Crowdsourcing; surveys; demographics; reliability.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI).

General Terms
Human Factors; Experimentation; Measurement.

INTRODUCTION
Lately human computation platforms such as Amazon's Mechanical Turk have been used as a venue for performing many types of user studies [17, 30]. Inexpensive, educated, and relatively reliable (especially as researchers identify best practices for using these platforms [2,6,15,17]), Turkers provide us with a convenient and diverse pool of prospective study participants [17,24]. Yet the use of Turkers as study participants is not without controversy: On one hand, in addition to being fairly diverse, Turkers appear to be patient, thoughtful, and committed to the work they accept; on the other hand, they may become tainted and cynical from performing a steady regimen of paid surveys.

WebSci'13, May 2–4, 2013, Paris, France.

In this paper, we describe six related studies about the ownership and control of online media that we performed using the Mechanical Turk platform. For each study, we recruited self-reported users of a different media type or genre, including tweets, photos, online reviews, recorded videoconferences, podcasts, and educational videos. Overall we collected data from 1377 prospective participants over the course of two years. Using stringent filtering standards, we retained data from 1252 of them for further analysis.

This paper aggregates and compares the data across studies to explore two overarching themes: (1) what we learned about the best practices as we applied and extended them and (2) what we discovered about the workers. To develop the first theme, we document methodological details and reflect on our experiences using this method; to develop the second, we aggregate and compare the data that was common across the studies. Our contributions thus fall under three rubrics: method; participants; and data reliability.

1) ***Method.*** To provide a backdrop, we describe the study strategy and reflect briefly on which aspects of it worked and which aspects fell short.
2) ***Participants.*** We use the data we collected to compare the characteristics of the Turkers who participated in the studies. How diverse are they? What do they do online besides crowdwork? How do they compare with participants in other studies with different recruiting requirements and effort levels?
3) ***Reliability.*** Reported experiences with Mechanical Turk vary [14, 24, 30]. How reliable is our data? How does reliability vary across surveys?

First we present the study method—how the surveys were structured and administered—and the data we collected. We go on to describe participant characteristics and assess data reliability by further analysis of the aggregate results. Finally we reflect on higher-level questions about the efficacy of this sort of study, and issues that may arise.

METHOD
The studies were designed to elicit respondents' attitudes toward the ownership, control, and reuse of digital content [4,12,18], particularly as it influences the creation and use of personal and institutional archives [10,26]. The six questionnaire-based studies were administered separately;

each was initially offered to workers during a two-week period. The first three (addressing tweets, personal photos, and book reviews) were separated in time—we analyzed the data and wrote up the results before embarking on the design of the next study—and the last three (covering educational recordings, recorded videoconferences, and podcasts) were performed more or less concurrently. The final three were set in motion over the year-end holidays, and had sparser participation. To compensate for this, we redeployed them serially four months later. The second dataset for each of these three studies was aggregated with the first datasets after we established their statistical indistinguishablity.

In this section, we describe the studies' design, their relative characteristics, and the data we collected from each. We include a short reflection on participant engagement, something we feel is important for crowd-sourced studies. We discuss methodological limitations later in the paper.

Study Design and Characteristics

All six surveys had comparable structures. At the outset, we asked between 9 and 12 demographic and background questions, including gender, birth decade, whether the participant is currently a student, how much education the participant has completed, native language, and how long the participant has been using the Internet. Participants were also asked about their Internet activities (first through checkboxes, then with an open-ended question) and what they publish on the Internet (again with an open-ended question). Three added questions addressed trends in the participant's social media use.

The second part of the survey presented scenarios, followed by related sets of Likert-scale statements designed to explore respondents' reactions to storing, sharing, publishing, and removing content. Interspersed were 2-3 reading comprehension questions to check data quality [17]. The scenarios and associated hypothetical statements (*what-ifs*) are borrowed from legal theory, where hypotheticals are used to help legal scholars explain doctrine and explore the moral underpinnings and consequences of legal rules [19]. Hypotheticals generally establish a fact pattern, then vary it one component at a time to pursue limits [28]. This use of hypotheticals is a salient feature of the method, since we are exploring questions aimed at eliciting emerging ethical norms and the reasoning that goes into their production.

Finally, we used a mix of open-ended and multiple choice questions to find out more about participants: what they do and their attitudes toward media reuse and institutional archiving. These questions ranged from the specific, e.g. "Describe the last online review you wrote" to the abstract, e.g. "Should the Library of Congress be able to archive social media?" As Kiesler and Sproull remind us, electronic surveys tend to elicit more self-absorbed and uninhibited responses than their paper predecessors. [16]

Table 1 compares survey characteristics; they are listed in deployment order. As the table shows, the responses to the first survey reassured us; we extended the scenarios and added more open-ended questions.

What Worked and What Didn't

After completing six surveys on Mechanical Turk, we can reflect on the efficacy of their structure. We discuss each portion, focusing on what we learned, with an eye toward anything that contradicts prevailing wisdom. We omit discussion of standard demographic questions; the only aspect of these questions to note is that the responses seemed consistent with open-ended questions that covered the same ground. For example, if respondents said they were students, schoolwork often came up in narrative responses.

