Article

SAGE Open
April-June 2017: 1–15

DOI: 10.1177/2158244017712774
journals.sagepub.com/home/sgo
SAGE

Intertemporal Differences Among MTurk Workers: Time-Based Sample Variations and Implications for Online Data Collection

Logan S. Casey[1], Jesse Chandler[2,3], Adam Seth Levine[4], Andrew Proctor[5], and Dara Z. Strolovitch[5]

Abstract

The online labor market Amazon Mechanical Turk (MTurk) is an increasingly popular source of respondents for social science research. A growing body of research has examined the demographic composition of MTurk workers as compared with that of other populations. While these comparisons have revealed the ways in which MTurk workers are and are not representative of the general population, variations among samples drawn from MTurk have received less attention. This article focuses on whether MTurk sample composition varies as a function of time. Specifically, we examine whether demographic characteristics vary by (a) time of day, (b) day of week, and serial position (i.e., earlier or later in data collection), both (c) across the entire data collection and (d) within specific batches. We find that day of week differences are minimal, but that time of day and serial position are associated with small but important variations in demographic composition. This demonstrates that MTurk samples cannot be presumed identical across different studies, potentially affecting reliability, validity, and efforts to reproduce findings.

Keywords

political methodology, political science, social sciences, politics and social sciences, research methods, data collection, research methodology and design, reliability and validity, political behavior/psychology, psychology

Background

Amazon Mechanical Turk (MTurk) is an online labor market in which people ("requesters") requiring the completion of small tasks ("Human Intelligence Tasks" [HITs]) are matched with people willing to do them ("workers"). MTurk has become a popular data collection tool among social science researchers: In 2015, the 300 most influential social science journals (with impact factors greater than 2.5, according to Thomson-Reuters InCites) published more than 500 articles that relied on MTurk data in full or in part (Chandler & Shapiro, 2016).

Reflecting the popularity of MTurk, considerable effort has been invested in evaluating data collected from it, with particular emphasis on documenting the demographic and psychological characteristics of its population, the quality of respondent data, and the methodological limitations of the platform. As a result, MTurk workers have become one of the most thoroughly studied convenience samples currently available to researchers (for a review, see Chandler & Shapiro, 2016), and researchers have learned a great deal about the ways in which MTurk respondents are and are not similar to the general population. There are reasons to suspect, however, that there are also important variations between different samples drawn from MTurk, and these variations have received far less attention. This article addresses this question, using data from a study of approximately 10,000 MTurk workers to examine whether sample composition varies as a function of the time that it is collected.

We begin by reviewing what extant research reveals about the demographic composition of the MTurk worker pool. Then, we describe the methods and measures that we use in our study, after which we present the results of our analyses,

[1]Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]University of Michigan, Ann Arbor, USA
[3]Mathematica Policy Research, Ann Arbor, MI, USA
[4]Cornell University, Ithaca, NY, USA
[5]Princeton University, NJ, USA

Corresponding Author:
Jesse Chandler, Research Center for Group Dynamics, Institute for Social Research, University of Michigan, 426 Thompson St., Ann Arbor, MI 48105, USA.
Email: jjchandl@umich.edu

which include a demographic description of the largest sample of MTurk workers we are aware of and an exploration of whether the demographic characteristics of MTurk respondent samples vary across day and time and earlier versus later in the data collection. We conclude with a discussion about the implications of the temporal variations we uncover for researchers using MTurk (and online data collection more generally).

How Representative of the General Population Are Samples of MTurk Workers?

The demographic characteristics of samples drawn from MTurk populations have been extensively studied. These studies show that most MTurk workers live in the United States and India (Paolacci, Chandler, & Ipeirotis, 2010), that U.S. MTurk workers are more diverse than many other convenience samples, and that they are not representative of the population as a whole (Paolacci & Chandler, 2014). However, while scholars caution that MTurk samples are typically less representative than commercial web panels that make explicit efforts to provide representative samples (Berinsky, Huber, & Lenz, 2012; Mullinix, Leeper, Druckman, & Freese, 2015; Weinberg, Freese, & McElhattan, 2014), they also agree that MTurk samples are more diverse than student samples or community samples recruited from college towns (Berinsky et al., 2012; Krupnikov & Levine, 2014).

Differences between the U.S. MTurk population and the U.S. general population parallel differences between samples recruited through other online methods and the U.S. population (Casler, Bickel, & Hackett, 2013; Hillygus, Jackson, & Young, 2014; Paolacci & Chandler, 2014). Most significantly, MTurk workers are typically younger than the general population (Berinsky et al., 2012; Paolacci et al., 2010), have more years of formal education, and are more liberal (Berinsky et al., 2012; Mullinix et al., 2015). MTurk workers are less likely to be married (Berinsky et al., 2012; Shapiro, Chandler, & Mueller, 2013), and more likely to identify as lesbian, gay, or bisexual (LGB; Corrigan, Bink, Fokuo, & Schmidt, 2015; Reidy, Berke, Gentile, and Zeichner, 2014; Shapiro et al., 2013). MTurk workers also tend to report lower personal incomes and are more likely to be unemployed or underemployed than members of general population (Corrigan et al., 2015; Shapiro et al., 2013). Whites and Asian Americans are overrepresented within MTurk samples, while Latinos and African Americans are underrepresented (Berinsky et al., 2012).

Are Samples of MTurk Workers Representative of MTurk Workers?

While the forgoing research makes clear that the U.S. MTurk population is not representative of the U.S. population as a whole, there are also reasons to suspect that samples recruited from MTurk are themselves not representative of the *MTurk population* as a whole. Different studies occasionally observe substantially different demographic characteristics. For example, the proportion of female respondents differed by about 10% across two studies that each recruited several thousand participants (Chandler & Shapiro, 2016).

There are many potential causes for sampling variation across studies. Anecdotal evidence suggests that MTurk sample composition might be influenced by the fact that workers share information about available studies and that reputation effects might lead workers to gravitate toward (and to avoid) particular requesters (Chandler, Mueller, & Paolacci, 2014). Some of this variation is also surely the result of MTurk workers self-selecting into the studies that interest them (for a discussion, see Couper, 2000). Design choices that are exogenous to a study design may also inadvertently influence sample composition. The effects of such exogenous choices are of particular interest to researchers because they are both within their control and typically irrelevant to the substance of the studies themselves.

The present study focuses on the impact of intertemporal variation on sample composition across (a) time of day, (b) day of week and serial position (i.e., earlier or later in data collection), both (c) across the entire data collection and (d) within specific batches. Extant evidence about sample differences across time and day are suggestive but limited by small sample sizes. Comparing samples of about 100 participants obtained within two different studies, Komarov, Reinecke, and Gajos (2013) observed that compared with workers recruited later in the evening, workers recruited during the daytime were older, more likely to be female, and less likely to use a computer mouse to complete the survey (suggesting that they were using mobile devices). Lakkaraju (2015) compared the gender, income, education and age of 700 workers across different times and days, finding that only gender varied as a function of the day a given HIT was posted.

Variation among participants who complete a research study early or later in the data collection process (referred to here as serial position effects) has been observed in other modes of data collection, but has not been examined on MTurk. Changes in sample composition between "early" and "late" responders have been observed in mail and email surveys, in part because the easiest to contact participants tend to complete surveys earlier (for a review, see Sigman, Lewis, Yount, & Lee, 2014). In general, people of color[1] are underrepresented among early respondents, as are men (Gannon, Nothern, & Carroll, 1971; Sigman et al., 2014; Voigt, Koepsell, & Daling, 2003), younger people, and people with fewer years of formal education (Voigt et al., 2003; for a discussion, see Sigman et al., 2014).

Examinations of lab studies of college students have also shown that sample compositions can vary over time. For example, women (Ebersole et al., 2016) and students with high GPAs (Aviv, Zelenski, Rallo, & Larsen, 2002; Cooper, Baumgardner, & Strathman, 1991) are more likely than men and students with lower GPAs to participate in lab studies at

the beginning of the semester. Personality variables also influence when students complete lab studies, with participants who report that they are less extraverted, less open to experience, and more conscientious more likely to respond at the beginning of the semester.

