Demographics and Dynamics of Mechanical Turk Workers

Djellel Difallah
New York University
New York, NY, USA
djellel@nyu.edu

Elena Filatova
City University of New York
Brooklyn, NY, USA
efilatova@citytech.cuny.edu

Panos Ipeirotis
New York University
New York, NY, USA
panos@nyu.edu

ABSTRACT

We present an analysis of the population dynamics and demographics of Amazon Mechanical Turk workers based on the results of the survey that we conducted over a period of 28 months, with more than 85K responses from 40K unique participants. The demographics survey is ongoing (as of November 2017) and the results are available at http://demographics.mturk-tracker.com: we provide an API for researchers to download the survey data.

We use techniques from the field of ecology, in particular, the capture-recapture technique, to understand the size and dynamics of the underlying population. We also demonstrate how to model and account for the inherent selection biases in such surveys. Our results indicate that there are more than 100K workers available in Amazon's crowdsourcing platform, the participation of the workers in the platform follows a heavy-tailed distribution, and at any given time there are more than 2K active workers. We also show that the half-life of a worker on the platform is around 12-18 months and that the rate of arrival of new workers balances the rate of departures, keeping the overall worker population relatively stable. Finally, we demonstrate how we can estimate the biases of different demographics to participate in the survey tasks, and show how to correct such biases. Our methodology is generic and can be applied to any platform where we are interested in understanding the dynamics and demographics of the underlying user population.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing**;

KEYWORDS

crowdsourcing; demographics; dynamics; surveys; selection bias; Amazon Mechanical Turk; capture-recapture

ACM Reference Format:
Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining , Marina Del Rey, CA, USA, February 5–9, 2018 (WSDM 2018),* 9 pages.
https://doi.org/10.1145/3159652.3159661

WSDM 2018, February 5–9, 2018, Marina Del Rey, CA, USA

ACM ISBN 978-1-4503-5581-0/18/02…$15.00
https://doi.org/10.1145/3159652.3159661

1 INTRODUCTION

Crowdsourcing, in general, and Amazon Mechanical Turk in particular, have been extensively used over the years for a wide variety of applications that require access to a large number of humans. Many of these tasks, especially tasks around social sciences (e.g., psychology, marketing, political science, and others) are sensitive to the demographics of the underlying population. Other tasks, e.g., surveys, require access to a large number of people to ensure that the results are representative and can be generalized from the crowdsourcing platform population to the general population.

Amazon does not provide any information about the evolving demographics of the workers population. While there exist multiple demographics studies[1] of the Mechanical Turk workers population [17, 26], these studies provide only a snapshot-in-time analysis of the MTurk workers population and, thus, do not contribute to the understanding of the evolution and dynamics of the MTurk workers population. In our work, we present a comprehensive, longitudinal study of Mechanical Turk. We present data of the demographics survey that was conducted over 28 months, with more than 85K responses, and approximately 40K unique participants. We present the evolution of the MTurk demographics over time and illustrate the composition of the MTurk workers population across demographics variables such as country, gender, age, income, marital status, and household size. We contrast and compare these results against the demographics of the general population.

The second question that we address is the size of the MTurk workers population. Amazon claims that there are hundreds of thousands of workers on MTurk.[2] However, studies claim that the real number of workers available to participate in academic experiments is much smaller [34], and is closer to 7300 workers. Such conflicting messages can lead to confusion, especially for those who want to leverage Mechanical Turk for conducting studies. For example, an experimenter would like to run a survey that requires an access to 10K distinct participants. Is MTurk population size sufficient for this experiment? We answer this question by building on top of the techniques from the field of ecology, specifically the *capture-recapture* techniques. Capture-recapture techniques operate by taking sample sets of the population individuals over time, and examining the overlap among these sample sets to provide the estimate of the overall population size. Prior studies [34] use capture-recapture models for estimating the size of Mechanical Turk but rely on the assumption that all workers are equally likely to participate in the posted tasks. However, when workers have *different propensities to participate in tasks* this assumption leads to significant underestimation of the true size of the MTurk workers

[1] See http://www.mturkgrind.com/threads/demographics-of-mechanical-turk.26341/ for a comprehensive list of demographic surveys of Mechanical Turk workers.
[2] According to https://requester.mturk.com/tour, MTurk provides "*Access more than 500,000 Workers from 190 countries*", as of August 2017.

population. In fact, we demonstrate that the number of available workers on Mechanical Turk is at least 100K, with approximately 2K workers being active at any given moment. We show that the MTurk workers' half-life is 12-18 months, indicating that the population refreshes significantly over time. This is good news for anyone willing to conduct wide-reaching studies using this platform.

Finally, we address a common problem with surveys: When researchers present results from surveys, a common concern is that these results are biased as they are obtained from users who choose to participate in this particular survey. We address this issue by modeling, for each worker, the *propensity* to complete a particular task. Then, we examine how demographics are correlated with propensities to participate, and we infer the hidden selection biases. Our results indicate that most demographics variables are not affected by selection biases; the notable exception being Indian workers, that demonstrate a significantly higher propensity to participate in our survey, and are, therefore, over-represented in the raw results. However, with our presented methods, we can now adjust the over-estimates to their true values.

In summary, the contributions of the paper are:

(1) A longitudinal analysis of various demographics indicators of Mechanical Turk workers (country, gender, age, income, marital status, and household size), showing which demographics variables remain stable over time, which ones change, and comparing these to the demographics indicators of the general US population (Section 3).
(2) A capture-recapture analysis to estimate the size of the overall MTurk workers population. Our capture-recapture analysis models the expected lifetime of the workers in the marketplace and also takes into consideration the different propensities to participate in a given task. We provide theoretical and experimental proofs indicating that models that assume equal propensities of participation generate population estimates can be off by orders of magnitude (Section 4).
(3) An analysis that correlates demographics variables with participation propensities and lifetime of workers. This technique removes the selection biases from surveys and allows us to generate unbiased estimates of the demographic profiles of Mechanical Turk workers (Section 5).

For reproducibility, our code is available at https://github.com/ipeirotis/mturk_demographics and the data can be downloaded through an API at http://demographics.mturk-tracker.com.

We believe that the results reported in this paper are not only interesting as a description of the MTurk population and its dynamics, but they also provide an insight into the MTurk demographics to those who hire their study participants via MTurk. Furthermore, the techniques presented in this paper on estimating unbiased population characteristics through biased surveys can be used to study a wide variety of systems that can only be sampled through convenience samples.