Scenarios. We developed detailed story-like scenarios that we based on real situations we had observed, followed by hypothetical statements about the story's characters and their actions. As in law, each hypothetical uses roughly the same fact pattern as the last, with one fact varying to test a single concept; hypotheticals with similar fact patterns could then be compared. For example in the photo study, after we showed participants a reference photo and told them a story about it, we posed a hypothetical: *Janice*, (the photo's subject) *should be able to post the photo to Facebook*. Then we varied the hypothetical in two ways: *Fred* (the photographer) *should be able to post the photo to Facebook* and *Janice should be able to post the photo to her public Flickr account*; these tested a distinction between

Survey	Participant ID	Premise of main scenarios	# Likert-scale	# demo/ practice	# reading comp.	# open-ended	Total
Twitter	TW###	User collects and reposts humorous and embarrassing tweets in different venues.	16	12	3	3	34
Photos	PH###	Photographer takes photo of two friends at a 25th birthday bash at a nightclub. Several women are visible in the background. The photo is reused by photographer, subject, woman in the background, and venue promoter.	18	14	3	6	41
Reviews	RE###	A kid's review and an educator's review of children's classic *Where the Wild Things Are* elicit comments on Amazon and are used by different parties for different purposes.	28	14	3	5	50
Podcasts	PC###	Longtime friends record a comedy podcast with a guest. An engineer edits and posts the podcast. A musical guest records a parody song. Different people reuse clips.	22	14	2	4	42
Recorded video-conferences	VC###	Recorded online job interview is repurposed in a variety of ways including instruction, satire, and a blog rant about the company performing the interview.	20	13	2	5	40
Educational recordings	ED###	Astronaut Sally Ride records university commencement address for educational service. The recording elicits comments and a recorded response from a scientific peer.	25	14	2	4	45

Table 1. Brief descriptions of the six surveys, listed in deployment order

the rights of the photographer and the subject, and between publishing in a semi-private venue and a public one.

Although the scenarios and hypotheticals exploited distinctions we thought were interesting, it's not clear that participants, who were working quickly, always understood the finer points of these distinctions, especially after they had gone through a number of hypotheticals. This potential for burn-out led us to cap the number of hypotheticals. Triangulation of participant responses to the hypotheticals was important to ensure response integrity.

Open-ended questions about practice. The open-ended questions that worked best were either concrete or fully aspirational. As established qualitative research methods suggest [7,35], drawing on recent incidents is effective; respondents answered these questions with extra details and seeming candor, although we recognize the potential for social desirability bias, especially as we ask about ethical matters [3]. Requiring respondents to refer to external facts about themselves (even concrete ones) did not work as well, especially if the respondents felt their study qualifications were being questioned or their privacy was threatened.

For example, asking respondents to describe the last review they had written or the last picture they had found online and reused were effective in eliciting detailed self-reports. PH032 answered the photo reuse question: "*Earlier this week I downloaded a picture of a dog wearing a party hat for a story I was doing on my pet blog about an event coming up. It was a photo included in a press release by the store holding the event.*" RE034 answered the review question that he had written a "*Quick, casual review of the new Thor movie. Liked the movie, like reviewing things, disagreed with some things I'd read.*"

On the other hand, asking Twitter survey respondents how many followers they had was problematic; even compliant participants seemed to be guessing (as evidenced by a preponderance of round numbers). Although these answers may have been accurate, that they even seemed suspect throws the question's effectiveness into doubt. Guessing, estimating, or making something up allowed respondents to answer the question without bringing up a Twitter client to check; they also may have felt compelled to demonstrate compliance with the survey's requirement that they be experienced Twitter users by inflating their numbers.

Response length to open-ended questions is one indication of participant engagement. Figure 1 shows the response length (in number of words) to the question about Internet activities; the y axis shows how often a response of this length occurred. Although the average response length varied slightly between the surveys, the relative values were consistent. For example, the average response to the question about Internet activities was about 16 words long for the podcast survey and about 19 words for the photo survey. Thus, most answers are brief, but longer answers are not uncommon.

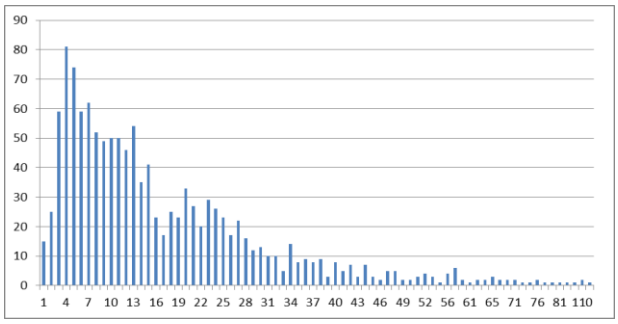

Figure 1. Length of Internet activities response (# of words)

Open-ended questions about values and attitudes. Open-ended questions about attitudes and values were generally taken seriously and elicited good results. Participants seemed genuinely pleased to share their opinions about important issues stemming from reuse of user-contributed material on the Internet and from archiving social media. For example, the photo survey asked about the practice of reusing online photos; answers often revealed an emerging personal system of rules, e.g. "*Personal photos from someone's personal album should probably not be reused. Perhaps if it is picture someone takes of a flower, or a sunset, and you find it pretty and want to save it, then that's fine. Pictures that have been circulating for a while, or not just limited to photos but also illustrations or drawings, could also be saved for reference or review, or to show a friend. However if the person specifically requests not to distribute the picture or try to make money from it, the request should be honored. If you had already been distributing the picture and the person asks you to stop, that request should also be honored.*" [PH227]

Detailed analyses of the open-ended responses are covered in our accounts of the individual studies. [20, 21, 3622, 23, 33] In these analyses, we used standard qualitative research methods such as open-coding [35] to discover patterns in respondents' open-ended responses; we were able to triangulate elliptical responses to resolve ambiguities.

Methodological Lessons

What methodological advice would we give to others who want to use US-based Turkers as study participants (beyond suggesting good design principles)? To our surprise, we found that Turkers had a high tolerance for completing open-ended questions, and their answers were articulate and consistent, as long as the questions fit the survey thematically (i.e. the questions sought the respondent's opinion on the topic at hand or were grounded in current practice), and answers could be written 'off the cuff' without consulting external resources (e.g. the respondent's own Twitter account). Scenarios and hypotheticals were most effective when they were entertaining, concrete, and had interesting details. Finally, the Turkers exhibit a capacity for completing long surveys as long as they are engaged and their efforts are respected; in line with Mason and Watts' results, we found that the quality of the work seems to be somewhat divorced from the pay [25].

