Intertemporal Differences Among MTurk Worker Demographics

Logan S. Casey[1], Jesse Chandler[2,3*], Adam Seth Levine[4],
Andrew Proctor[5], Dara Z. Strolovitch[5,6]

[1] Department of Political Science, University of Michigan, Ann Arbor, Michigan, United States of America
[2] Research Center for Group Dynamics, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, United States of America
[3] Survey Research Division, Mathematica Policy Research, Ann Arbor, Michigan, United States of America
[4] Department of Government, Cornell University, Ithaca, New York, United States of America
[5] Department of Politics, Princeton University, Princeton, New Jersey, United States of America
[6] Program in Gender and Sexuality Studies and Department of Politics, Princeton University, Princeton, New Jersey, United States of America

* Corresponding author
Email: jjchandl@umich.edu (JC)

Abstract

The online labor market Amazon Mechanical Turk (MTurk) is an increasingly popular platform for generating samples of respondents for social science research. A growing body of research has examined the demographic composition of MTurk workers, typically comparing samples of MTurk workers to samples of respondents drawn from other populations. While these comparisons have revealed important information about 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 paper focuses on whether MTurk sample composition varies as a function of time. Using an original dataset of nearly 10,000 MTurk workers, we examine whether demographic characteristics vary by (1) time of day, (2) day of week, and (3) serial position (i.e., earlier or later in data collection). 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, including characteristics known to impact political attitudes and psychological processes. This demonstrates that MTurk samples cannot be presumed identical across different studies, and we suggest several forms of variation to which researchers using MTurk ought to be attentive.

1. Background

Amazon Mechanical Turk (MTurk) is an online labor market in which people ("requesters") requiring the completion of small tasks ("Human Intelligence Tasks," hereafter "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 over 500 papers that relied on Mechanical Turk data in full or in part [1].

Reflecting the popularity of MTurk, considerable effort has been invested in assessing the characteristics and quality of the data collected using it by, for example, 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 best-studied convenience samples currently available to researchers (for a review see [1]), and we have learned a great deal about the ways in which MTurk respondents are and are not representative of the general population. There are reasons to suspect, however, that there are also important variations *among* samples drawn from MTurk, and these variations have received far less attention. This paper takes up this question, using data from a study of approximately 10,000 MTurk workers to examine whether MTurk sample composition varies as a function of the time that it is collected.

We begin below by reviewing what extant research has revealed about the demographic composition of the MTurk worker pool and about how samples drawn from it compare to samples of respondents drawn from other populations. Next, we describe the methods and measures that we use in our study, after which we present the results of our analyses exploring whether the demographic characteristics of MTurk respondent samples vary by (1) time of day,

(2) day of week, and (3) serial position (i.e., whether a respondent completes the survey earlier or later in the data collection process). We conclude with a discussion about what the temporal variations we uncover suggest about several issues to which researchers using MTurk ought to be attentive.

1.1. How Representative of the General Population are Samples of MTurk Workers?

The demographic characteristics of samples drawn from MTurk populations - particularly U.S. workers – have been extensively studied. These studies show that most MTurk workers live in the United States and India [2], that MTurk workers are more diverse than many other convenience samples, and that they are not representative of the population as a whole [3]. However, while scholars caution that MTurk samples are typically less representative than commercial web panels that make explicit efforts to provide representative samples [4-6], they also agree that MTurk samples compare favorably in terms of diversity to student samples or community samples recruited from college towns [4,7].

Scholars have also found that 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 [3,8,9]. Most significantly, MTurk workers are typically younger than the general population [2,4]. MTurk workers therefore also differ from the population as a whole in ways that correlate with age and cohort differences. For example, MTurk workers tend to report that they have more years of formal education and that they are more liberal [4,5], less likely to be married [4,10], and more likely to identify as lesbian, gay or bisexual (LGB) [10-12]. MTurk workers also tend to report lower personal incomes and are more likely to be unemployed or underemployed than members of general population [10,11].

Whites and Asian Americans are overrepresented within MTurk samples, while Latinos and African Americans are underrepresented [4]. These latter differences may reflect the "digital divide" in which internet access (particularly high speed internet) varies across racial and economic groups [3].

1.2. Are Samples of MTurk Workers Representative of MTurk Workers?

While the forgoing research makes clear that the MTurk population is not representative of the 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. Significant differences in the demographic characteristics are occasionally observed. For example, the proportion of female respondents differed by about 10% across two studies that each recruited several thousand participants [1]. Some of this variation is likely the result of self-selection biases, as participants in MTurk research likely self-select into studies that interest them (for a discussion see Couper [13]).

Anecdotal evidence suggests that MTurk sample composition might also be influenced by the fact that workers share information about available studies and that reputation effects might lead workers to gravitate towards (and to avoid) particular requestors [14]. Indeed, there are several online communities dedicated to publicizing or discussing particular tasks or surveys posted to MTurk and to rating various dimensions of the requesters who publish these tasks.[1] Design choices that are exogenous to a study itself may also inadvertently influence sample composition. The effects of such choices are of particular interest to researchers because they are both within their control and typically irrelevant to the substance of the studies themselves.