Investigating whether samples vary over the course of a survey fielding period is critical, because researchers tend to recruit small samples for their research (Fraley & Vazire, 2014). In fact, most of the existing studies of the characteristics of MTurk workers rely on relatively small samples ($N < 500$) that capture only a small proportion of the approximately 16,000 active MTurk workers (Stewart et al., 2015). If researchers use only small samples, the samples they recruit may differ systematically from the worker pool as a whole. In addition, if researchers recruit unique workers to participate in a series of related experiments (as they should; see Chandler et al., 2014; Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015), sample composition may vary systematically across the experiments, compromising both the reliability and validity of their studies, and possibly complicating efforts to reproduce findings.

A second potential serial position effect on MTurk is differences between people who complete HITs shortly after they are posted or later on. This factor is independent from early versus late responding to the study because study data can be collected through any number of batch postings. In practice, researchers often collect data from MTurk by posting more than one batch of HITs, either to speed up data collection (data collection is faster immediately after an HIT is posted; Peer, Brandimarte, Samat, & Acquisti, 2017) or to circumvent the fee Amazon charges for a batch that recruits more than nine participants. When more (but smaller) batches are posted, the average batch will, by default, be closer to the front of the queue, which could affect sample composition for at least three reasons. First, a batch closer to the front of the queue reduces the amount of work it takes to find it, especially for workers who rely on the default sort order. Second, smaller batches might limit the number of workers who discover the survey through links on worker forums, because the link will be valid for a shorter period of time. Third, some workers use automated scripts or other tools to be alerted about the availability of new work. In this study, we post multiple batches that allow us to disentangle serial position effects within batches of posted HITs from serial position effects across the data collection as a whole.

Method

To explore whether MTurk worker demographics vary intertemporally, we crafted a brief HIT (average completion time was approximately 5 min) that contained demographic questions that are of interest to scholars across an array of disciplines.

We first posted our HIT on March 19, 2015, and data collection concluded on May 14, 2015, so it was active for a total of 56 days (or 8 weeks). We began by posting the HIT twice daily, at 3 p.m. and 10 p.m. Eastern Time (ET). After the first week, we added a third posting at 10 a.m. ET.[2]

Only U.S.-based workers with a HIT acceptance ratio (HAR) greater than 95% and who had completed at least 100 HITs were eligible to participate. We selected workers with a 95% HAR because this subsample of workers has been shown to result in higher quality data (Peer, Vosgerau, & Acquisti, 2014) and, in our experience, to be favored by researchers. We prevented workers from completing this survey more than once across the entire fielding period.

For the first 3 weeks, workers were paid US$0.25 to complete the survey. After learning that the average time to completion was roughly 5 min, we increased the pay rate to US$0.50 for the remainder of the fielding period to comply with recommended pay norms of US$0.10 per minute (see "Guidelines for Academic Requesters," 2014). By the end of the study, we had posted the HIT 162 times and sampled 9,770 unique respondents.

Measures

At the beginning of the study, we collected measures of age and the U.S. state in which respondents lived. Participants were then asked to report demographic information including their highest level of education, current employment status, and current occupation. We also asked a series of questions about their current relationship status, sexual orientation, sex assigned at birth, and current gender identity. In addition, we asked questions about household size, race and ethnicity, household income, religious denomination, how often they attend religious services, and self-perceived socioeconomic status (see Howe, Hargreaves, Ploubidis, De Stavola, & Huttly, 2011; Ravallion & Lokshin, 1999).

We also included a 10-item measure of the "Big Five" personality factors (Ten Item Personality Measure or TIPI; Gosling, Rentfrow, & Swann, 2003). The "Big Five" is among the most widely accepted taxonomy of personality traits within psychology (for a review, see John & Srivastava, 1999) and conceptualizes personality as consisting of five bipolar dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The questionnaire and other materials are available online on the Open Science Framework (osf.io/tg7h3).

Prior to completing the survey, participants were asked whether they learned about the survey on MTurk or somewhere else. Those who indicated somewhere else were asked to specify where they learned about it.

Finally, using a database of more than 100,000 HITs submitted over 3 years immediately prior to the present study (reported in Stewart et al., 2015), we were able to estimate individual workers' relative experience completing MTurk tasks. Workers with no recorded experience during the Stewart et al.'s study ($N = 4,746$) were assigned a value of one and all other workers were assigned a value equal to their

total number of previously completed HITs plus one. Although this measure does not capture total HITs a respondent has completed, it does allow us to analyze temporal variations in workers' relative levels of experience (see Chandler et al., 2015).

Results

Data Cleaning and Survey Metadata

Data collection resulted in 10,121 survey attempts, of which 169 attempts (generated by 147 workers) were identified as duplicate responses. Duplicate responses were defined as any submission from a WorkerID in excess of one. For workers with duplicate responses, the most complete response was taken. When both responses were of equal length (typically complete), the first response was taken. An additional 182 responses that came from non-U.S. IP addresses and one respondent without a WorkerID were also identified and deleted, resulting in 9,770 valid survey attempts.

Of the valid attempts, 780 (8%) were identified by Qualtrics as incomplete. A visual inspection of these responses found that 724 of these respondents answered the last question in the survey and were functionally complete. Only 56 respondents (0.6%) dropped out of the survey after providing only partial data. These partial responses were included for analysis.

Of all valid attempts, 518 (5.3%) came from an IP address shared by at least one other response. The majority of IP addresses (n = 196) contributed two responses, with 10 contributing three responses, three contributing four responses, two contributing 10 responses, one contributing 26 responses, and another 39 responses. All responses from duplicate IP addresses were left in for this analysis, as shared IP addresses do not necessarily indicate the same worker repeating a task.

For example, the 433 responses from IP addresses that contributed four or fewer responses were examined. Of these, 233 were almost certainly unique respondents from the same household: They came from people who listed the exact same household size, the same age of household members (±2 years in aggregate) and reported an age that corresponded to an age that matched an age of a person that the other respondent reported that they lived with. An additional 49 respondents were likely from the same household, reporting approximately the same total age of members (±5 years in aggregate), or who appeared to have neglected to report a household member (usually a child or much older adult).

Three of the four IP addresses that generated the most responses were servers registered to Amazon. It is likely that participants from these addresses are using either a proxy server, or an ISP hosted on Amazon Web Services. These responses varied in the time they were attempted, the specific browser and operating system configuration used, and the content of the survey responses.

Characteristics of the MTurk Sample

Tables 1 to 4 present summary data about the entire sample, about participants in the first two batches only, and for national estimates when available. The entire sample represents the largest sample of MTurk workers we are aware of, and likely measures about two thirds of the available worker population (Stewart et al., 2015). The sample size of the first two batches (N = 438) approximates a sample slightly larger than those typically used in behavioral science research (Fraley & Vazire, 2014) and is presented to enable comparisons of this study to other, typically sized data collections.

The demographic data are reported in Table 1, including information about worker experience and where they learned about the survey. Differences between this sample and the U.S. population as a whole are generally consistent with those reported in previous analyses of smaller surveys (Berinsky et al., 2012; Krupnikov & Levine, 2014; Paolacci et al., 2010; Shapiro et al., 2013). For example, the workers in our sample are younger and more likely to be white than the U.S. population as a whole. Workers residing in the Eastern Time Zone are overrepresented compared with those in other parts of the United States. This variation is likely because the times that HITs were posted aligned most closely with the times that workers in the time zone were likely to be active.

Almost all (90.9%) workers reported finding the survey on MTurk. Of the 868 workers who found the survey elsewhere, most (n = 671) named HitsWorthTurkingFor (a Reddit forum), 29 listed Hit Scraper (an automatic alerting service), and virtually all other respondents listed other MTurk discussion forums (e.g., TurkerNation).

Table 2 summarizes the socioeconomic characteristics of our sample. Respondents to our survey generally reported more years of formal education than the population as a whole. Although Americans residing in the wealthiest households are underrepresented in our data, household income was much closer to the median U.S. income than would be expected from previous measurements of individual worker income (Berinsky et al., 2012; Paolacci et al., 2010). A portion of this difference is likely due to the fact that 16.5% of the respondents in our sample are under 30 and living with someone at least 18 years older than they are, suggesting that our sample includes a substantial number of millennials with low individual income but who are living with their higher income parents.