PARTICIPATION

Mechanical Turk offers many surveys to qualified US-based workers. Do the Turkers who participate represent a coherent group, or do their characteristics differ substantially from study to study? As we described earlier, each questionnaire included demographic questions. Although comprehensive studies have been performed to describe US Turkers [13, 28], we wanted to understand the effects of our recruiting requirements and study content on participation.

Demographic Profile

In the demographic profile, we looked at several standard characteristics—age, gender, education level, whether or not they were students, and how long they reported having used the Internet. We also collected a baseline description of participants' Internet activities. This was useful for two purposes: it confirmed the primary recruiting requirement (use of the media type under investigation) and it gave us a sense of what else workers did online. Finally, we were interested in what workers published online, since this experience may influence their attitudes toward reuse [21].

Table 2 shows the participants' characteristics. According to Ipeirotis's demographic survey, females are over-represented among US-based Turkers: 65% report as female and 35% report as male. Our respondents are closer to parity at 55% female and 44% male (1% did not specify). Except for the photo-sharing study, our female participation rates are below his. What might be the source of this discrepancy?

In both cases, workers self-select to participate. Because we allowed participants to fill out a questionnaire for each media type, we must factor this in: if we eliminate extra reports from these participants, we have 1090 unique workers, 604 (55%) of which are female and 475 (44%) are male—in other words, the proportion is the same. Hence, the difference must lie elsewhere. Ipeirotis's survey was shorter than any of ours, and just collected basic demographic information—we seem to have targeted a different segment of the Mechanical Turk population, one that is more motivated by interest and a desire to be heard.

Indeed, if our participants were more motivated by the payment, they would not have written such extensive responses to open-ended questions (see Figure 1), particularly the abstract ones enquiring about their reuse ethos and their attitudes toward institutional archiving of social media. In support of this interpretation, his socioeconomic explanation of female over-participation (that the Turkers tend to not be employed outside the home) does not wholly align with the open-ended responses we will explore later in the section.

Education seems to be an orthogonal factor: in agreement with Ipeirotis's findings, 60% of our respondents have finished college, and 91% have attended at least some college. Although Ipeirotis does not check whether his respondents are currently students, about 1/3 of our participants report being students.

Our population also skews slightly younger than Ipeirotis's, although this may be part of the trend Ross et al. document toward younger US Turkers [28]. While 12% of his population reports being born before 1960, and 17% reports a birthdate in the 1960s, the combined figure for our survey respondents is substantially lower at 13%. We pass Ipeirotis in the younger groups—he reports about 38% born in the 1980s vs. 53% for our participants. While he only reports a little over 5% as born in the 1990s, 13% of our respondents are in that category (just old enough to Turk by the terms of use, and considered "digital natives" by most). The age difference between our results and Ipeirotis's may be because we are attracting younger people who are more apt to create digital media as well as consume it. But it is also possible that Mechanical Turk's constituency has changed significantly between Ipeirotis' survey period (2009/2010) and our last three (late 2011 to mid-2012).

Internet Activities

We thought it was important to get a structured snapshot of respondents' Internet activities and a more complete self-report (in case we overlooked anything). We first asked respondents to select all of the activities that they engaged in regularly: email, Facebook, online shopping, video-sharing (e.g. YouTube), instant messaging, photo-sharing (e.g. Flickr), videoconferencing (including Skype), Twitter, and multi-player online gaming. Figure 2 shows participation levels in these activities. The most frequent are email and Facebook: nearly all participants report using email, and 1121/1252 report using Facebook.

Survey Media	screened responses	% female/ male/ no response	have college degree	have some college	current students	born before 1960	born 1960-1969	born 1970-1979	born 1980–1989	born after 1989
Twitter	173	61/39/0 (105/68/0)	54% (94)	88% (152)	did not ask	4% (7)	4% (7)	17% (29)	64% (110)	11% (19)
Photos	242	71/27/1 (173/66/3)	55% (133)	91% (221)	34% (82)	1% (4)	10% (25)	22% (53)	57% (137)	10% (23)
Reviews	203	59/41/0 (119/84/0)	62% (125)	92% (186)	32% (64)	3% (7)	12% (25)	23% (47)	50% (101)	11% (23)
Podcasts	225	44/55/1 (99/123/3)	58% (130)	90% (202)	31% (70)	3% (7)	8% (18)	20% (45)	52% (118)	16% (37)
Videos	200	47/53/1 (93/105/2)	69% (137)	93% (185)	24% (48)	5% (10)	9% (18)	26% (51)	47% (94)	13% (26)
Educational recordings	209	50/50/0 (105/104/0)	57% (120)	94% (196)	36% (75)	6% (13)	8% (17)	19% (40)	51% (108)	14% (30)
Total	1252	55/44/1 (694/550/8)	60% (739)	91% (1142)		4% (48)	9% (110)	21% (265)	53% (668)	13% (158)

Table 2. Demographic data from the six surveys.

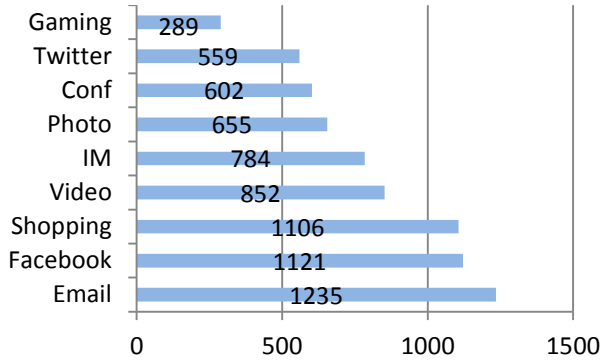

Figure 2. Participation in Internet-based activities.