Extant evidence about intertemporal variation is suggestive but limited by small sample

[1] See, for example, Turker Nation, Turkopticon, MTurk Grind and Reddit's HITsWorthTurkingFor.

sizes. Comparing samples of about 100 participants obtained within two different studies, Komarov and colleagues [15] observed that compared to 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 completing surveys using mobile devices instead of computers). In a study using a more systematic but nonetheless still small sample, Lakkaraju [16] 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.

Another possible source of temporal variation among participants might be observed between those who complete a research study early or later in the data collection process (referred to here as serial position effects). Changes in sample composition between "early" and "late" responders have been observed in other modes of data collection including both mail and email surveys, which seem to be a function of how difficult it is to recruit particular populations of respondents (for a review see [17]). In general, people of color are underrepresented among early respondents, as are men [17-19], younger people, and people with fewer years of formal education [19] (for a discussion see [17]). Examinations of lab studies also show that sample compositions can vary over time. For example, women [20] and students with high GPAs [21,22] 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 than their more extroverted, open, and careless counterparts [22].

Investigating whether samples vary dynamically is critical because researchers tend to

recruit small samples for their research [23]. Further, 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 [24], calling into question whether they can be used as the basis for general claims about the MTurk worker pool. If researchers also prevent workers from completing multiple experiments (as they should; see Chandler et al [14,25]), sample composition is also likely to vary systematically across experiments, compromising both the reliability and validity of their studies and complicating efforts to reproduce findings.

2. Method

To explore whether MTurk worker demographics vary intertemporally, we crafted a brief HIT (average completion time was approximately five minutes) that contained demographic questions that are of interest to scholars across an array of disciplines. Our central goal was to explore whether the demographics of MTurk worker samples vary based on the day of the week and the time of day that a HIT is posted. We also sought to examine differences based on the serial position in which respondents complete they survey. 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 3pm and 10pm EST. After the first week we added a third posting at 10am.

Only U.S. based workers with a HIT acceptance ratio (HAR) greater than 95% who had completed at least 50 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 [26] and, in our experience, to be favored by researchers. The relatively low minimum requirement of 50

completed HITs was adopted to ensure that workers with a wide range of experiences were eligible to complete the HIT. We prevented workers from completing this survey more than once across the entire fielding period. For the first three weeks, workers were paid $0.25 to complete the survey. For the remainder of the fielding period, workers were paid $0.50 to increase the completion rate. This latter pay rate is consistent with minimum acceptable pay norms of $0.10 per minute. By the end of the study, we had posted the HIT 162 times and sampled 9,770 unique respondents.

3. Measures

To understand the contours of the MTurk worker pool, our study included questions about a number of demographic variables that are of interest to a wide social science audience. We also included several measures that are of interest to particular groups of researchers, including the "Big Five" personality factors (popular among psychologists), party identification and political ideology (of interest to many political scientists), and several measures that allowed us to identify members of hard-to-reach populations (such as lesbian, gay, bisexual, and transgender -- or LGBT -- people) so that we could recruit them for future studies.

As a condition of participation, subjects had to be at least 18 years old and U.S. residents. Thus, at the very beginning of the study we collected measures of age and the U.S. state in which the respondent resides. 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 [27,28]). We also asked about respondents' contact with lesbian, bisexual, and gay people, contact with transgender people, and the contexts in which respondents knew LGBT people.

We also included a series of questions intended to measure the "Big Five" personality factors. The "Big Five" is among the most widely accepted taxonomy of personality traits within psychology (for a review, see John & Srivastava [29]), and conceptualizes personality as consisting of five bipolar dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. To operationalize these dimensions, we use the "Ten Item Personality Index," or TIPI -- a ten item measure of these dimensions that has been shown to have adequate reliability [30; please see supplementary materials for a full list of questions and exact wordings].

Finally, using a database of over 100,000 submitted HITs (reported in Stewart et al [24]), we were able to identify the number of HITs each worker had completed and used this information to estimate their prior experience completing MTurk surveys. This measure allows us to analyze potential temporal variations in workers' levels of experience (see [25]).

4. Results

4.1. Characteristics of the MTurk sample

Tables 1-4 present summary data about the entire sample and data for participants in the first two batches. The entire sample represents the largest sample of MTurk workers we are aware of, and likely measures about two thirds of the active worker population [24]. The sample size of the first two batches (N=438) approximates a sample slightly larger than those typically used in behavioral science research [23]. The demographic data are reported in Table 1.

Differences between this sample and the U.S. population as a whole are generally consistent with those reported in previous analyses of smaller surveys [2,4,7,10]. For example, the workers in our sample are younger and whiter than the U.S. population as a whole. Workers residing in the Eastern Standard Time Zone are overrepresented compared with those in other parts of the U.S. 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.