Table 3 summarizes the relationship status and characteristics of respondents, revealing that approximately a third of respondents are married and another third are single. In addition, we find that 1.5% of our sample reports are currently engaged in a consensually nonmonogamous relationship (see Haupert, Gesselman, Moors, Fisher, & Garcia, 2016). As has been observed in other studies of other MTurk workers (Corrigan et al., 2015; Reidy et al., 2014; Shapiro et al., 2013), the proportion of lesbian, gay, and particularly bisexual

Table 1. Demographic Characteristics of Workers.

Characteristic	Total sample (*N* = 9,770)	First respondents (*N* = 438)	National estimates
Mean age	33.51 [32.3, 33.7]	33.59 [32.6. 34.58]	47.01[a]
Female	51.7% [50.8, 52.7]	46.8% [42.2, 51.6]	50.8%[b]
Transgender	0.5% [0.3, 0.6]	0.2% [0.0, 0.6]	0.6%[c]
Gender queer	0.9% [0.7, 1.1]	0.2% [0.0, 0.6]	—
Mean worker experience (Prior HITs completed)	3.67 [3.5, 3.9][d]	6.94 [5.82, 8.06][d]	—
Found HIT outside of MTurk	9%	14.8%	—
U.S. time zone			
Eastern	52.2% [51.2, 53.2]	56.2% [51.6, 60.9]	47.3%[e]
Central	25.3% [24.4, 26.2]	23.3% [19.3, 27.3]	29.0%[e]
Mountain	5.9% [5.4, 6.4]	3.9% [2.1, 5.7]	6.5%[e]
Pacific	15.9% [15.2, 16.6]	16.4% [12.9, 19.9]	16.6%[e]
Other	0.6% [0.5, 0.8]	0.2% [0.0, 0.6]	0.6%[e]
Race and ethnicity			
White/Caucasian	82.9% [82.2, 83.7]	79.5% [75.7, 83.3]	73.6%[f]
African American	8.6% [8.0, 9.2]	7.8% [5.3, 10.3]	12.6%[f]
Asian American	7.7% [7.2, 8.2]	11.2% [8.3, 14.1]	5.1%[f]
American Indian or Alaskan Native	2.1% [1.8, 2.4]	3.2% [1.6, 4.9]	0.8%[f]
Native Hawaiian or Pacific Islander	0.6% [0.4, 0.8]	1.6% [0.4, 2.8]	0.2%[f]
Other	1.3% [1.1, 1.5]	0.9% [0.0, 1.8]	4.7%[f]
Latino	5.5% [5.1, 6.0]	6.4% [4.1, 8.7]	17.1%[f]

Note. 95% CI indicated in parentheses. HITs = Human Intelligence Tasks; MTurk = Amazon Mechanical Turk; CI = confidence interval.
[a]U.S. Census Bureau (2016; mean age of adult population).
[b]U.S. Census Bureau (2011-2015).
[c]Flores, Herman, Gates, and Brown (2016).
[d]Stewart et al. (2015).
[e]U.S. Census Bureau (2016) population estimates.
[f]U.S. Census Bureau (2011-2015).

respondents is higher than it is in the U.S. population as a whole. This is likely because online populations are disproportionately young, and younger people are also more likely to identify as LGB (Gates & Newport, 2012; Moore, 2015).

Finally, summary statistics for the attitudinal and personality measures are summarized in Table 4. Consistent with earlier research, workers were more likely to identify as Democrats than are members of the general population (Berinsky et al., 2012; Mullinix et al., 2015). Relatively few workers identified as religious, a disproportionate number identified as atheists, and reported rates of church attendance were generally low. Relative to normed data obtained from a large convenience sample of Internet users (Gosling, Rentfrow, & Potter, 2014), MTurk workers reported being about two thirds of a standard deviation less extraverted, about a third of a standard deviation less open to new experiences, and only slightly less agreeable, conscientious, or emotionally stable.

The vast majority (92.5%) of participants in our study completed the survey on a computer. Of the remaining participants, 2% completed the survey using a tablet, 4.5% using a phone, and the rest using other devices (e.g., game consoles) or devices that could not be identified. Rates of mobile device use are somewhat lower than have been noted in other online panels (de Bruijne & Wijnant, 2014a, 2014b).

Sample Differences by Time of Completion

The focus of our investigation is how the composition of the MTurk worker pool varied across days of the week, across time of day, and across the serial order in which they participated. Main findings of these analyses are summarized in Table 5. We looked for variations within the following variables: age, gender identity, education, employment, household income, household size, race, Latino ethnicity, socioeconomic status, sexual orientation, relationship status, party identification, religion, and religiosity. Our survey design allowed respondents to identify as more than one race, so we treated each racial category (White, Black or African American, Asian American, American Indian or Alaskan Native, Native Hawaiian or Pacific Islander, or Other) as a single binary dependent variable. We also looked for differences in the Big Five personality traits: extraversion, agreeableness, conscientiousness, emotional stability, and openness. Finally, we examined workers' prior experience and where they reported finding the survey.

In two instances, similar and highly correlated variables were collected for purposes irrelevant to the present study. In each case, only one variable was selected for analysis. The first instance was marital status and relationship status. We

Table 2. Socioeconomic Characteristics of Workers.

Characteristic	Total sample (*N* = 9,770)	First respondents (*N* = 438)	National estimates
Household income			
<14,999	11.7% [11.1, 12.3]	11% [8.1, 13.9]	12.5%[a]
15,000-29,999	17.5% [16.8, 18.3]	17.4% [13.9, 21.0]	15.6%[a]
30,000-49,999	24.6% [23.8, 25.5]	25.8% [21.7, 29.9]	18.5%[a]
50,000-74,999	20.7% [19.9, 21.5]	21.5% [17.7, 25.3]	17.8%[a]
75,000-99,999	12.2% [11.6, 12.9]	12.6% [9.5, 15.7]	12.1%[a]
>US$100,000	12.9% [12.2, 13.6]	11.7% [8.7, 14.7]	23.5%[a]
Household size	2.82 [2.79, 2.85]	2.76 [2.62, 2.90]	2.64[a]
Living with parents	16.5% [15.8, 17.2]	15.1% [11.8, 18.5]	—
Employment status			
Employed full-time	48.5% [47.5, 49.5]	55.3% [50.6, 60.0]	48.4%[b]
Working part-time	15.7% [15.0, 16.4]	14.2% [10.9, 17.5]	10.8%[b]
Homemaker	8.6% [8.0, 9.2]	8% [5.5, 10.5]	5.4%[b]
Unemployed	9.4% [8.8, 10.0]	9.1% [6.4, 11.8]	4.8%[b]
Retired	2.2% [1.9, 2.5]	1.4% [0.3, 2.5]	15.4%[b]
Student	11.9% [11.3, 12.5]	7.5% [5.0, 10.0]	6.4%[b]
Permanent disability	1.9% [1.6, 2.2]	2.3% [0.9, 3.7]	6.5%[b]
Other	1.7% [1.4, 2.0]	2.3% [0.9, 3.7]	1.2%[b]
Education			
Less than high school	0.7% [0.5, 0.9]	1.4% [0.3, 2.5]	16.1%[c]
High school or equivalent	10.2% [9.6, 10.8]	10.5% [7.6, 13.4]	27.6%[c]
Some college	31.4% [30.5, 32.3]	22.4% [18.5, 26.3]	18.1%[c]
2-year college degree	11.7% [11.1, 12.3]	9.6% [6.8, 12.4]	9.1%[c]
4-year college degree	34.8% [33.9, 35.7]	44.5% [39.9, 49.2]	18.5%[c]
Postgraduate degree	11.1% [10.5, 11.7]	11.6% [8.6, 14.6]	10.6%[c]

Note. 95% CI indicated in parentheses. CI = confidence interval.
[a]U.S. Census Bureau (2011-2015).
[b]Bureau of Labor Statistics, U.S. Department of Labor (2016).
[c]U.S. Census Bureau (2016).

Table 3. Relationship Characteristics of Workers.