An open-ended question about Internet activities acted as a second venue for profiling participants. They specified an average of 4.2 Internet activities in this alternate venue. Table 3 shows the top 10 categories of participant activities. Unsurprisingly, using social media, consuming digital media, and communication were specified as the three top activities (contrast this ordering with Figure 2, where email—i.e. communication—came out on top).

Many respondents are free-lancers, stitching together a variety of skilled online jobs (illustration, graphic design, programming, IT support, research, writing, and editing). Others use Turking as a respite from computer-centric jobs (clerical work, office management, legal research); some are working remotely. Still others are students, making a little extra money or distracting themselves from homework or lectures. About 40 of the participants were job-seekers. All of these spins on participants' professional lives support Ipeirotis's assertion that a significant proportion of US-based Turkers are doing HITs as a distraction or for entertainment. Posts in forums such as Turker Nation (http://turkernation.com/) suggest that some respondents participate in surveys to offer their opinions about hot-button issues like privacy, permission, pseudonymity, and anonymity; we discuss such forums later in the paper.

Category	Subcategories	Total
Social media	social networking; Facebook; keep/stay in touch; Twitter; Forums; Reddit; Myspace	1020
Consume	reading; watching videos/tv/movies; listening to music/podcasts/radio; surfing or browsing	972
Communicate	email; talking, Skyping, or videoconferencing; communicating/contacting; IM/chat	757
Research	research/search; researching specific topics; using specific resources	548
Work and school	work/job; specific work-related activities; school, learning, or homework; looking for jobs	460
Shopping	shopping/buying; shopping (specific stores); shopping for specific items; find coupons	432
Publish media	photo or video sharing; art; blogging; website development; media aggregation	297
Gaming	gaming (casual, online), multiplayer gaming; fantasy sports; specific games (e.g. WoW)	257
"Get Paid To…"	Mechanical Turk/HITs; surveys; other; Etsy and eBay selling	185
Entertainment	entertainment/fun; killing time/leisure; hobbies and crafts; porn	147

Table 3. Summary of responses to open-ended question about Internet activities.

Because we were investigating emerging social norms about the ownership of online content, we were interested in what participants reported publishing online. What surprised us was the degree to which participants counted self-description (e.g. Facebook profiles) as publishing.

Table 4 shows the categories and counts. Over a half of the respondents (728/1252, or 58%) reported publishing visual media (mostly photos and videos) to the Internet. Although a significant number (206/1252, or about 16%) reported publishing creative content, most of this was in the form of personal journals, memoirs, or personal blogs (as opposed to topical blogs). Only 25 said they had published their own artwork, and 21 respondents said they had published fiction, short stories, or poetry online. Fewer reported other types of creative efforts of their own including films (2), music videos (3), drawings (1), or humor (2). The most common type of content respondents reported publishing (805/1252, or 64%) was social media status updates. More surprisingly, 19% of the respondents considered their social media profiles to be a form of publication.

Publishing/Contributing Content

Primary type published	#	Subgenre examples	Example from survey
Social media status updates	805	*Pictures from events or daily life; descriptions of everyday activities; thoughts*	*" "I publish picture of my children and some status updates of how my day is going or how it was, but never what my plans are." [ED077]*
Photo/video	728	*pictures; photos; pix; videos; videoclips*	*"Pictures are what I share most." [RE095]; Since im [sic] a part time model, i mostly post my portfolio pictures on facebook.[PH234]*
Original factual content	388	*Topical blog posts; tutorials; answers; comments; reviews; videogame walkthroughs; etsy listings*	*"I have a blog devoted to my favorite college hockey team." [PH154]; "How to: Electrical wiring and Auto Mechanics" [PC168]*
Republish	282	*Funny or interesting articles; (links to) articles, videos, or blog posts*	*"videos from funny or die or from Youtube - also share videos reated [sic] to education for nurses (my former career)" [ED068];*
Social media profiles	233	*Personal data (e.g. name, age, location, email address); profiles; credit card data); likes and dislikes; resume; favorites*	*"I share … personal information such as likes, dislikes, name, location, DOB, occupation, etc." [PH140]; "Minimal contact information (such as email), first name, general area (state), hobbies and interests, stuff like that." [PH127]*
Original creative content	206	*fan fiction; original videos and music; stories; artwork*	*"I am a member of Deviantart and publish my artwork on that site regularly." [VC149]*
None	51	*None*	*"I don't share much but I observe others through facebook, etc." [PC202]*
Other (OTH)	38	*scientific data; code; design patterns; school assignments*	*"Scientific data" [RE119]; "Code, small scripts or programs." [PC147]; "WISH LISTS ON AMAZON" [PH156]*

Table 4. Participants' online publishing activities

Participation lessons

We take away three important lessons from the detailed picture of the workers who participated in our surveys:

(1) MTurk can provide a diverse set of participants for many different types of studies of online behavior. Turkers are a good source of reliable self-reports of many nascent phenomena (for example, the emerging view of social media profiles as publishing or the relative penetration of various online technologies). Although participants are better educated and more Internet-savvy than the general online population (e.g., see [36]), they may represent an important growing sector of information workers.

(2) Workers take recruiting requirements seriously; however, triangulation helps guarantee full compliance.

(3) It is difficult for any researcher to get a stable picture of the MTurk population. As we see from the difference between our study population and Iperotis's, different workers may take different types of surveys. Furthermore, certain workers are more attracted to survey-taking tasks.