2 RELATED WORK

MTurk labor marketplace is widely used by researchers in a variety of fields to recruit participants for experiments on data collection, data annotation, survey completion. Sociology, psychology, behavioral, economics, political science researchers actively recruit study participants on various crowdsourcing platforms [2, 3, 5, 23, 32, 33, 35]. Many studies show that the data collected using MTurk workers is comparable in quality to the data collected using undergraduate students or professional/commercial panels [19]. High correlation between the data collected using MTurk workers and participants recruited using other methods (e.g., college students) occurs despite the MTurk workers being significantly more socioeconomically and ethnically diverse than test participants recruited using other methods [6].

The validity of the experiments that use human subjects depends on the understanding of participants demographics. Chandler and Shapiro [8] evaluate the validity of experiments using MTurk in clinical psychological research from the point of view of sample composition and collected data quality. Other researchers are interested in getting access to experiment participants from a particular location [12], or, on the contrary, from around the world [13]. Arechar et al. [3] investigate the use of MTurk for interactive experiments in the field of economics. On the one hand, they demonstrate that MTurk can be used to replicate the results obtained in the physical laboratory, but on the other hand, they emphasize the importance of knowing the MTurk participants demographic profile.

Given that while running experiments using MTurk participants, researchers have minimal control over who volunteers to participate in the study, concerns regarding the quality of the data submitted by MTurk workers have been discussed [7, 30, 32]. According to this discussion, several methods exist that allow to filter out most of the low-quality results from MTurk workers.

Several researchers raise the concern regarding the ethics of using MTurk workers for research experiments [15, 16]. One of these concerns is considered particularly problematic, namely the low wages that MTurk workers get for their participation in the experiments. Often this issue is associated with particular locations of MTurk workers.

One work that attempts to estimate the number of MTurk workers who are ready for work at a particular moment is described in [34]. However, as mentioned in the introduction (Section 1), we believe that the presented number underestimates the MTurk workers population size. There are two reasons for this underestimation: 1) the experiment set-up where MTurk workers participate in a variety of unrelated tasks with different pay rates, that ran in irregular time intervals, and 2) insufficient calibration of the capture-recapture model. Both of these issues rely on the assumption that different MTurk workers have the same probability or propensity to participate in every experiment. This bias is a typical issue for the ecology capture-recapture models that ignore the population heterogeneity [1, 28] resulting in population size underestimation. We build on the findings from the work in the field of ecology [4, 9, 24, 25] regarding the population heterogeneity and how it affects the probability to be caught during the capture-recapture experiment.

Within the context of web-related research, Trushkowsky et al. [36] use capture-recapture models while trying to measure the cardinality of relations extracted from the web. Lu and Li [21] incorporate the heterogeneity issue into their application of the capture-recapture model for the deep web size estimation: different documents have different probabilities to be retrieved given a random query.

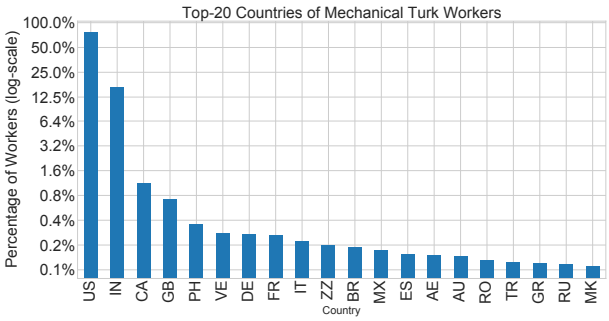


Figure 1: The top-20 countries of origin for Mechanical Turk workers. Most of the workers are from the USA (75%), with India (16%) being second, followed by Canada (1.1%), Great Britain (0.7%), Philippines (0.35%), and Germany (0.27%).

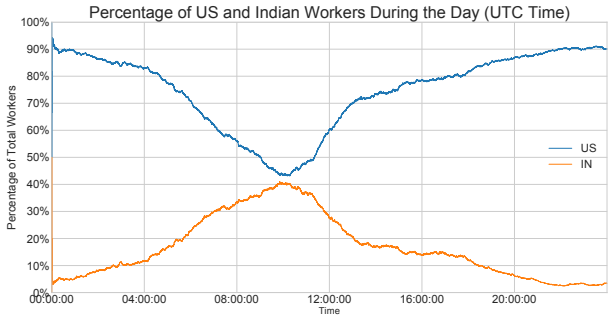


Figure 2: Percentage of US and Indian workers throughout the day.

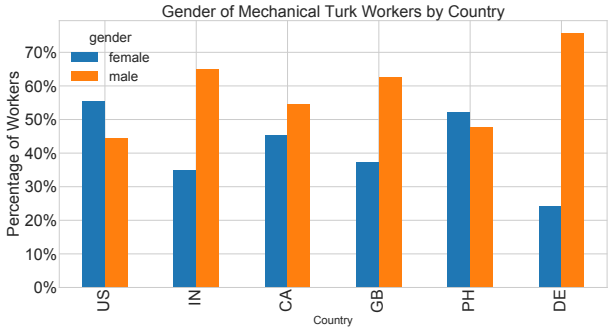


Figure 3: Gender breakdown across countries.

In our previous work [14], we focus on understanding the dynamics of MTurk from a market perspective, where we analyze the demand and supply on the marketplace, and examine the features that drive the speed of completion of a batch of tasks. In this work, we are interested in modeling the size and the demographics of the MTurk worker population by accounting for the workers' propensity to participate in a survey that we regularly post on the platform. We show that failing to account for the task's specific propensity leads to the underestimation of the MTurk workers population size when applying classical capture-recapture models.

3 DEMOGRAPHICS OF MTURK WORKERS

In this section, we describe our survey data collection methodology, and present the basic results for the demographics of the Mechanical Turk workers.

3.1 Data Collection

We collect basic demographics information about Mechanical Turk workers, by periodically posting a survey task asking workers to submit the following information: (a) Gender, (b) Year of Birth, (c) Marital Status, (d) Household Size, (e) Household Income, and (f) Location (City, Country). [3]

We post one survey task every 15 minutes. The survey can be completed by a single MTurk worker and takes on average 30 seconds. For every submitted survey response, we pay 5 US cents. Each worker on the platform can participate in our survey once every 30 days. As of now (August 12, 2018), the survey is continuously posted every 15 minutes, starting on March 26, 2015. For the purpose of this paper, we report the results obtained from all the surveys posted between March 26, 2015 and July 31, 2017 (859 days). In this time frame, we collected a total of 84,511 responses, submitted by 39,461 unique workers.