Characteristic	Total sample (*N* = 9,770)	First respondents (*N* = 438)	National estimates
Relationship status			
Single	32.3% [31.4, 33.2]	36.5% [32.0, 41.0]	—
Casually dating	5% [4.6, 5.4]	5.7% [3.5, 7.9]	—
Monogamous	60.6% [59.6, 61.6]	56.8% [52.2, 61.4]	—
Consensually nonmonogamous	1.5% [1.3, 1.7]	0.7% [0.0, 1.5]	—
Other/refused	0.3% [0.2, 0.4]	0.0% [0.0, 0.3]	—
Marital status			
Never married	42.8% [41.8, 43.8]	46.1% [41.4, 50.8]	32.8%[a]
Married	34.9% [34.0, 35.9]	29.2% [24.9, 33.5]	48.2%[a]
Partnered	14.2% [13.5, 14.9]	16.4% [12.9, 19.9]	—
Separated	1.2% [1.0, 1.4]	0.5% [0.0, 1.2]	2.1%[a]
Divorced	6% [5.5, 6.5]	7.3% [4.9, 9.7]	11.0%[a]
Widowed	0.8% [0.6, 1.0]	0.5% [0.0, 1.2]	5.9%[a]
Sexual orientation			
Lesbian or gay	3.8% [3.4, 4.2]	2.3% [0.9, 3.7]	1.7%[b]
Bisexual	6.9% [6.4, 7.4]	6.6% [4.3, 8.9]	1.8%[b]
Straight	86.8% [86.1, 87.5]	88.8% [85.9, 91.8]	96.5%[b]
Other	2.2% [1.9, 2.5]	2.1% [0.8, 3.4]	—

Note. 95% CI indicated in parentheses. CI = confidence interval.
[a]U.S. Census Bureau (2011-2015).
[b]General Social Survey (as reported and summarized in Gates, 2014).

Table 4. Attitudinal and Personality Characteristics of Workers.

Characteristic	Total sample (*N* = 9,770)	First respondents (*N* = 438)	National estimates[a]
Political affiliation			
Identifies as republican	17.90% [17.1, 18.7]	18.3% [14.7, 21.9]	28.8% [27.0, 30.5]
Identifies as democrat	41.30% [40.3, 42.3]	47% [42.3, 51.7]	34.9% [33.1, 36.7]
Ideology (1 = extremely liberal, 7 = extremely conservative)	3.39 [3.36, 3.42]	3.31 [3.16, 3.46]	4.26 [4.20, 4.31]
Religion			
Christian—Mainline Protestant	16% [15.3, 16.7]	13.3% [10.1, 16.5]	11.7% [10.5, 12.8]
Christian—Evangelical	8.5% [8.0, 9.1]	8.6% [6.0, 11.3]	21.3% [19.7, 22.8]
Christian—Catholic	11.4% [10.8, 12.0]	14.3% [11.0, 17.6]	22.4% [20.9, 24.0]
Christian—Other/not specified	10% [9.4, 10.6]	7.3% [4.9, 9.7]	13.8% [12.5, 15.1]
Jewish	1.2% [1.0, 1.4]	0.5% [0.0, 1.2]	2.2% [1.7, 2.8]
Muslim	0.6% [0.5, 0.8]	1.4% [0.3, 2.5]	1.0%[b]
Atheist	20.4% [19.6, 21.2]	25.5% [21.4, 29.6]	3.1%[c]
Nothing in particular	24.6% [23.8, 25.5]	23.6% [19.6, 27.6]	24.0% [22.3, 25.6]
Other	7% [6.5, 7.5]	5.3% [3.2, 7.4]	4.7% [3.9, 5.5]
Religiosity			
Attends at least weekly	9.2% [8.6, 9.8]	6.9% [4.5, 9.3]	21.4% [19.8, 22.9]
Attends at least monthly	12.1% [11.5, 12.8]	13.1% [9.9, 16.3]	11.3% [10.1, 12.5]
Attends a few times per year	24.2% [23.4, 25.1]	22.6% [18.7, 26.5]	24.1% [22.5, 25.8]
Never attends	54.1% [53.1, 55.1]	57.4% [52.8, 62.0]	43.2% [41.3, 45.1]
Big Five personality traits (1 = *low*, 7 = *high*)			
Extraversion	3.58 [3.55, 3.61]	3.48 [3.33, 3.63]	4.13 [4.09, 4.18]
Agreeableness	5.11 [5.09, 5.13]	5.18 [5.06, 5.30]	5.11 [5.07, 5.15]
Conscientiousness	5.24 [5.21, 5.27]	5.40 [5.28, 5.52]	5.63 [5.59, 5.67]
Emotional stability	4.70 [4.67, 4.73]	4.90 [4.76, 5.04]	4.92 [4.87, 4.97]
Openness	5.09 [5.07, 5.11]	4.86 [4.74, 4.98]	4.81 [4.77, 4.85]

Note. 95% CI indicated in parentheses. CI = confidence interval.
[a]Population estimates derived from American National Election Studies 2012 time series unless otherwise noted.
[b]Pew Research Center (2016a).
[c]Pew Research Center (2016b).

selected marital status for analysis because this variable is more typically recorded in national surveys and therefore more relevant for this demographic analysis. The second instance was political ideology and party affiliation. We conducted the analyses using political ideology, but results are identical when party identification is used instead.

To limit the number of comparisons, some response options were collapsed into broader categories (e.g., specific denominations of Christianity were collapsed into a single category). In total, given the coding, our final analysis included 31 different demographic variables.

For all continuous, ordinal, and binomial variables, generalized linear modeling (GZLM) was used to regress (a) the day of the week (categorical), (b) the time of day the batch was posted (categorical), (c) the serial position of the batch within the data collection run (continuous), (d) the serial position of the individual response within the batch (continuous), and (e) a dichotomous variable representing the amount of compensation (categorical) to control for possible effects of increasing payment part way through the study. Interval dependent measures were treated as linear effects, except for worker experience (i.e., the total number of MTurk HITs already completed), which was modeled using a negative binomial distribution. This approach was adapted to multinomial regression to evaluate differences in religion, as SPSS' implementation of GZLM cannot be used for multinomial variables.

Including so many independent and dependent variables brings with it the risk of false positives. To mitigate this risk, we limited the number of comparisons by not including interactions in the model. We also limited the comparisons of each time or day to the grand mean for all times and days (rather than individual comparisons against all other times or days). For example, we compared the mean percentage of college graduates in batches posted on Tuesdays with the mean percentage of college graduates in all batches (including Tuesdays). This approach led to a total of 13 significance tests for each of the 29 demographic variables and two MTurk behavior variables (worker experience and where they found the study), for a total of 403 comparisons.

To further reduce the potential for false positives, we set the alpha criterion at .01, rather than the more typical .05, and used the Benjamini–Hochberg adjustment (Benjamini & Hochberg, 1995) to hold the false discovery rate across all

Table 5. Significant Results by Time of Day, Day of Week, Serial Position, and Pay Rate.