RELIABILITY

Crowdsourcing researchers have investigated a number of techniques to ensure data quality [2,14]. We incorporated their suggestions in our survey designs, but we felt that the most effective ways to ensure data quality were to maintain worker engagement (by developing interesting scenarios), to reduce worker frustration (by ensuring that questions were easy to interpret), and to respect workers' opinions (by asking questions about personal ethics). This strategy is consistent with Eickhoff and de Vries' observation that the best way to discourage malicious workers is to offer creative, non-repetitive tasks [8].

Although reading comprehension questions help detect scammers, they are also difficult to design well and we suspected they were mildly insulting to committed workers. We found that there were enough other 'tells' (e.g., nonsense answers to open-ended questions and impossibly short completion times) that the reading comprehension questions were mostly redundant for fraud detection. On the other hand, comprehension questions might promote careful reading, so we would approach removing these questions with care.

Formally, we used three pre-engagement screening criteria: (1) we requested workers who had performed in the past with 95% reliability [15]; (2) we paid workers at rates established for comparable surveys, 50 cents per HIT [17]; (3) we requested that workers be familiar with the media type that was the survey's focus. We also used a point system to remove bad data from the mix; responses received one point each for any of the following anomalies: (1) minimal time spent on the survey; (2) each wrong answer to a reading comprehension question; (3) unanswered questions or nonsense answers (e.g. a few respondents pasted the instructions into the response box); suspicious patterns in the Likert-scale responses (e.g. all values being the same). We were conservative about data hygiene; if we doubted a response's veracity, we threw out the results for that respondent.

Even with stringent quality tests, we detected relatively few fraudulent participants. Of course, some spent less time on the open-ended responses than others, but even so, we were pleasantly surprised by how forthcoming the respondents were (especially given both authors' prior experience administering surveys). We paid all respondents regardless of whether we discarded the data.

We concluded each survey with a question about whether participants would be willing to do another survey "like this one." If we got an appreciable number of "no" answers, we would know that we had upset the balance of questions, attention, and payment. Generally, there were only a sprinkling of "nos" (from 1 to 7), with the 7 stemming from dissatisfaction with the educational videos survey. Note that this is one of the longer surveys, and it is the one with the lowest time spent per question. It is also the study in which we discarded the most suspect data.

Low work times did not necessarily signal reduced data quality. Some of the faster completion times were associated with participants who were likely to have been focusing on the survey rather than multi-tasking (e.g. watching TV or listening to a classroom lecture while Turking). In fact, as we discuss later, sometimes Turkers hold surveys to avoid potential rejection during requestors' survey quality screening processes.

Table 5 shows the relative times workers spent on the surveys, and the number of good and bad responses we received. We also report the minimum and maximum work times on data we kept and data we discarded. The higher level of bad responses on the final surveys may reflect changing demographic characteristics of US-based Turkers or may be the result of requestors approving lower quality

Survey	# Respondents	Avg. work time (sec)	Max work time (sec)	Min work time (sec)	Length (# of questions)	Time per question (sec)
Twitter	173 (190)	522 (557)	2295 (2742)	187 (154)	34	15.35 (16.39)
Photos	242 (250)	801 (394)	2896 (579)	226 (228)	41	19.53 (9.62)
Reviews	203 (216)	890 (457)	2577 (1373)	223 (135)	50	17.79 (9.14)
Podcasts	225 (239)	724 (267)	3181 (760)	158 (38)	42	17.24 (6.35)
Video	200 (229)	656 (315)	2644 (1054)	153 (53)	40	16.39 (7.88)
Edu .Rec.	209 (250)	681 (502)	3351(3597)	207 (33)	45	15.13 (11.16)
Total	1252 (1377)	720	3351	153	252	

Table 5. Overview of survey performance and data cleaning. Parenthetical values reflect the data, pre-cleaning.

work (i.e. a 95% acceptance rate may no longer be a good indicator of Turker reliability [8]). It may also stem from the fact that the HITs were exposed longer: they were available to Turkers for up to a month.

It pays to take a closer look at the bad responses; there weren't that many of them on the early surveys, just 38 out of 643 total responses, or a little under 6%. Many of these weren't out-and-out fraud either: a long survey was left unfinished; reading comprehension questions were misinterpreted; recruiting requirements (e.g. English as a first language) were fairly harmlessly violated; in fact, the initial open-ended questions were, without exception, answered in a wholly acceptable way.

It wasn't until the three later surveys that we saw a higher rate of suspect data; 84 out of 718 responses or a little under 12%—about twice as many—were probably bad. The open-ended responses on those surveys were more apt to be unacceptable; 21/84 (25%) were clearly bad (they were blank or nonsense). We were left wondering whether the Turker population had changed, whether the time of the year influenced the number of scammers (the last three surveys were in place over the winter holidays), whether administering three surveys at once was provoking fraud, or whether the 95% prior acceptance rate was no longer an effective screening metric.

We looked for patterns in the bad data. The educational recordings survey is by far the worst in terms of bad data. It is also one of the longer surveys, and arguably (by our own admission) a scenario that may be more difficult to relate to—two scientists disagree on how data might be interpreted. Although this situation is familiar to us, it is evidently less so to the workers. The videoconferencing survey elicited the second worst performance, although the scenario should be more familiar since many of the respondents report they watch instructional videos on YouTube. The workers generating this data seemed more disengaged: the minimum work time for the three later surveys stands out as being considerably shorter than the work times of the earlier surveys.

A second plausible hypothesis might be that the availability of multiple surveys stretched the goodwill of our participants or attracted spammers. If this is so, then we'd expect the bad responses on the later surveys to come from multiple survey takers. This hypothesis is probably untrue, since 14 out of 104 workers (or about 13.5%) with discarded data submitted more than one survey, as compared to 121/1090, or about 11% of workers, who produced acceptable data as they completed multiple surveys. It may also be that we fatigued workers, so we should check overall multiple survey participation (that is, mix the data back together); in that case, the data is an unsurprising 12% (138/1185)--in other words, workers who performed poorly on one survey probably were not burn-outs from other surveys. Instead we see that good workers probably perform well on multiple surveys and bad workers perform poorly on multiple surveys.