3.2 Survey Results Analysis

We now present the analysis of the survey results.

3.2.1 Country: Figure 1 shows the top-20 countries from which MTurk workers completed our survey. Most of the workers are from the US (75%), with India (16%) being second, followed by Canada (1.1%), Great Britain (0.7%), Philippines (0.35%), and Germany (0.27%).[4] The dominance of US and India among the worker population is well-documented in prior studies [17, 22, 26]. As expected, due to the time difference, the activity levels of US and Indian workers are different. Figure 2 shows the percentage of the US and Indian workers during the day on MTurk platform, illustrating that at 10am UTC the percentage of workers from the US is at a minimum at around 45%, while midnight UTC is at a maximum at around 90%. Our analysis did not detect any other significant periodicities in the data (e.g., day of the week, etc.)

A notable effect that we noticed in our surveys is the increase of international workers, happening around May 2016. Figure 4 shows the percentage of the MTurk workers from US, India, Canada, and Great Britain over time. Note that around May 2016 there is a sharp drop in the percentage of US workers on MTurk, and a corresponding increase in the percentage of workers from Canada, Great Britain, and other countries. We believe, this is due to the reverse of the Amazon policy from 2012, which restricted the enrollment of international workers. However, as we can see in the data, after the initial spikes in participation, the percentage of international workers has been steadily falling and is slowly converging towards April 2016 levels.

3.2.2 Gender: Our results indicate a generally balanced workforce, with 51% female workers and 49% male. However, we detected significant deviations from the average across countries. Among

[3] We also use geolocation tools to verify the location of each participating worker.

[4] India is over-represented in the raw results presented here due to participation bias. We discuss this in detail in Section 5.

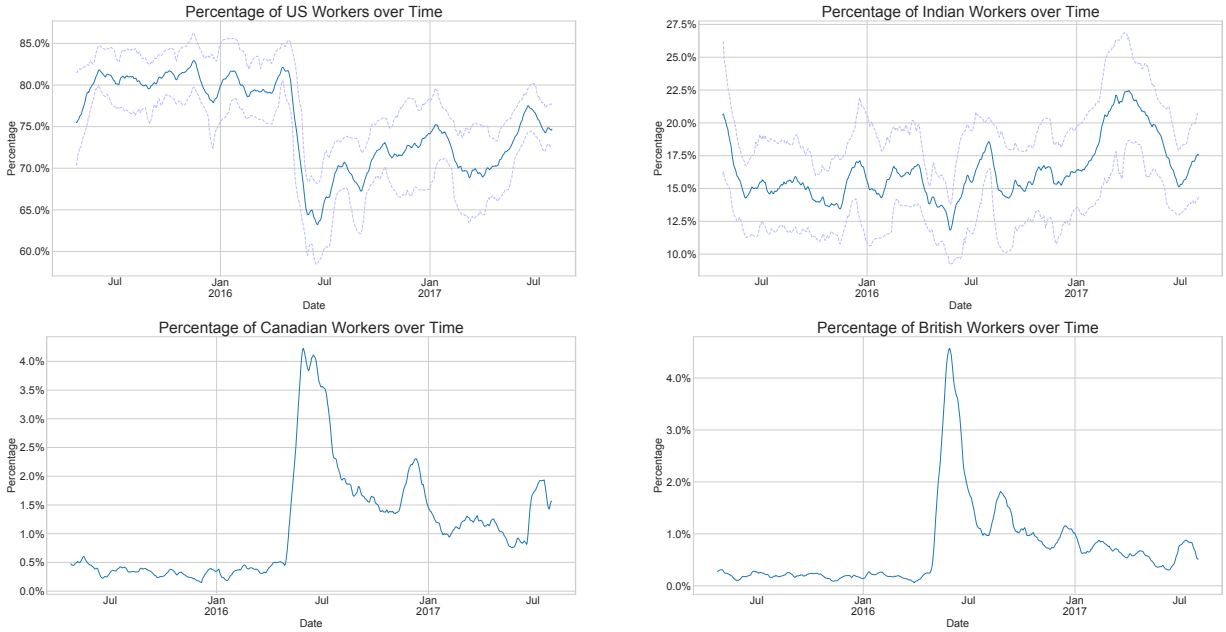


Figure 4: The average percentage of workers from various countries, over time. For US and India, we also show the the 25% and 75% percentiles to illustrate the typical noise levels in our survey measurements.

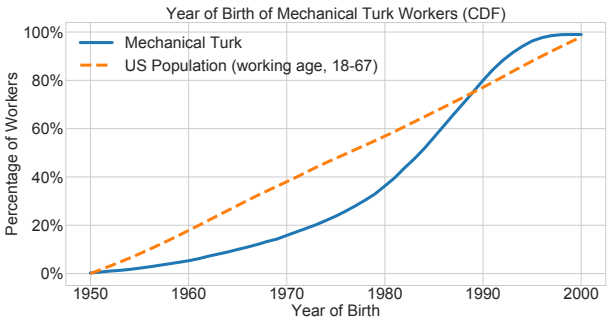


Figure 5: Year of birth.

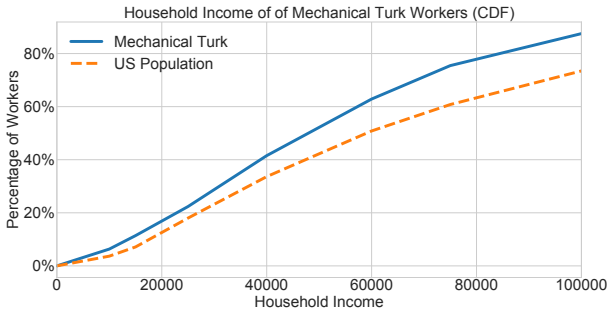


Figure 6: Household Income.

US workers, we observed that females constitute 55% of the participants, while for most other countries we observed the opposite bias. Figure 3 shows the breakdown among the top-6 countries.

3.2.3 Age: The population of MTurk workers tends to be younger than the overall population (Figure 5): 20% of the MTurk workers are born after 1990, 60% are born after 1980, and 80% are born after 1970. When compared to the *working age, adult* US population, 20% of the adult population is born after 1990, 40% are born after 1980, and 60% of the adult population is born after 1970. When comparing the population of US workers vs India, we also observe that Indian workers are a bit younger than their US counterparts.