Outcome	Contrast	Wald	*p*	*d*	Interpretation
Time of day effects					
Time zone	10 a.m. vs. mean	71.93	<.00001	0.17	More workers from eastern time zones at 10 a.m. ET
Time zone	10 p.m. vs. mean	68.12	<.00001	0.17	More workers from western time zones at 10 p.m. ET
Worker experience	10 p.m. vs. mean	43.67	<.00001	0.13	Workers are less experienced at 10 p.m. ET
Worker experience	10 a.m. vs. mean	27.78	<.00001	0.11	Workers are more experienced at 10 a.m. ET
% completed by smartphone	10 p.m. vs. mean	18.01	<.00001	0.09	Workers more likely to use phones at 10 p.m. ET
Relationship status	10 a.m. vs. mean	16.91	<.00001	0.08	Workers less likely to be single at 10 a.m. ET
Relationship status	10 p.m. vs. mean	16.63	<.00001	0.08	Workers more likely to be single at 10 p.m. ET
Found HIT outside of MTurk	10 a.m. vs. mean	16.01	<.00001	0.08	Workers less likely to find the HIT outside of MTurk at 10 a.m. ET
% Asian American	10 p.m. vs. mean	15.51	<.00001	0.08	Workers more likely to be Asian American at 10 p.m. ET
% Asian American	10 a.m. vs. mean	15.24	<.00001	0.08	Workers less likely to be Asian American at 10 a.m. ET
Found HIT outside of MTurk	3 p.m. vs. mean	13.07	.0003	0.07	Workers more likely to find the HIT outside of MTurk at 3 p.m. ET
Conscientiousness	10 p.m. vs. mean	11.53	.0007	0.07	Workers are less conscientious at 10 p.m. ET
Day of week effects					
Found HIT outside of MTurk	Sat vs. mean	35.87	<.00001	0.12	Workers less likely to find the HIT outside MTurk on Saturday
Found HIT outside of MTurk	Thurs vs. mean	35.52	<.00001	0.12	Workers more likely to find the HIT outside MTurk on Thursday
Age	Sat vs. mean	35.08	<.00001	0.12	Workers were older on Saturdays
Age	Thurs vs. mean	32.14	<.00001	0.11	Workers were younger on Thursdays
Employment status	Sun vs. mean	14.01	0.0002	0.08	Workers more likely to have full-time jobs; less likely to lack formal employment altogether (no change in part-time status)
Age	Wed vs. mean	12.47	0.0004	0.07	Workers were younger on Wednesdays
Found HIT outside of MTurk	Sun vs. mean	12.12	0.0005	0.07	Workers less likely to find the HIT outside MTurk on Sunday
Overall serial position effects					
Worker experience	Linear effect	460.68	<.00001	0.44	Workers more experienced earlier in the data collection
Emotional stability	Linear effect	38.20	<.00001	0.13	Workers more emotionally stable earlier in the data collection
Age	Linear effect	26.67	<.00001	0.1	Workers were older earlier in the data collection
Conscientiousness	Linear effect	23.96	<.00001	0.1	Workers more conscientious earlier in the data collection
Agreeableness	Linear effect	23.44	<.00001	0.1	Workers more agreeable earlier in the data collection
Employment status	Linear effect	12.55	.0004	0.07	Workers more likely to have full-time jobs earlier in the data collection
Household size	Linear effect	12.36	.0004	0.07	Workers come from smaller households earlier in the data collection
Within-batch serial position effects					
Worker experience	Linear effect	35.27	<.00001	0.12	More experienced workers respond to an available HIT faster
Sex	Linear effect	26.99	<.00001	0.09	Female workers respond to an available HIT faster
% Asian American	Linear effect	18.52	<.00001	0.08	Asian workers respond to an available HIT slower
Age	Linear effect	14.06	.0002	0.08	Younger workers respond to an available HIT slower
Found HIT outside of MTurk	Linear effect	159.38	<.0001	0.26	Workers who completed the HIT sooner were less likely to have found it outside MTurk
Pay effects					
Worker experience	High vs. low pay	78.69	<.00001	0.18	Workers more experienced once pay was increased
Emotional stability	High vs. Low Pay	13.35	.0003	0.07	Workers more emotionally stable once pay was increased

Note. This table includes the 33 comparisons that revealed statistically significant differences. We only report effect sizes for statistically significant results. The entries in the table are sorted by type of temporal variation, and then by ascending order of effect size. As noted in the text, we used the Benjamini–Hochberg adjustment for multiple comparisons and consider all *p*-values less than .0007 to be statistically significant (this ensures that the false discovery rate across all comparisons is held constant at .01). ET = Eastern Time; HITs = Human Intelligence Tasks; MTurk = Amazon Mechanical Turk.

comparisons constant at .01 across all tests. Following these adjustments, no results with an unadjusted p value above .0007 are reported as statistically significant, and of the significant results that we report, only four are expected to be false positives observed by chance alone.[3] Table 5 includes the 33 statistically significant differences among the 403 comparisons.

Day of week effects. Of our 217 day-of-week comparisons, we found seven instances in which the attributes of participants recruited on a particular day of the week significantly differed from the sample as a whole.[4] These findings are summarized in Table 5.

The average age of respondents varied as a function of the day of the week. Participants on Wednesday ($M = 32.4$, $SD = 10.78$) and Thursday ($M = 32.46$ $SD = 10.67$; $\beta = -1.04$, Wald $\chi^2 = 12.47$, $p < .001$, $d = .07$ and $\beta = -1.44$, Wald $\chi^2 = 32.14$, $p < .0001$, $d = .11$, respectively) were somewhat younger than the sample as a whole ($M = 33.51$, $SD = 11.31$). Respondents completing the survey on Saturday were somewhat older than average ($M = 35.84$, $SD = 12.47$; $\beta = 1.88$, Wald $\chi^2 = 35.09$, $p < .0001$, $d = .12$).

People completing HITs on Sundays were more likely to be employed full-time (52%) than the sample as a whole (48.5%; $\beta = .21$, Wald $\chi^2 = 14.01$ $p = .0002$, $d = .08$), with a corresponding decrease in the proportion of individuals without any formal employment (31.2% as compared with 35.7%). The proportion of workers employed part-time was roughly the same across all days of the week.

Workers were less likely to find the survey outside of MTurk on Saturday (3.4%) or Sunday (6%) than the sample as a whole (9%; $\beta = -.04$, Wald $\chi^2 = 35.87$ $p < .0001$, $d = .12$ and $\beta = -.02$, Wald $\chi^2 = 12.12$ $p = .0005$, $d = .07$, respectively). Workers who completed the survey on a Thursday were much more likely to have found it on a source outside of MTurk (15.3%; $\beta = .04$, Wald $\chi^2 = 35.52$ $p < .0001$, $d = .12$).

Time of day effects. Of our 93 time-of-day comparisons, we found 12 instances in which attributes of participants recruited at a particular time of day differed significantly from the grand mean.[5] These differences generally reflected linear trends in the composition of the MTurk workforce throughout the day, and are summarized in Table 5.

As might be expected, one of the most pronounced consequences of posting at different times was variation in the proportion of workers from different time zones. People in earlier time zones were more likely than average to complete HITs posted at 10 a.m. ($\beta = -.15$, Wald $\chi^2 = 71.92$, $p < .0001$, $d = .17$). Conversely, people in later time zones were more likely to complete HITs posted at 10 p.m. ($\beta = .13$, Wald $\chi^2 = 68.11$, $p < .0001$, $d = .17$). As an illustration of the consequences of this shift, 56.8% of respondents at 10 a.m. Eastern Time were from the U.S. Eastern time zone while only 10.9% of workers were from the Pacific Time zone. In contrast, 48.6% of workers at 10 p.m. Eastern Time reside in the U.S. Eastern time zone, while 18.9% of workers were from the U.S. Pacific time zone.

The proportion of Asian American respondents also increased over the course of the day, growing from 5.9% at 10 a.m. to 7.6% at 3 p.m. to 9% at 10 p.m. The proportion of Asian Americans was significantly lower than average at 10 a.m. ($\beta = -.016$, Wald $\chi^2 = 15.24$, $p < .0001$, $d = .08$) and significantly higher than average at 10 p.m. ($\beta = .016$, Wald $\chi^2 = 15.49$, ps $< .0001$, $d = .08$). This effect was no longer significant, however, when controlling for time zone, suggesting that this difference reflects that more Asian American workers live on the west coast.

Other differences were observed that were not an artifact of time zone. The proportion of single workers increased linearly throughout the day from 29.1% at 10 a.m. to 32.2% at 3 p.m. to 34.9% at 10 p.m. The proportion of workers who are single was significantly lower than average at 10 a.m. ($\beta = -.03$, Wald $\chi^2 = 16.91$, $p < .0001$, $d = .08$) and significantly higher than average at 10 p.m. ($\beta = .03$, Wald $\chi^2 = 16.62$, $p < .0001$, $d = .08$).

More workers who completed the survey at 10 p.m. used smartphones (5.8%) than across the sample as a whole (3.7%; $\beta = .014$, Wald $\chi^2 = 18.01$, $p < .0001$ $d = .09$). Workers recruited at 10 p.m. also reported being less conscientious ($M = 5.18$, $SD = 1.31$) than the sample as a whole ($M = 5.24$, $SD = 1.27$; $\beta = -.06$, Wald $\chi^2 = 11.53$, $p = .0007$, $d = .07$).

Workers who completed the HIT at 10 a.m. were less likely to report having found the HIT outside of the MTurk interface (8.5%) than the sample as a whole (9%; $\beta = -.014$, Wald $\chi^2 = 16.01$, $p < .0001$, $d = .08$). Workers who completed the HIT at 3 p.m. were more likely (9.7%) to have found the HIT outside of the MTurk interface ($\beta = .013$, Wald $\chi^2 = 13.07$, $p = .0003$, $d = .07$).

Finally, relative to the sample as a whole ($M = 4.67$, $SD = 10.04$), more experienced workers tended to participate in the morning ($M = 5.02$, $SD = 10.57$; $\beta = .43$, Wald $\chi^2 = 27.77$, $p < .0001$, $d = .11$) and less likely to do so at night ($M = 4.75$, $SD = 8.29$; $\beta = -.43$, Wald $\chi^2 = 43.67$, $p < .0001$, $d = .13$).