Dates	Media	Bad	Total	% Bad
April/May 2010	tweets	17	190	8.90
August 2010	photos	8	250	3.20
June/July 2011	reviews	13	216	6.02
Dec 2011 - Jan 2012	videos	34	404	8.42
April 2012	videos	60	314	19.11

Table 6. Discarded data rates for the 5 time periods.

Is the US-based Turker population changing? A rising number of male respondents and a higher (and more conspicuous) fraud rate in the later surveys drove us to examine the responses by survey date. Our first survey, conducted while we were still in the throes of developing the method and learning Mechanical Turk best practices, resulted in almost 9% bad data. As we already mentioned, much of this data was not clearly bad; some misdetection resulted from ambiguities in the reading comprehension questions. The bad response rate went down for the next two surveys (personal photos and reviews). The most recent surveys were fielded in two distinct time periods – over the 2011-2012 holiday season and again in April 2012. Here we see the largest change: the percentage of bad responses rose above 19% during April (see Table 6). We might surmise that this rise in fraud was caused by an influx of new Turkers or the perception that surveys are easy to game and potentially more lucrative than other HITs (as shown in worker forums such as Turker Nation).

However, as Table 7 shows, the effect was not uniform across the three surveys. Responses to the podcast survey were high quality during both periods. This survey was also available for only a few days in April since the response quota was met quickly. On the other hand, the remaining two surveys were available for 1-2 weeks, potentially providing more time for communication among Turkers about the survey and exposing them longer to fraudsters.

The Effect of Forums

Turking is not done in isolation. Turkers may talk to one another in external forums (or directly—some surveys were completed in apparent collusion). To find out what Turkers talk about, we monitored forum posts in Turker Nation (especially posts about surveys) and in the Subreddit HITsWorthTurkingFor, a group in which crowdworkers recommend appealing (well-paying and interesting) tasks to one another. Other research has conducted more thorough analysis of the content of crowdworker forums, and some experienced HIT designers suggest regular visits to these

	Dec./Jan. (simultaneous)			April (serial)		
Media	Bad	Total	% Bad	Bad	Total	% Bad
podcasts	10	180	5.56	4	59	6.78
videoconference	9	107	8.41	20	122	16.39
educational	15	117	12.82	26	133	19.55

Table 7. Bad data rates over two data collection periods.

venues to discover what workers are saying [2].

From the forum posts, we learned that survey completion time can be an unreliable metric. Sophisticated Turkers are aware that surveys that are completed too quickly stand a good chance of being rejected. Hence they advise each other to hold back on survey submission. TurkerNation member jdowling advised, "*Sometimes if it seems a little quick i'll let it sit for a few minutes before submitting.*" To avoid returning an undesirable (or broken) survey, F1rsTxLas7 said he "*complete[s] a survey first, then take the code, accept the HIT, put in the code, and submit.*" This strategy will result in a deceptively low completion time.

However, the forums also reveal that some crowdworkers understand the forces acting on the researchers too. Via the same forum, amaeru said "*IMO, requesters can reject for whatever reason they want. Especially on surveys, which are often academic, the requesters have to ensure that the data they are getting is accurate--finishing too quickly may show them that you're not paying attention and that the data you've provided is therefore unusable.*" But then the writer went on to suggest, "*after you complete the survey, let the timer run until it reaches a more reasonable time, and then submit.*" Worker paperprincess70 agreed, saying "*…I usually let the HIT sit for a minute after completing the survey so that it doesn't appear that I've sped through it."*

The Turker Nation members are sufficiently aware of constraints on the researchers to raise the possibility of turning in negligent requestors to their institutions' IRBs: "*After becoming annoyed last night at seeing her [the requestor under discussion] re-post this study under a new requester name I reported her to the Princeton IRB. …*" This poster went on to publish a URL that would enable other Turkers to contact the researcher's IRB.

Forums also reveal that survey topic matters to Turkers. As we suspected, at least some Turkers are looking for surveys on topics they care about, both because it's easier to complete these surveys and because they want their opinions to be heard. For example, Joeturker said, "*…Both surveys I did were about 12 questions, a sentence or two for each question. Since both surveys I did were on subjects where I have strong opinions anyway (Global Warming and Privatizing Social Security) they were very easy to write…*" While we don't want to bias results by revealing our specific interests, it is important to remember that workers may welcome an opportunity to vent in a meaningful venue, since they are aware that researchers publish their results.

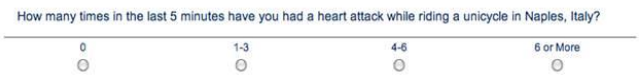

Figure 3. Excerpt from Turker Nation post demonstrating participant appreciation of funny or unusual attention checks.

Similarly, in line with our experiences suggesting that maintaining participant engagement and respecting their skills and commitment are as important (if not more important) than catching fraudsters, forum posts confirm that humor and efforts to engage workers matter. There are entire threads on the discussion boards devoted to the 'quality assurance' questions that workers find funny or interesting. For example, Figure 3 shows a screen capture that poster BoomMike did to show his fellow Turkers "*a creative attention check*." On the Subreddit, workers alerted each other to ACs (attention checks) and MCs (memory checks), treating them much the way one driver might alert another to a highway patrol speed trap ("*Short and simple HIT. Couple AC's, One MC.*"—lampshade3).