3.2.4 Marital Status and Household Size: In terms of marital status, we observed that 40% of the workers report being single and 42% report being married; another 10% reports cohabitating, 5% being divorced, and 3% being engaged. When examining the combination of household sizes and marital status, 15% of the workers are single and live in a household of one, while the categories of singles living in household sizes of two, three, four, and five-plus account, each, for 6% of the workers. The corresponding numbers of married workers are household sizes of two (11%), three (11%), four (11%), and five-plus (8%).

3.2.5 Income: In Figure 6, we show the income distribution for workers based in the US, compared to the income distribution for the general US population. We observe that MTurk workers have household incomes that are below the average US population. For example, the median household income for the US is around $57K, while for MTurk workers the median household income is around $47K. Similarly, while 26.5% of US households make more than $100K per year, for MTurk workers this percentage falls at 12.5%.

We presented raw data statistics for the MTurk worker population. This presentation is restricted to percentages of the population, as opposed to an absolute number of workers. In the next section, we show how to estimate the absolute number of workers in the marketplace by using techniques from the field of ecology. Later

on, we demonstrate how to measure the selection and participation biases and explain how to ensure that our results are representative of the overall worker population, and not biased towards workers that are likely to participate in this particular task.

4 ESTIMATING POPULATION SIZE

In this section, we focus on estimating the *number* of workers on Amazon Mechanical Turk. We build on techniques from the field of ecology, specifically on a family of techniques commonly referred to as "*capture-recapture*." We start with simple techniques, and we proceed to more advanced ones, by identifying the various assumptions of the basic models that result in inconsistencies. Specifically, we highlight that for online populations some of the techniques from ecology lead to highly inaccurate results because the *MTurk worker propensity of participating in online surveys* follows a highly skewed distribution, while in ecology the probability of capturing an animal is relatively more homogeneous.

4.1 Two Occasion, Closed Population Model

The simplest capture-recapture model is the "two-occasion" model, which results in the *Lincoln estimator* [20]. In the two-occasion model there are two stages:

- The **capture** phase, in which a set of N_1 animals are captured, marked, then released. In our scenario, the capture phase is a 30-day period and we consider as "captured" the workers that participate in the survey during that period.
- The **re-capture** phase, where a set of N_2 animals are captured. Among these N_2 animals, $M = N_1 \cap N_2$ were marked in the capture phase. In our scenario, the recapture phase is a different 30-day period and we consider as "recaptured" the M workers that participate in both surveys.

Given $n_1 = |N_1|$, $n_2 = |N_2|$, and $m = |M|$, we estimate the total number of workers N. If there are N workers available during the recapture period, and n_1 among them are marked from the capture period then, when we sample a single worker, the probability that this worker was previously marked is n_1/N. When we sample n_2 workers, the expected number of marked workers is $\frac{n_1 \cdot n_2}{N}$. Since we counted m marked workers, we set $m = \frac{n_1 \cdot n_2}{N}$ and therefore we get the estimator:

$$N = \frac{n_1 \cdot n_2}{m} \tag{1}$$

Example 4.1. Consider the following simple, two-occasion measurement. During the first month of our experiment, in April 2015, our survey was completed by 2812 unique workers.[5]. During the third month of our experiment, in June 2015, our survey was completed 2828 times. Among the 2828 MTurk workers who completed the survey during June 2015, we have 593 MTurk workers who also completed the survey during April 2015. This means that approximately $593/2828 \approx 21\%$ of the recaptured workers were marked. Therefore, according to the Lincoln estimator, our population estimate is $N = \frac{2812 \cdot 2828}{593} = 13410$ workers. □

The Lincoln estimator relies on two main assumptions:

- The *closed population* assumption: None of the workers from the capture occasion have left the MTurk platform before the recapture occasion (in ecology, this corresponds to no deaths or immigration in the animal population). This ensures that, during the recapture occasion, all n_1 marked MTurk workers are present in the population and can be recaptured.
- The *equal catchability* assumption: The probability of capturing each worker is uniform across the population of MTurk workers.

In Figure 7, we present the population size estimations generated using the results of 100,000 randomly selected two-occasion measurements: each measurement takes two random 30-day periods, that have at least *diff* > 60 days between the beginning of the first and the second time periods. The *diff* > 60 ensures no "interference" between the two sample periods, as we allow a worker to participate only once every 30 days.

The results highlight a violation of the closed population assumption for the Lincoln estimator: In Figure 7a, we see that the population estimates are different as the *diff* value increases (the colors encode the *diff* value), while the population estimates remain relatively stable when the *diff* value remains fixed (e.g., follows the bands with the same color). While it would not be surprising to see an increase in population over time, such an increase should have been observed even when keeping *diff* fixed, something that we do not see in our results. Figure 7b is a variation of Figure 7a with *diff* in the x-axis, and colors encoding the beginning of the recapture period. Figure 7b makes more explicit the increase of the population estimates as the *diff* increases, for the *same* recapture period. This increase means that the overlap between two samples decreases as the time difference between samples increases, highlighting that marked workers depart from the population, and therefore violating the closed population assumption.

4.2 Open Population Model

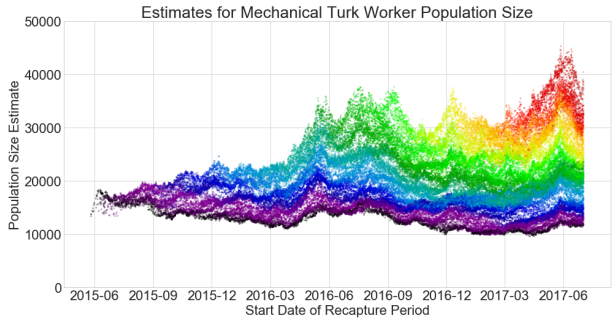
The assumption of the closed population model was a very restrictive one in ecology, and multiple models have been developed that allow for arrivals and departures of animals in the population (e.g., see the Jolly-Seber and the Cormack-Jolly-Seber models [10, 18, 29]). In our setting, an open population model allows for the arrival of new workers on the platform, and departure of workers permanently from the platform.

To move from the closed capture-recapture model to the open one we remove the assumption of *no departure* and instead we introduce the notion of *lifetime* of MTurk workers in the platform. For simplicity, we assume that the survival probability of MTurk workers is constant over time and across the population. In this case, the survival probability $S(t)$ is defined by the survival function $S(t) = \exp(-\lambda \cdot t)$ where t is the time (in days) since the last time the survey is answered by a particular MTurk worker; and λ is the decay rate, which corresponds to the rate in which workers depart from the platform.