Overall serial position effects. Of our 31 positional comparisons, we found seven instances in which the attributes of participants differed over time.[6] Workers who completed HITs earlier in the data collection process reported higher levels of emotional stability, conscientiousness, and agreeableness. Participants who completed earlier batches of HITs also tended to be older were more likely to have a full-time job and live in smaller households. Workers who completed HITs earlier were also substantially more experienced than workers recruited later in the study (Table 6).

Within-batch serial position effects. Of our 31 positional comparisons within batch, we found five instances in which the attributes of participants recruited earlier in a given batch

Table 6. Worker Characteristics as a Function of Serial Position Across Study.

	Respondent 2065 (−1 *SD*)	Respondent 7706 (+1 *SD*)	Linear trend
Age	34.79 [34.41, 3.16]	32.62 [32.01, 33.22]	β = −.00038, Wald χ^2 = 26.67, p < .001, d = .10
Household size	2.73 [2.69, 2.78]	2.92 [2.85, 2.99]	β = .00003, Wald χ^2 = 12.36, p < .001, d = .07
Employed full-time	50% [48, 52]	45% [42, 47]	β = −.00004, Wald χ^2 = 12.55, p < .001, d = .07
Conscientiousness	5.34 [5.29, 5.38]	5.10 [5.03, 5.17]	β = −.00004, Wald χ^2 = 23.96 p < .001, d = .10
Agreeableness	5.20 [5.15, 5.24]	4.97 [4.91, 5.04]	β = −.00004, Wald χ^2 = 23.44 p < .001, d = .10
Emotional stability	4.83 [4.78, 4.88]	4.49 [4.41, 4.57]	β = −.00006, Wald χ^2 = 38.20 p < .001, d = .13
Worker experience	6.52 [6.29, 5.6.76]	2.66 [2.51, 2.83]	β = −.00016, Wald χ^2 = 460.68, p < .0001, d = .44

Table 7. Worker Characteristics as a Function of Serial Position Within Batches.

	First respondent in batch (−1 *SD*)	100th responder in batch (+1 *SD*)	Linear trend
Age	34.14 [33.81, 34.48]	33.26 [32.85, 33.67]	β = −.0089, Wald χ^2 = 14.06, p < .0001, d = .08
Female	56% [44%, 57%]	50% [48%, 52%]	β = .0022, Wald χ^2 = 26.99, p < .0001, d = .11
Asian American	7% [6%, 8%]	9% [8%, 10%]	β = .0031, Wald χ^2 = 18.52, p < .0001, d = .09
Found survey outside of Mechanical Turk	5% [4%, 6%]	10% [9%, 11%]	β =.0074, Wald χ^2 = 159.38, p < .0001, d = .26
Worker experience	4.46 [4.31, 4.61]	3.89 [3.74, 4.06]	β = −.0013, Wald χ^2 = 35.27, p < .0001, d = .12

differed from the attributes recruited later in the same batch. Workers who completed an available HIT earlier in a given batch were on average older, more likely to be female, and less likely to be Asian American. Workers who completed HITs sooner were also less likely to have found the survey on a source outside of MTurk but tended to be more experienced than workers recruited later in the study (Table 7).

Pay effects. Pay effects were included primarily to control for a change in design part way through data collection. Of the 31 payment comparisons, we found evidence of only two characteristics that changed once we offered to pay more. Controlling for other variables, workers in the high-pay condition reported higher emotional stability (M = 4.77 SD = 1.90) than workers paid less (M = 4.56, SD = 2.30; β = .32, Wald χ^2 = 13.35, p = .0003, d = .07). Workers were also more experienced when pay was higher (M = 4.77, SD = 1.90) than when pay was lower (M = 4.77, SD = 1.90; β = .47, Wald χ^2 = 78.69, p < .0001, d = .18). These results and all other significant intertemporal differences are summarized in Table 5.

Discussion

In this article, we have described demographic characteristics of a large sample of MTurk workers and examined differences across time, day, and serial position. Of our 403 demographic comparisons, we found 33 differences (8.2% of tested effects), and significant effects had an average effect size of d = 0.11. These findings provide evidence that MTurk samples vary intertemporally, but that in general these differences are small. An important caveat to these findings is that we recruited workers without allowing for replacement—that is, workers could only participate once. Differences between samples may be larger or smaller if workers are not restricted from participating more than once.

Demographic Differences by Day and Time

Day of the week influenced few (2%, or 4/203) demographic characteristics, and these effects were small (M_d = 0.09). To the extent that these effects were detectable, they suggest that samples collected over the weekend are more likely to include older and more fully employed respondents. These differences seem plausible, but the lack of differences across other characteristics suggests that potential day of week effects can be safely ignored.

Time of day resulted in similarly small effects (M_d = 0.10) but within a larger proportion (9%, or 8/87) of measured variables. In almost all cases, these differences represented linear trends in sample composition across the day, and thus when considering the potential impact of recruiting in the morning or in the evening, the combined impact of both effect size estimates should be considered.

Of particular note, contrary to previous research (Komarov et al., 2013), we found that workers were more likely to use mobile devices late at night (5.8% of HITs posted at 10 p.m. were submitted from mobile phones, compared with 3.7% of HITs submitted during the rest of the day). Mobile device use can have adverse effects on data quality, including increased rates of attrition (Mavletova, 2013; Sommer, Diedenhofen, & Musch, 2016; Wells, Bailey, & Link, 2013) and shorter and fewer open-ended responses (Mavletova, 2013; Struminskaya, Weyandt, & Bosnjak, 2015). As a result,

researchers might consider adjusting the time of day at which they post research studies or collect data if they hope to optimize mobile completion or collect open-ended responses.

The large proportion of observed differences suggest that time of day effects might be a fruitful area of future research, both through expanding the range of variables that are examined and with a particular effort to understand how regional differences, differences in the active user population across time within regions, and changes in individual responses throughout the day combine to produce these differences.

Demographic Differences by Serial Position

The effects of serial position were more extensive than time-of-day and day-of-week effects; 21% (6/29) of across-sample serial position effects were significant, with an average effect size of M_d = = .10 and 10% (3/29) of within-batch serial position effects were significant, with M_d = = .09. Many of these across-sample findings are compatible with earlier studies of serial position effects. As observed in university subject pools, early respondents report higher levels of conscientiousness (Aviv et al., 2002; Ebersole et al., 2016). In general population samples, those who responded to surveys first tended to be older (Filion, 1975; Sigman et al., 2014). We observe similar results both across our entire sample and within individual batches of HITs. While other studies find that women are more likely to respond to requests to complete both mail surveys (Gannon et al., 1971) and web surveys (Sigman et al., 2014) quickly (Cooper et al., 1991; Ebersole et al., 2016), we find that women respond more quickly within batches, but not across the sample as a whole. Contrary to studies of race and serial position effects in other modes (Gannon et al., 1971; Sigman et al., 2014; Voigt et al., 2003), we found little evidence that racial diversity increased over time. Typically, later survey respondents belong to groups that are possible but difficult to contact. Only those who register with MTurk can take part in surveys posted on the platform. African American and Latino populations are underrepresented on MTurk, and so it may be that those individuals who may be possible but difficult to contact through other modes of survey data collection are simply impossible to reach on MTurk.

When sampling error is unsystematic, larger samples more closely approximate the population. This is not so in the presence of systematic bias. As our sample increased, some biases (e.g., the democratic tendencies of respondents) remained the same. In other cases, biases actually increases (e.g., age, employment, conscientiousness, and emotional stability). Thus, it is not a given that making a sample more representative of the U.S. MTurk worker population will also make it more representative of the U.S. population as a whole. Variations in demographic characteristics across the entire sample are also relevant to researchers who recruit workers from the available pool without replacement (e.g., to prevent workers from completing the same study twice). Of particular relevance, we found variations in the "Big Five" personality factors as a function of serial position. Workers who completed HITs earlier in the data collection process reported being slightly more emotionally stable, more conscientious, and more agreeable. These traits are associated with and may moderate other important variables including respondent data quality, or political behaviors and attitudes that might bias samples (for an excellent review, see Gerber, Huber, Doherty, & Dowling, 2011), or data quality.