Reliability Lessons

In general, the response quality was good. Even the data we discarded from the early surveys was submitted in good faith. Assumptions must be examined carefully when researchers build in various quality mechanisms, including comprehension questions (which, as we saw in the Subreddit, may irritate or confuse respondents) [17], data cleaning methods [32], and feedback [5]. One possibility that seems worth investigating is that longer exposure poses additional opportunities to attract fraudsters. Unlike most other survey venues, methods for improving data reliability need to take into account communication among workers and their understanding of researcher practices. Turker interest and engagement, coupled with good survey design, still seems to be the best assurance of high-quality data.

LIMITATIONS

We acknowledge certain limitations to our method and its results. After performing six of these studies, and comparing them to our other experiences with Mechanical Turk and other crowdsourcing platforms, we feel relatively comfortable with the constraints that are built into our method and the effects of these constraints.

Methodological limitations. There are several inherent limitations to this method. In our surveys, the goal has been to elicit respondents' beliefs, attitudes, and self-reports of recent actions. Although we skirt around the edge of certain legal taboos (e.g. the reuse of digital content in a manner that the respondent may feel violates copyright restrictions), we are relying to some extent on worker anonymity to allow them to express attitudes and describe actions that are marginal. If the topics significantly violate legal or cultural expectations, this method may not elicit truthful answers. For example, we suspect that scenarios involving downloaded music or porn (and other significant legal or social transgressions) might not be as successful as our relatively benign scenarios.

Furthermore, if the behavioral questions are too detailed and out of the reach of normal memory, respondents may not go out of their way to verify their answers. As we discussed earlier, our questions about Twitter followers were probably not wholly accurate.

Naturally the topic must be amenable to developing concrete scenarios. We feel that omitting a scenario's

details might cause respondents to act on varying assumptions. For example, if we developed generic family scenarios, we might be making assumptions about respondents' relationships with, say, their siblings.

The population of interest must be available within the pool of reliable Turkers. For example, if an insufficient number of Turkers were podcast listeners, our podcast survey HITs might go unfulfilled. Similarly, we rely on workers to be relatively truthful about whether they meet a survey's requirements. Needless to say, finding participants who are not computer users or who are very inexperienced computer users, would be problematic. Similarly, finding participants who are uniformly high earners, or who are very busy, would fall outside the scope of this method.

Finally, we are aware that privacy concerns introduce a very real tension. Specifically, Amazon's terms of service (as well as those of other platforms) protect crowdworkers' privacy. Yet researchers' minimal demographic questions, coupled with other open-ended questions, might indeed force workers to decide whether they are surrendering more privacy than they intend to. It's difficult for both researcher and respondent to predict what combination of answers when taken together will reveal the respondent's identity or sacrifice some other aspect of his or her privacy [27].

Respondents are sensitive about their privacy. Although our open-ended question about online publishing did not ask about privacy, 43 responses mentioned privacy explicitly, and others alluded to having published more information than they had been told was prudent; for example, ED168 began his response by saying, "*Likely [I've published] too much, hah!*" We are aware that some desire for privacy is aspirational, and that privacy may be readily surrendered when respondents are faced with real situations [1]. At the same time, researchers don't wish to violate respondents' rights or Amazon's terms of service, but they do need to ask enough demographic questions—and questions about the participants' practices—to satisfy a study's requirements; this data may be easily aggregated (as we have shown) and brought together with other online data sources. Technical data curation solutions (e.g. differential privacy [7]) might be brought to bear on this tension, but it is unlikely that a technical solution (especially one that assumes a closed world) will address broader socio-technical concerns, especially when many Turkers rely on privacy through obscurity [11]. Communication about privacy in forums like Turker Nation may be more effective and realistic [34].

Limits to our results. Our results are limited to the US Turker population, which has some specific properties we take advantage of, such as workers' desire to use HITs as entertainment [13, 28].

We also acknowledge that it is difficult to compare or generalize our results to a ground truth collection (a so-called gold set) or to similar tasks (as one would in a relevance judgment situation [2]). On the upside, our experiences suggest that respondents are more engaged by this sort of survey than they are by relevance judgment tasks, which may be both difficult and frustrating. Because we have paid significant attention to entertaining the workers, our results may not generalize to surveys with drier content (e.g. straightforward demographic surveys).

CONCLUSION

The Turkers provided us with a substantial glimpse into their online ownership, control, and reuse behavior and attitudes—that was our primary reason for performing the individual studies. In this paper, we have aggregated six studies' worth of Mechanical Turk data for three reasons: First we wanted to document our method and reflect on its strengths and limitations. Second, we sought to characterize the US-based Turkers who participated, both to better understand them and to show the effects of varying recruiting requirements. Finally, because data reliability is such a persistent question when researchers survey the crowd, we felt it was important to take it on from different angles, including how the Turkers are changing in the face of increasing survey research.

One finding that has surprised us is the participants' high level of engagement, as demonstrated by the data quality. Of course there is no way to guarantee that this level of quality is sufficiently stable to expect it indefinitely; but even as the constituency of US-based Turkers changes, we continue to gather useful data. A seventh study, completed after this paper was written, confirms our sense that topics of greater interest to the target population (in this case, massively multiplayer online games) elicit better data, both because recruiting requirements are met more quickly and because participants are invested in the subject matter.

The analysis of our data has revealed a new (or alternative) demographic to the one Ipeirotis originally identified, a modern information labor force, one that pieces together work from many sources, diverts itself in front of the same screen as it works, and vacillates between a mild sense of exploitation and control. The demographic characteristics of survey-takers have been changing over the two years we have been conducting MTurk studies. It may be that we are reading symptoms of larger changes afoot as Turker subcommunities develop—survey-takers talk to other survey-takers and frictions develop between Turkers and researcher-requestors—and aspects of the workplace, such as the reliability metric, shift in meaning and utility.