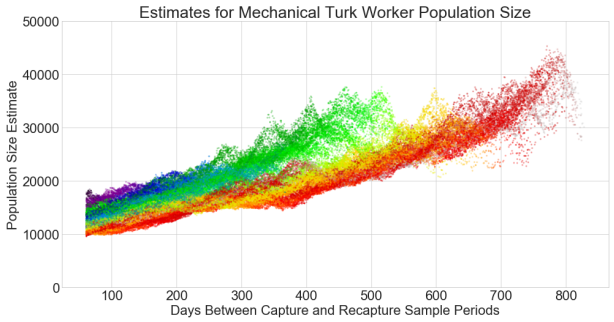
Given the new assumption about the MTurk population dynamics we can refine the estimator described in Section 4.1. We have a capture period at time $d - t$ and a recapture period at time d.

After the n_{d-t} workers are captured and marked during the first part of the two-occasion experiment, there is a $S(t)$ probability for each MTurk worker to survive until a recapture part of the two-occasion experiment that is t days away. On expectation, out of n_{d-t}

[5]Note that, by design, a worker can complete our survey only once every 30 days.



(a) Population estimates as a function of the start time of the recapture period. The colors, using a spectral colormap, encode the time distance between the two samples. Red colors correspond to large time distance and black corresponds to small time distance.



(b) Population estimates as a function of the time distance between the two samples. The colors, using a spectral colormap, encode the time corresponding to the beginning of the recapture period. Red encodes recapture periods close to July 2017 and black those close to May 2015.

Figure 7: Estimates of the size of the MTurk worker population, using the Lincoln estimator. Each data point corresponds to a comparison of two sample periods. Note that the population estimates increase as the distance between sample periods increases, signaling violation of the closed population assumption.

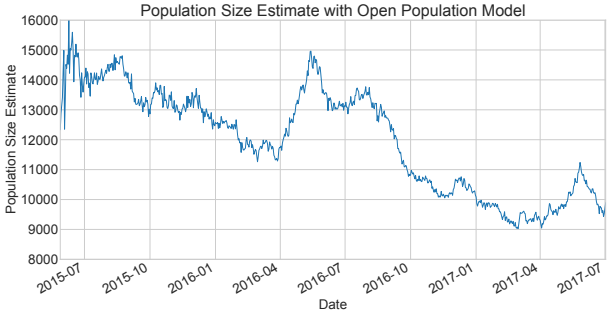


Figure 8: Estimated MTurk population using Open Population Model.

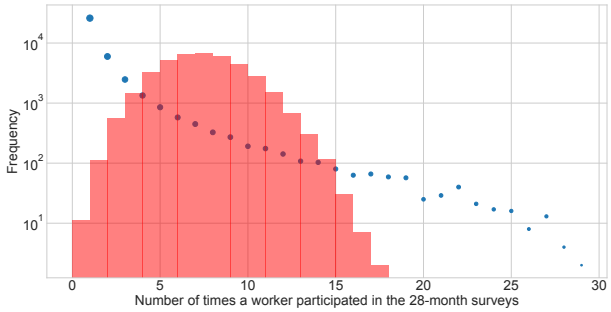


Figure 9: The dots shows the actual frequency of seeing workers in our samples during the study. The shaded histogram shows the expected distribution under the assumption of equal catchability, indicating the equal catchability assumption is violated in our data.

MTurk workers marked during the capture occasion, $n_{d-t} \cdot S(t)$ workers are still present in the population during the recapture occasion at time d. Therefore, the probability of picking a marked worker when picking a random MTurk worker is $\frac{n_{d-t} \cdot S(t)}{N_d}$, with N_d being the population during the recapture period. Therefore, when sampling n_d MTurk workers during the recapture, on average we expect to see $m_{d,t} = S(t) \cdot \frac{n_{d-t} \cdot n_d}{N_d}$ marker MTurk workers. Thus, estimate of population size N_d becomes:

$$N_d = \exp(-\lambda \cdot t) \cdot \frac{n_{d-t} \cdot n_d}{m_{d,t}} \quad (2)$$

In the equation above the unknowns are the N_d values and the λ, which can be estimated through a simple OLS regression after we take the logs.

Based on the model above, our estimated half-life of the MTurk worker population is 404 days, which means that roughly half of the workers leave the platform every year. Figure 8 shows the population estimates N_d over time. The results indicate that the average worker population is around 12K workers. This result is qualitatively similar with the estimate from [34], which calculates that there are around 7.3K workers available on Mechanical Turk for academic experiments.

However, this number is a severe underestimate of the actual number of workers in the platform, due to the varying propensities of users to participate. We can detect this by analyzing the frequency of observing users in our samples. If we assume equal probability of capturing a worker in each sample, the probability of capturing a user each month is about 0.25 (3K users sampled per month, out of 12K available users). Therefore the distribution of frequencies of participation should roughly follow a binomial distribution with $p = 0.25$ and $n = 28$ samples. Figure 9 shows the observed frequencies in our study, contrasted with the expected binomial distribution. We observe that we have very significant deviations at both ends of the frequency spectrum: We see too many workers participating only one, and we also see an unusual number of workers participating "too many" times. This is a signal that the equal catchability assumption is violated, and we need to use models that account for this heterogeneity. Stewart et al. [34] dealt with the heterogeneity in participation by eliminating the heavy-hitter workers from the experiment and assuming that everyone else has an equal probability of participating. Unfortunately, this shortcut is incorrect, and leads to significant underestimates, especially in an

online environment such as Amazon Mechanical Turk. We demonstrate and quantify the problem next, and we present techniques that deal with the issue.

4.3 Accounting for Propensity of Participation

An assumption of the models presented so far is that the probability of a worker participating in the survey is equal among all workers (the "equal catchability" assumption). This assumption is often incorrect in ecology (e.g., females during nesting times do not move and are unlikely to be captured) but it is a much bigger problem when applied to online environments. In many online environments, the activity levels of different users tend to follow heavy-tailed, power-law-type distributions. As we demonstrate below, when faced with such user behaviors, the estimators that assume equal catchability generate severe population underestimates.