Variations in demographic characteristics associated with serial position within batches of HITs are important when considering whether to recruit respondents in large batch or small batches. It is particularly important to understand potential within-batch serial position effects because several third-party solutions (e.g., TurkGate, Goldin & Darlow, 2013; and TurkPrime, Litman, Robinson, & Abberbock, 2017) make it easy to divide data collection efforts into a large number of very small batches. By and large, we find that smaller batches will lead samples to be older and have more women, but will attenuate the overrepresentation of Asian American workers.

Differences in Worker Experience and Forum Use

Time of day and serial position were strongly related to how much MTurk experience respondents had and how workers found the survey. More experienced workers completed the survey earlier in data collection (both within and across batches). Variations in worker experience may be associated with greater exposure to survey tactics, experimental manipulations, which can have various effects on data quality. On one hand, more experienced workers are more familiar with common research questions, leading to practice effects (Chandler et al., 2014), potentially smaller effect sizes on commonly used experimental paradigms (Chandler et al., 2015) and potentially more extreme and less malleable attitudes toward topics that respondents are frequently asked about (Sturgis, Allum, & Brunton-Smith, 2009). On the other hand, more experienced workers may be more attentive and therefore may provide higher quality responses.

We also observed substantial intertemporal variation in workers using forums, with more referrals from links shared outside of MTurk happening in the afternoon and on Thursdays and less in evenings and weekends. These differences may be relevant if researchers are concerned about respondents who have potentially seen information about a study prior to completing it. The longer a HIT is available, the more opportunity workers have to find it on an outside forum.

Although we did not vary pay rates experimentally, we nonetheless found that when we increased pay, there was a concomitant increase in the experience of survey participants. Together, we thus observed two separate patterns: (a) Early responders to the survey tended to be more experienced workers and (b) when we increased the pay,

the proportion of more experienced workers increased even further. If researchers are concerned that worker savviness might affect their findings (Krupnikov & Levine, 2014), they should be attentive to these possibilities when they post their studies.

Conclusion

This study is the largest and most comprehensive description of MTurk demographics that we are aware of and the first large-scale effort to examine intertemporal differences in sample composition (however, for a similar project, see Arechar, Kraft-Todd, & Rand, 2016). Data from our study of approximately 10,000 MTurk workers have allowed us to examine three key possible sources of temporal variation in MTurk sample composition: (a) time of day, (b) day of week, and serial position both (c) across the entire data collection and (d) within specific batches.

Taken as a whole, our results should serve as a source of both comfort and caution to scholars who use MTurk to recruit participants for their research. On one hand, we found only minimal day-of-week differences. However, we also showed that there are small but significant time-of-day variations in demographic composition—variations that bear closer scrutiny. The effects of serial position also warrant further study, as they emerged as persistent influences across multiple variables, including characteristics known to affect political and psychological attitudes (e.g., Big Five personality traits; Dietrich, Lasley, Mondak, Remmel, & Turner, 2012; Gerber et al., 2011). Differences in sample composition can compromise claims to generalizability and might lead to challenges with reproducing research findings as well (Peterson & Merunka, 2014). As is often the case, larger samples (and/or those recruited in such a way to be more representative) are especially critical when researchers are concerned about heterogeneous treatment effects may reduce the external validity of a given sample.

Researchers should bear our findings in mind as they consider how best to recruit samples from MTurk. The intertemporal dynamics we have detailed are likely to be most relevant to researchers attempting to collect representative samples of the MTurk worker population, such as studies of MTurk worker behavior and attitudes that attempt to understand the dynamics of contract labor and piece-work in the "gig economy" (Aguinis & Lawal, 2013; Brawley & Pury, 2016). But researchers interested in other topics should pay attention to relationships such as those between serial position and psychological characteristics and consider including information about when and how many times they posted their HIT when reporting results.[7] Perhaps most importantly, these findings demonstrate that the number of workers recruited and the size of batches used to recruit them can have a large effect on the average experience of sample respondents.

As MTurk and other similar online convenience samples become more widely used, it is increasingly important that we better understand who participates in these subject pools and when certain kinds of respondents are more likely to opt-in relative to others. Such examinations will help researchers assess published results, especially (though not limited to) their generalizability across populations and over time.

This project suggests several directions for future research. Beyond extending the analysis of temporal effects to new variables, or examining intertemporal variation in other sources of data, future work could examine how other design choices affect sample composition, including whether researchers with poor ratings or tasks with low pay get substantively different samples than researchers with better ratings or tasks with higher pay. This is an important area for future research to examine, particularly as researchers continue or increase reliance on online data collection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. *People of color* is a commonly used umbrella term denoting racial and ethnic minorities in America, including African Americans, Latinos, Asians, and others.
2. We followed this general procedure when it was time to repost the HIT: first, close the existing HIT; second, prevent the workers who participated in the existing HIT from participating in future postings (using qualifications; see Chandler, Mueller, & Paolacci, 2014); third, post the new HIT.
3. The Benjamini–Hochberg adjustment does not identify specific false positives, but rather holds the number of false positives constant across many tests to a specified level.
4. Seven days by 31 variables produces 217 comparisons.
5. Three times of day (10 a.m., 3 p.m., 10 p.m.) by 31 demographic variables produces 93 comparisons.
6. Thirty-one demographic variables, treating time as a linear effect by batch number.
7. The size of these effects will depend on both the magnitude of difference between the samples on a given variable and the magnitude of the moderating effect this variable has on the theoretical relationship of interest (Ho, Imai, King, & Stuart, 2007).

References

Aguinis, H., & Lawal, S. O. (2013). eLancing: A review and research agenda for bridging the science–practice gap. *Human Resource Management Review*, *23*, 6-17. doi:10.1016/j.hrmr.2012.06.003

American National Election Studies. (2012). *The ANES 2012 Time Series Study* [dataset]. Stanford University and University of

Michigan [producers]. Available from www.electionstudies.org

Arechar, A. A., Kraft-Todd, G. T., & Rand, D. G. (2016). *Turking overtime: How participant characteristics and behavior vary over time and day on Amazon Mechanical Turk*. Retrieved from https://ssrn.com/abstract=2836946

Aviv, A. L., Zelenski, J. M., Rallo, L., & Larsen, R. J. (2002). Who comes when: Personality differences in early and later participation in a university subject pool. *Personality and Individual Differences*, *33*, 487-496. doi:10.1016/S0191-8869(01)00199-4

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B: Methodological*, *57*, 289-300. doi:10.2307/2346101

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351-368. doi:10.1093/pan/mpr057

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

Bureau of Labor Statistics, U.S. Department of Labor. (2016). Reasons people give for not being in the labor force, 2004 and 2014 on the Internet. *The Economics Daily*. Retrieved from https://www.bls.gov/opub/ted/2016/reasons-people-give-for-not-being-in-the-labor-force-2004-and-2014.htm

Casler, K., Bickel, L., & Hackett, E. (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*, 2156-2160. doi:10.1016/j.chb.2013.05.009

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaive among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods Science*, *46*, 112-130.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnative participants can reduce effect sizes. *Psychological Science*, *26*, 1131-1139. doi:10.1177/0956797615585115

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*, 53-81. doi:10.1146/annurev-clinpsy-021815-093623

Cooper, H., Baumgardner, A. H., & Strathman, A. (1991). Do students with different characteristics take part in psychology experiments at different times of the semester? *Journal of Personality*, *59*, 109-127. doi:10.1111/j.1467-6494.1991.tb00770.x

Corrigan, P. W., Bink, A. B., Fokuo, J. K., & Schmidt, A. (2015). The public stigma of mental illness means a difference between you and me. *Psychiatry Research*, *226*, 186-191. doi:10.1016/j.psychres.2014.12.047

Couper, M. P. (2000). Review: Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*, 464-494. doi:10.1086/318641

de Bruijne, M., & Wijnant, A. (2014a). Improving response rates and questionnaire design for mobile web surveys. *Public Opinion Quarterly*, *78*, 951-962. doi:10.1093/poq/nfu046

de Bruijne, M., & Wijnant, A. (2014b). Mobile response in web panels. *Social Science Computer Review*, *32*, 728-742. doi:10.1177/0894439314525918