As time goes on, we have come to realize that our biggest worry is not the spammers; unlike relevance judgment tasks (or even surveys consisting wholly of Likert-scale questions), we can rely on responses to open-ended questions—questions that work well in online surveys [16]—to separate bad quality data from the good. Instead, our chief concern is maintaining goodwill and promoting the Turkers' engagement via entertaining (but realistic) situations and provocative open-ended questions.

REFERENCES

1. Acquisti, A. and Grossklags, J. Privacy Attitudes and Privacy Behavior, in J. Camp and S. Lewis (Eds.) *The Economics of Information Security*, Kluwer, 165-178.
2. Alonso, O. Implementing crowdsourcing-based relevance experimentation: An industrial perspective, *Information Retrieval Journal* (2012), in press.
3. Antin J. & Shaw, A. Social desirability bias and self-reports of motivation: a study of amazon mechanical turk in the US and India. *Proc. CHI '12*. 2925-2934.
4. Boyle, J. *The Public Domain: Enclosing the Commons of the Mind*, Yale University Press, New Haven, 2008.
5. Dow, S., Kulkarni, A., Klemmer, S., and Hartmann, B. 2012. Shepherding the crowd yields better work. *Proc. of CSCW '12*. 1013-1022.
6. Downs, J., Holbrook, M., Sheng, S., and Cranor, L. Are your participants gaming the system?: Screening Mechanical Turk workers. *Proc. CHI'10*. 2399-2402.
7. Dwork, C. Differential privacy. *ICALP*, 2006, 1–12.
8. Eickhoff, C. and de Vries, A.P. How Crowdsourceable is Your Task? *Proc. Workshop on Crowdsourcing for Search and Data Mining*, 2011.
9. Fetterman, D. *Ethnography*. Sage, 1989.
10. Greengard, S. Digitally Possessed. *Communications of the ACM*, 55 (5), 2012, 14-16.
11. Herzog, W. & Stutzman, F. The Case for Online Obscurity, *California Law Review*, Vol. 101.
12. Hill, B., Monroy-Hernandez, A., and Olson, K. 2010. Responses to Remixing on a Social Media Website. *Proc. AAAI Conf. on Weblogs and Social Medi*a. 74-81.
13. Ipeirotis, P. *Demographics of Mechanical Turk*. NYU Tech Report, 2010.
14. Ipeirotis, P., Provost, F. and Wang, J. Quality Management on Amazon Mechanical Turk. *KDD-HCOMP*, 2010.
15. Jakobsson, M. Experimenting on Mechanical Turk: 5 How Tos. *ITWorld*, September 3, 2009.
16. Kiesler, S., and Sproull, L. Response Effects in the Electronic Survey. *Public Opin Q* 50 (3): 402-413.
17. Kittur, A., Chi, E., and Suh, B. Crowdsourcing User Studies with Mechanical Turk. *Proc. CHI'08*. 453-456.
18. Lessig, L. *Remix*, Penguin, New York, 2008.
19. MacCormick and Summers, (eds.) *Interpreting Precedents*, Ashgate/Dartmouth, 1997, pp. 528-9.
20. Marshall, C.C., and Shipman, F.M. Social media ownership: Using Twitter as a window onto current attitudes and beliefs. *Proc. CHI'11*, ACM, 1081-1090.
21. Marshall, C.C., and Shipman, F.M. The ownership and reuse of visual media. *Proc. JCDL'11*. ACM, 157-166.
22. Marshall, C.C., and Shipman, F.M. On the institutional archiving of social media. *Proc. JCDL'12*. ACM, 1-10.
23. Marshall, C.C. and Shipman, F.M. Saving, reusing, and remixing web video: using attitudes and practices to reveal social norms. *Proc. WWW'13*. ACM.
24. Mason, W. and Suri, S. Conducting Behavioral Research on Amazon's Mechanical Turk (October 12, 2010). *Behavior Research Methods*, Forthcoming.
25. Mason, W. & Watts, D. Financial incentives and the performance of crowds. *Proc. SIGKDD 2009 workshop on human computation* pp. 77–85.
26. Odom, W. Sellen, A., Harper, R., and Thereska, E. Lost in Translation: Understanding the Possession of Digital Things in the Cloud. *Proc. CHI'12*. 781-790.
27. Palen, L. and Dourish, P. 2003. Unpacking "privacy" for a networked world. *Proc CHI 2003*, 129-136.
28. Rissland and Ashley, "Hypotheticals as Heuristic Device." Proceedings of Strategic Computing Natural Language Workshop, Marina del Rey, California, May 1-2, 1986, p. 168.
29. Ross, J., Irani, L., Silberman, M.S., Zaldivar, A., & Tomlinson, B. Who are the crowdworkers?: shifting demographics in mechanical turk. *Proc. CHI EA '10*. 2863-2872.
30. Rzeszotarski, J. and Kittur, A.. 2011. Instrumenting the crowd: using implicit behavioral measures to predict task performance. *Proc. UIST '11*. 13-22.
31. Schmidt, L., Crowdsourcing for Human Subjects Research. *Proc. CrowdConf 2010*, SF, CA.
32. Schnoebelen T. and Kuperman, V. Using Amazon Mechanical Turk for linguistic research, *PSIHOLOGIJA 43*, 4, 441–464.
33. Shipman, F.M. and Marshall, C.C. Are user-contributed reviews community property? exploring the beliefs and practices of reviewers, *Proc. WebSci*, 2013.
34. Silberman, M., Irani, L., and Ross, J. Ethics and Tactics of Professional Crowdwork. *XRDS 17*, 2, 39-43
35. Strauss, A. and Corbin, J. *Basics of Qualitative Research*, Sage Publications, 1998.
36. United States Census Bureau. *Computer and Internet Use in the United States: 2010*. http://www.census.gov/hhes/computer/publications/2010.html. Retrieved 29 January 2013.