4.3.1 Propensity in Two Occasion Model. We now take the simple two-occasion model, but allow users to have different levels of activity. Formally, we endow each user with a propensity parameter a, which captures the probability that a user will be *active and willing to participate in the survey*, at any given time. The propensity parameter is a random variable with a prior distribution $p(a)$. If we have a population of N users (not all of them *active* at the same time), each of them with a propensity a_i, then each user participates in the survey with a probability proportional to a_i and therefore the probability that we "capture" worker i in a sample of n_1 workers is:

$$P(capture|a_i) = 1 - (1 - \frac{a_i}{\sum_j^N a_j})^{n_1} \approx \frac{n_1}{N} \cdot \frac{a_i}{E[a]} \quad (3)$$

Given the propensity a_i, the probability of re-capturing in a second sample with n_2 workers is *conditionally independent given* a_i:

$$P(capture, recapture|a_i) = \frac{n_1 \cdot a_i}{N \cdot E[a]} \cdot \frac{n_2 \cdot a_i}{N \cdot E[a]} = \frac{n_1 \cdot n_2 \cdot a_i^2}{N^2 \cdot E[a]^2}$$

By integrating over the population with N workers and a distribution of propensities $p(a)$, we get that the expected intersection between two samples of size n_1 and n_2 is:

$$m = N \int \frac{n_1 n_2 a^2}{N^2 E[a]^2} p(a) da = \frac{n_1 \cdot n_2}{N} \cdot \frac{E[a^2]}{E[a]^2} = \frac{n_1 \cdot n_2}{N} \cdot \left(1 + \frac{Var[a]}{E[a]^2}\right)$$

The baseline Lincoln estimator is $\frac{n_1 \cdot n_2}{N}$, so the difference is the $\frac{Var[a]}{E[a]^2}$ factor. When the factor $\frac{Var[a]}{E[a]^2}$ is small then the Lincoln estimator is close to the true value. However, when the variance is high compared to the average propensity, the underestimates can be significant. Intuitively, this happens because the users with the highest propensity are significantly more likely to appear in both the capture and the recapture phase. Therefore the intersection of the sample tends to be dominated by the highly-returning users; the higher the variance in propensity, the more significant the effect.

Example 4.2. Consider the following three cases, where the *average* propensity to participate is $E[a] = 0.5$, i.e., workers on average have 50% probability of being active and willing to participate in the survey.

- *Constant a:* In the simplest case, where *all* workers have identical propensity, then the Lincoln estimator gives accurate results.
- *Uniformly distributed a:* Consider the case where a is uniformly distributed in the $[0, 1]$ interval ($E[a] = 0.5$). In this case, $Var[a] = 0.08$, $\frac{Var[a]}{E[a]^2} = 0.33$ and the Lincoln estimator will be $0.75 \cdot N$ when the true population is N.
- *Beta-distributed a:* Consider the case where a is distributed following the $Beta(0.001, 0.001)$ distribution ($E[a] = 0.5$). In this case $Var[a] = 0.245$, and the the Lincoln estimator will be $0.5 \cdot N$ when the true population is N.

The cased above, all with mean propensity $E[a] = 0.5$, illustrate that increased variance can lead to significant population underestimates. The underestimates are much more extreme under heavily skewed distributions that are common in online environments. For example, if propensity follows a $Beta(0.05, 20)$ distribution, then the underestimate is a factor of 20! Such degree of heterogeneity of catchability is rare in ecology, but common online; as we will see later, such a skewed Beta distribution is a good fit for the propensities observed in the MTurk population.

Our analysis for the simple two-occasion model shows that eliminating the heavy hitters from our data (as done by Stewart et al. [34]) is not a solution, as even in the simple two-occasion model (where we cannot eliminate any worker), we still observe significant underestimates. Unfortunately, simple two-occasion models are also not sufficient to estimate the characteristics of the underlying propensity distribution $p(a)$. We show next how to use a multiple-capture setting to estimate $p(a)$ and get correct population estimates.

4.3.2 Modeling Propensity with Multiple Captures. Below, we introduce a model that leverages the multiple samples of our study to measure the underlying propensity distribution.

Our model assumes that the propensity $P(capture|a_i)$ (Equation 3) follows a Beta distribution $B(\alpha, \beta)$, which is very flexible and is commonly used to model random variables that take values in a closed interval.

In addition, we assume a super-population [27] of N^* workers available over the span of the study. This assumption simplifies away the arrival and departure variables of workers and focuses on estimating an aggregated population size which can be used to derive finer-grained estimates.

We model our capture-recapture data collection by sampling n times from an underlying super-population where each worker has a capture probability $P(capture|a_i)$ that follows a $B(\alpha, \beta)$ distribution, then the probability of seeing a worker k times follows a *Beta-Binomial distribution*, with the following pdf:

$$f(k|n, \alpha, \beta) = \binom{n}{k} \frac{B(k + \alpha, n - k + \beta)}{B(\alpha, \beta)} \quad (4)$$

We can estimate the parameters of the Beta-Binomial distribution by using either the method of moments, or by using the maximum likelihood estimation approach. For our data, through MLE estimation, we obtain $\alpha = 0.29$ and $\beta = 20.9$.

The next step is to estimate the number of workers N^* from this model. As described in [9], we focus on estimating the probability a worker having frequency $k = 0$. In our data, we observe all the S sampled workers that have frequencies $k \geq 1$. We know that:

$$S = N^* \cdot \sum_{k=1}^{n} f(k|n, \alpha, \beta) = N^* \left(1 - f(0|n, \alpha, \beta)\right)$$

In our experiments, we sampled a total of $S = 39,461$ distinct workers with frequencies $k \geq 1$. For $n = 28$, $\alpha = 0.29$ and $\beta = 20.9$, we have that $f(0|n, \alpha, \beta) = 0.7793$, and therefore the estimate for the super-population of workers N^* is:

$$N^* = \frac{S}{1 - f(0|n, \alpha, \beta)} = \frac{39,461}{1 - 0.7793} \approx 178,800 \text{ workers}$$

Similarly, we can get additional estimates by examining the number of workers that appeared once (N_1^*), twice (N_2^*), thrice, and so on:

$$N^* = \frac{N_1^*}{f(1|n, \alpha, \beta)}, \; N^* = \frac{N_2^*}{f(2|n, \alpha, \beta)}, \; N^* = \frac{N_3^*}{f(3|n, \alpha, \beta)}, \ldots$$

Our technique relies on the assumption of a Beta distribution for the probabilities of capture. To check whether our technique produces population estimates that are reasonable, we also used the technique of Chao [9, Eq. 10], which yields a lower bound for the size of the population. By using our data, the Chao approach gives a *lower bound* of 97,579 workers, which is compatible with our results.