Dietrich, B. J., Lasley, S., Mondak, J. J., Remmel, M. L., & Turner, J. (2012). Personality and legislative politics: The Big Five trait dimensions among U.S. state legislators. *Political Psychology*, *33*, 195-210.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J., Banks, J. B., . . . Nosek, B. A. (2016, June 28). *Many labs 3: Evaluating participant pool quality across the academic semester via replication*. Retrieved from osf.io/ct89g

Filion, F. L. (1975). Estimating bias due to nonresponse in mail surveys. *Public Opinion Quarterly*, *39*, 482-492. doi:10.1086/268245

Flores, A., Herman, J. L., Gates, G. J., & Brown, T. N. T. (2016). *How many adults identify as transgender in the United States?* The Williams Institute. Retrieved from https://williamsinstitute.law.ucla.edu/research/how-many-adults-identify-as-transgender-in-the-united-states/

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, *9*(10), e109019. doi:10.1371/journal.pone.0109019

Gannon, M. J., Nothern, J. C., & Carroll, S. J. (1971). Characteristics of nonrespondents among workers. *Journal of Applied Psychology*, *55*, 586-588. doi:10.1037/h0031907

Gates, G. (2014). *LGBT demographics: Comparisons among population-based surveys*. The Williams Institute. Retrieved from https://williamsinstitute.law.ucla.edu/wp-content/uploads/lgbt-demogs-sep-2014.pdf

Gates, G., & Newport, F. (2012, October 18). *Special report: 3.4% of U.S. adults identify as LGBT*. Retrieved from http://opiniontoday.com/2012/10/18/special-report-3-4-of-u-s-adults-identify-as-lgbt/

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). The Big Five personality traits in the political arena. *Annual Review of Political Science*, *14*, 265-287. doi:10.1146/annurev-polisci-051010-111659

Goldin, G., & Darlow, A. (2013). TurkGate (Version 0.4. 0) [Software]. Retrieved from http://gideongoldin.github.io/TurkGate/

Gosling, S. D., Rentfrow, P. J., & Potter, J. (2014). *Norms for the Ten Item Personality Inventory* (Unpublished data).. Retrieved from http://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*, 504-528.

Guidelines for academic requesters. (2014). Version 1.1. Retrieved from https://irb.northwestern.edu/sites/irb/files/documents/guidelinesforacademicrequesters.pdf

Haupert, M. L., Gesselman, A. N., Moors, A. C., Fisher, H. E., & Garcia, J. R. (2016). Prevalence of experiences with consensual nonmonogamous relationships: Findings from two national samples of single Americans. *Journal of Sex & Marital Therapy*, *22*, 1-17.

Hillygus, D. S., Jackson, N., & Young, M. (2014). Professional respondents in nonprobability online panels. In M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research* (pp. 219-237). John Wiley. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118763520.ch10/summary

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence

in parametric causal inference. *Political Analysis*, *15*, 199-236. doi:10.1093/pan/mpl013

Howe, L. D., Hargreaves, J. R., Ploubidis, G. B., De Stavola, B. L., & Huttly, S. R. A. (2011). Subjective measures of socio-economic position and the wealth index: A comparative analysis. *Health Policy and Planning*, *26*, 223-232. doi:10.1093/heapol/czq043

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research* (Vol. 2, pp. 102-138). New York: The Guilford Press. .

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013). Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 207-216). New York, NY: ACM. doi:10.1145/2470654.2470684

Krupnikov, Y., & Levine, A. S. (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science*, *1*, 59-80. doi:10.1017/xps.2014.7

Lakkaraju, K. (2015). A study of daily sample composition on Amazon Mechanical Turk. In N. Agarwal, K. Xu, & N. Osgood (Eds.), *Social computing, behavioral-cultural modeling, and prediction* (pp. 333-338). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-16268-3_39

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433-442. doi:10.3758/s13428-016-0727-z

Mavletova, A. (2013). Data quality in PC and mobile web surveys. *Social Science Computer Review*, *31*(6), 725-743.. doi:10.1177/0894439313485201

Moore, P. (2015). *A third of young Americans say they aren't 100% heterosexual*. YouGov. Retrieved from https://today.yougov.com/news/2015/08/20/third-young-americans-exclusively-heterosexual/

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*, 109-138. doi:10.1017/XPS.2015.19

Paolacci, G., & Chandler, J. (2014). Inside the turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*, 184-188. doi:10.1177/0963721414531598

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). *Running experiments on Amazon Mechanical Turk* (SSRN Scholarly Paper No. ID 1626226). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1626226

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

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*, 1023-1031. doi:10.3758/s13428-013-0434-y

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research*, *67*, 1035-1041. doi:10.1016/j.jbusres.2013.08.010

Pew Research Center. (2016a). *A new estimate of the U.S. Muslim population*. Retrieved from http://www.pewresearch.org/fact-tank/2016/01/06/a-new-estimate-of-the-u-s-muslim-population/

Pew Research Center. (2016b). *10 facts about atheists*. Retrieved from http://www.pewresearch.org/fact-tank/2016/06/01/10-facts-about-atheists/

Ravallion, M., & Lokshin, M. (1999). *Subjective economic welfare*. The World Bank. Retrieved from http://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-2106

Reidy, D. E., Berke, D. S., Gentile, B., & Zeichner, A. (2014). Man enough? Masculine discrepancy stress and intimate partner violence. *Personality and Individual Differences*, *68*, 160-164. doi:10.1016/j.paid.2014.04.021

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213-220. doi:10.1177/2167702612469015

Sigman, R., Lewis, T., Yount, N. D., & Lee, K. (2014). Does the length of fielding period matter? Examining response scores of early versus late responders. *Journal of Official Statistics*, *30*, 651-674. doi:10.2478/jos-2014-0042

Sommer, J., Diedenhofen, B., & Musch, J. (2016). Not to be considered harmful: Mobile-device users do not spoil data quality in web surveys. *Social Science Computer Review*, *35*(3): 378-387. doi:10.1177/0894439316633452

Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, *10*, 479-491.

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The effects of questionnaire completion using mobile devices on data quality: Evidence from a probability-based general population panel. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology*, *9*, 261-292. doi:10.12758/mda.2015.014

Sturgis, P., Allum, N., & Brunton-Smith, I. (2009). Attitudes over time: The psychology of panel conditioning. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 113-126). John Wiley. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470743874.ch7/summary

U.S. Census Bureau. (2011-2015). *American community survey 5-year estimates*. Retrieved from https://www.census.gov/data/developers/data-sets/acs-5year.html

U.S. Census Bureau. (2016). *Current population survey, annual social and economic supplement*. Retrieved from https://www.census.gov/cps/data/cpstablecreator.html

Voigt, L. F., Koepsell, T. D., & Daling, J. R. (2003). Characteristics of telephone survey respondents according to willingness to participate. *American Journal of Epidemiology*, *157*, 66-73. doi:10.1093/aje/kwf185

Weinberg, J., Freese, J., & McElhattan, D. (2014). Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample. *Sociological Science*, *1*, 292-310. doi:10.15195/v1.a19

Wells, T., Bailey, J., & Link, M. (2013). Filling the void: Gaining a better understanding of tablet-based surveys. *Survey Practice*, *6*(1). Retrieved from http://www.surveypractice.org/index.php/SurveyPractice/article/view/25

Author Biographies

Logan S. Casey received his PhD from the University of Michigan. He is a research analyst in Public Opinion at the Harvard Opinion Research Program in the Harvard T.H. Chan School of Public Health. His research examines political psychology, emotion, and public opinion, particularly in the context of LGBTQ politics.

Jesse Chandler received his PhD from the University of Michigan. He is a researcher at Mathematica Policy Research and adjunct faculty at the Institute for Social Research. He is interested in survey methodology, online research studies, decision-making, and human computation.

Adam Seth Levine is an assistant professor of Government at Cornell University.

Andrew Proctor is a doctoral candidate in the Department of Politics at Princeton University. His primary research interests are in LGBT politics, identity and political mobilization.

Dara Z. Strolovitch is an associate professor at Princeton University, where she holds appointments in Gender and Sexuality Studies, African American Studies, and the Department of Politics. Her teaching and research focus on interest groups and social movements, political representation, and the intersecting politics of race, class, gender, and sexuality. Her book, *Affirmative Advocacy*, addressed these issues through an examination of the ways in which advocates for women, people of colour, and low-income people represent intersectionally marginalized subgroups of their constituencies.