Similarly, we also experimented with a technique from Pledger et al. [25] that uses finite mixtures to model heterogeneity in an open population. The technique operates with the assumption that the population is split into a predefined set of underlying clusters, each having unique survival and capture characteristics. Our numerical estimates were sensitive to the choice of the number of clusters (the technique did not converge when we used larger number of clusters) nonetheless, it estimated the size of the super-population to 125K-130K using 3 clusters with highly heterogeneous capture probabilities across the three groups, and with a survival probability similar to the one we estimated in Section 4.2.

Having estimated the distribution of propensities, we now move to our final question: Are the demographics estimates, reported in Section 3, affected by selection bias? We address the issue next.

5 DEMOGRAPHICS AND SELECTION BIASES

A natural concern when reporting the results of a survey is that the survey answers may be biased due to the type of people that are willing to participate in the survey. Our setting, where we survey workers multiple times over a long period of time, allows us to infer the propensity of participation for the workers. This, in turn, allows us to examine whether the demographic variables that we measure are correlated with the propensity to participate, and therefore need to be adjusted.

In our model, the propensities, and hence the probability of capture, are distributed according to a Beta distribution. Therefore, to examine if there is a correlation of the propensities with the demographic variables, we use a *Beta regression* [11, 31], which is explicitly designed to handle dependent variables that are Beta distributed and exhibit heteroskedasticity and skewness.

We examined the correlation of the six demographic variables, described in Section 3.1, with the observed propensity dependent variable computed from the frequency of participation over the course of the study periods. Since the computed propensity includes values of 1.0, indicating workers who participate every month since the beginning of the study, we perform an additional transformation [31] to shift the distribution into the open interval (0,1) using $(propensity \cdot (n-1) + 0.5)/n$, where $n = 28$ is the sample size. By

Dependent variable: Propensity			
(Intercept)	−3.44	***	(-3.53, 3.53)
Age	0.006	***	(0.005, 0.007)
Location_Country = India	0.383	***	(0.291, 0.475)
Marital_Status = Divorced	−0.064	**	(-0.121, -0.006)
Constant(ϕ)	−3.443	***	(-3.553, -3.333)
Observations	39,453		
Pseudo R^2	0.029		
Log Likelihood	208,080.100		

Table 1: Beta Regression Results for Demographic Variables that affect propensity. Note: We only include parameter estimates when: **p<0.05; ***p<0.01 (95% CI in parentheses).

inspecting the distribution of the computed propensities, we note that 88% of the participants have $a < 0.1$ propensity.

We fit a mean-only beta regression model with the logit link function. Since we have a large number of categorical variables as independent regressors, and due to space constraints, we present only on variables with statistically significant effects at the $p < 0.05$ level. (See Table 1.) We did not detect any statistically significant correlation with gender, household income, and household size.

We detected a *statistically* significant effect of age (year of birth), with older workers having a slightly higher propensity to participate; while the effect was statistically significant, the magnitude of the effect was small ($\beta_{age} = 0.006$). By exponentiating the Beta coefficient of the parameter *Age* and fixing the other parameters we can obtain the rate of increase in the odds of the mean propensity for each unit change in *Age* following $\exp(\beta_{age} \cdot Age)$. Analytically, this indicates that the error in Figure 5 would be at most 10%, in *relevant* (not absolute) terms, hinting to a *slightly* younger audience on the platform. A similarly statistical significant, but weak, effect is observed for divorced workers who are slightly underrepresented by around 6%, in relative terms.

The most significant effect was detected for the independent variable *country of origin* where we observe a significant increase in propensity for workers coming from India with ($\beta_{India} = 0.38$). Here, the odds of the mean propensity are $\exp(0.38) = 1.46$ times higher if the worker is from India. This indicates that the Indian workers are over-represented in the sample and the percentage of Indian workers in Figure 1 are inflated by 46%; the real percentages are to be closer to 10%-14% and not 15%-20%.

6 DISCUSSION AND FUTURE DIRECTIONS

We presented an analysis of population estimation for Amazon Mechanical Turk using capture-recapture techniques from ecology. We demonstrated that using simple techniques, without understanding the limitations and assumptions of these models can lead to inconsistent and inaccurate results. Our analysis indicates that the population of the Mechanical Turk worker population remains relatively stable over time and it consists of at least 100K-200K workers, and possibly more. We also see that the population of the workers renews over time, with an average half-life of 400 days; this means that there are tens of thousands of new workers arriving on the platform every year.

For employers that rely on having access to a large number of workers, our study indicates that the MTurk platform *does provide access to hundreds of thousands of workers.* Out of these, and

based on our empirically estimated propensity distribution, we can compute the number of available workers at any given time, which we interpret from calculating the expected propensity times the super-population size as follows: $E[a] \cdot N^* = \frac{\alpha}{\alpha+\beta} \cdot N^* = \frac{0.29}{0.29+20.9} \cdot 178,800 \approx 2,450$ *workers*. This is in sharp contrast to the conclusions of [34], which estimated that there are only 7,500 workers available to experimenters for social science research. That incorrect estimate was due to the incorrect assumption that all workers have the same degree of activity in the platform, and the same propensity to participate in a certain type of task.

In the future, we plan to consider models that allow time-varying survival and propensities. Our current models consider both to be stable over time, and our attempts to allow time-varying parameters did not indicate significant variations over time for our data. It would be also interesting to collect participation data for other long-running tasks, and examine cross-task propensities: this will allow us to understand whether propensity to participate is a global trait across tasks, or whether workers have task-specific propensities.

We would also like to use our technique to study other online environments: By monitoring user activity on various platforms (e.g., comments in threads) we can infer the total number of users in these communities, or even identify the *potential* number of users by estimating the number of users that have made no visible contribution so far. As Trushkowsky et al. [36] have shown, it is possible to use such models also in fields like information extraction and online databases, where population sizes do not necessarily correspond to humans but are important to estimate for a variety of applications. By learning the state-of-the-art from the techniques in ecology, we can adapt these models for application in domains where the behavior is different than the behavior of animals, but the fundamental ideas behind the models remain useful.

REFERENCES

[1] Steven C. Amstrup, Trent L. McDonald, and Bryan F. J. Manly. 2010. *Handbook of Capture-Recapture Analysis.* Princeton University Press, Princeton, NJ.

[2] Joanna J. Arch and Alaina L. Carr. 2016. Using Mechanical Turk for research on cancer survivors. *Psycho-Oncology* Forthcoming (2016), –. PON-15-0731.R1.

[3] Antonio A. Arechar, Simon Gächter, and Lucas Molleman. 2017. Conducting interactive experiments online. *Experimental Economics* Forthcoming (09 May 2017), 1–33.

[4] K. P. Burnham and W. S. Overton. 1978. Estimation of the Size of a Closed Population when Capture Probabilities vary Among Animals. *Biometrika* 65, 3 (1978), 625–633.

[5] Chris Callison-Burch and Mark Dredze. 2010. Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1–12.

[6] Krista Casler, Lydia Bickel, and Elizabeth Hackett. 2013. Separate but Equal? A Comparison of Participants and Data Gathered via Amazon's MTurk, Social Media, and Face-to-Face Behavioral Testing. *Computers in Human Behavior* 29 (2013), 2156–2160.

[7] Jesse Chandler, Gabriele Paolacci, and Pam Mueller. 2013. Risks and Rewards of Crowdsourcing Marketplaces. In *Handbook of Human Computation*. Springer, New York, 377–392.

[8] Jesse Chandler and Danielle Shapiro. 2016. Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annual Review of Clinical Psychology* 12 (2016), 35–81.

[9] Anne Chao. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43, 4 (Dec. 1987), 783–791.

[10] R. M. Cormack. 1964. Estimates of Survival From the Sighting of Marked Animals. *Biometrics* 51 (1964), 429–438.

[11] Francisco Cribari-Neto and Achim Zeileis. 2010. Beta Regression in R. *Journal of Statistical Software, Articles* 34, 2 (2010), 1–24.

[12] Damien L. Crone and Lisa A. Williams. 2017. Crowdsourcing participants for psychological research in Australia: A test of Microworkers. *Australian Psychological Society* 69 (2017), 39–47. Issue 1.

[13] Joost de Winter, Miltos Kyriakidis, Dimitra Dodou, and Riender Happee. 2015. Using CrowdFlower to Study the Relationship between Self-reported Violations and Traffic Accidents. *Procedia Manufacturing* 3 (2015), 2518–2525.

[14] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, ACM, New York, NY, USA, 238–247.

[15] Karën Fort, Gilles Adda, and K. Bretonnel Cohen. 2011. Amazon Mechanical Turk: Gold Mine or Coal Mine? *Computational Linguistics* 37 (2011), 413 – 420. Issue 2.

[16] Ilka H. Gleibs. 2017. Are all research fields equal? Rethinking practice for the use of data from crowd-sourcing market places. *Behavior Research Methods* 49 (August 2017), 1333–1342. Issue 4.

[17] Panagiotis Ipeirotis. 2010. *Demographics of Mechanical Turk.* Technical Report CeDER-10-01. New York University.

[18] G. M. Jolly. 1965. Explicit Estimates from Capture-Recapture Data With Both Death and Immigration- Stochastic Model. *Biometrika* 52 (1965), 225–247.

[19] Jeremy Kees, Christopher Berry, Scot Burton, and Kim Sheehan. 2017. An Analysis of Data Quality: Professional Panels, Student Subject Pools, and Amazon's Mechanical Turk. *Journal of Advertising* 46, 1 (2017), 141–155.

[20] Frederick Charles Lincoln. 1930. *Calculating waterfowl abundance on the basis of banding returns.* US Dept. of Agriculture, Washington, DC.

[21] Jianguo Lu and Dingding Li. 2010. Estimating deep web data source size by capture-recapture method. *Information Retrieval* 13, 1 (February 2010), 70–95.

[22] Winter Mason and Duncan J. Watts. 2009. Financial Incentives and the "Performance of Crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09)*. ACM, New York, NY, USA, 77–85.

[23] Joshua D. Miller, Michael Crowe, Brandon Weiss, Jessica L. Maples-Keller, and Donald R. Lynam. 2017. Using online, crowdsourcing platforms for data collection in personality disorder research: The example of Amazon's Mechanical Turk. *Personality Disorders: Theory, Research, and Treatment* 8, 1 (2017), 26–34.

[24] Shirley Pledger, Kenneth H. Pollock, and James L. Norris. 2003. Open Capture-Recapture Models with Heterogeneity: I. Cormack-Jolly-Seber Model. *Biometrics* 59, 4 (Dec. 2003), 786–794.

[25] Shirley Pledger, Kenneth H. Pollock, and James L. Norris. 2010. Open Capture-Recapture Models with Heterogeneity: II. Jolly-Seber Model. *Biometrics* 66, 3 (September 2010), 883–890.

[26] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems (CHI EA '10)*. ACM, New York, NY, USA, 2863–2872.

[27] Carl James Schwarz and A. Neil Arnason. 1996. A General Methodology for the Analysis of Capture-Recapture Experiments in Open Populations. *Biometrics* 52, 3 (1996), 860–873.

[28] George Seber. 1982. *The Estimation of Animal Abundance and Related Parameters.* Charles Griffin, London, UK.

[29] G. A. F. Seber. 1964. A Note on the Multiple Recapture Census. *Biometrika* 52 (1964), 249–259.

[30] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. ACM, New York, NY, USA, 614–622.

[31] Michael Smithson and Jay Verkuilen. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods* 11 (2006), 54–71. Issue 1.

[32] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and Fast—but is It Good?: Evaluating Non-expert Annotations for Natural Language Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 254–263.

[33] Vanessa V. Sochat, Ian W. Eisenberg, A. Zeynep Enkavi, Jamie Li, Patrick G. Bissett, and Russell A. Poldrack. 2016. The Experiment Factory: Standardizing Behavioral Experiments. *Frontiers in Psychology* 7 (2016), 610.

[34] Neil Stewart, Christoph Ungemach, Adam J. L. Harris, Daniel M. Bartels, Ben R. Newell, Gabriele Paolaccik, and Jesse Chandler. 2015. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making* 10, 5 (2015), 479–491.

[35] Keela S. Thomson and Daniel M. Oppenheimer. 2016. Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making* 11, 1 (2016), 99–113.

[36] Beth Trushkowsky, Tim Kraska, Michael J. Franklin, and Purnamrita Sarkar. 2016. Answering Enumeration Queries with the Crowd. *Commun. ACM* 59, 1 (2016), 118–127.