The socioeconomic characteristics of the workers in our sample are summarized in Table 2. 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 US income than would be expected from previous measurements of individual worker income [2,4]. 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 disproportionate number of millennials with low individual income but who are living with their higher income parents. The proportion of people working full time in this sample is roughly ten percent higher than the proportion reported in Shapiro et al's 2013 study [10], suggesting that improving labor conditions may have also contributed to the higher income of the respondents in our study.

Relationship status and characteristics of respondents are summarized in Table 3, and reveal that approximately a third of respondents are married and that a third are single. As has been observed in other studies of other MTurk workers [10-12], the proportion of lesbian, gay, and particularly bisexual respondents is higher than it is in the US population as a whole. This is likely because online populations are disproportionately young, and younger people are also

more likely to identify as LGB [31,32].

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 [4,5]. Relatively few workers identified as religious, and reported rates of church attendance are generally low. Relative to normed data obtained from a large convenience sample of Internet users [33], MTurk workers were 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, while only 4.5% of participants completed the survey using a tablet, 2% using a phone, with the remaining using other devices (such as 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 [34,35].

4.2 Demographic Differences by Time of Completion

The focus of our investigation is to determine whether the composition of the MTurk worker pool varied across days of the week, across times of day, or across the serial order in which they participated -- that is whether those who were first to complete the survey differed from those who complete the survey later. To determine whether any such deviations are evident in our sample, we looked for variations within the following variables: age, gender identity, education, employment, household income, household size, Latino ethnicity, socio-economic 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 variable. We also looked for differences in the Big 5 personality traits: extraversion, agreeableness, conscientiousness, emotional stability, and openness. Finally, we estimated each worker's prior experience by recording the total number of times they appeared in the dataset of completed HITs (described above; [24]). When several variables examining the same topic were likely to be highly correlated (e.g., political ideology and party affiliation), only one was selected for analysis. To further reduce the number of comparisons, some response options were collapsed into broader categories (e.g. religious identifications). In total, given the coding, we ended up with 30 different variables.

For all continuous, ordinal, and binomial variables, generalized linear modeling (GZLM) was used to regress the day of the week and time of day the batch was posted (categorical variables) on the variable of interest. The serial position of the batch within the data collection run was included as a continuous variable. A dichotomous variable representing the amount of compensation was included to control for possible effects of increasing payment part way through the study. Continuous 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 limit 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 to individual comparisons against all other days or times). For

example, we compared the mean percentage of college graduates in batches posted on Tuesdays to the mean percentage of college graduates in all batches (including Tuesdays). While we took steps to limit the number of comparisons, this analysis strategy still led to a total of 12 significance tests for each of the 30 variables, for a total of 360 comparisons.

Concerned about the possibility of alpha inflation from so many significance tests on a large sample, we took several additional steps to reduce the potential for false positives. First, we set the alpha criterion at 0.01 instead of the more typical liberal alpha criterion at 0.05 that is typically used. Second, we used the Benjamini-Hochberg adjustment [36] to hold the false discovery rate across all comparisons constant at .01. Following these adjustments, no results with an unadjusted p-value above .00084 are reported as statistically significant, and of the significant results that we report, fewer than four are likely to be false positives observed by chance alone.[2] Table 5 includes the 25 statistically significant differences among our 360 comparisons.

4.2.a. Time of day effects. Of our 90 time-of-day comparisons, we found 10 instances in which attributes of subjects recruited at a particular time of day differed significantly from the grand mean.[3] These differences generally reflected linear trends in the composition of the MTurk workforce throughout the day.

As we might expect, 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 to complete HITs posted at 10am (β = -.16 Wald χ^2 = 74.10, $p < .0001$, d = .17). Conversely, people in later time zones were more likely to complete HITs posted at 10pm (β = .13 Wald χ^2 = 70.01, $p < .0001$, d = .17). As an illustration of the consequences of this shift,

[2] The Benjamini-Hochberg adjustment does not identify specific false positives, but rather holds the number of false positives across many tests to a specified level.
[3] Three times of day (10am, 3pm, 10pm) by thirty demographic variables produces ninety comparisons.

56.8% of respondents at 10am 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 10pm Eastern Time reside in the U.S. Eastern time zone, while 18.9% of workers were from the U.S. Pacific time zone.

The time of day is also related to the proportion of workers completing the survey on smartphones. Workers completing the survey later in the day were more likely to do so using cellphones than those responding earlier ($\beta = .014$ Wald $\chi^2 = 18.37$, $p < .001$ $d = .09$), with 5.8% of HITs posted at 10pm submitted from mobile phones as compared to 3.7% of HITs submitted during the rest of the day. This effect remained significant when controlling for time zone.

The proportion of single workers increased linearly throughout the day from 29.1% at 10am to 32.2% at 3pm to 34.9% at 10pm. The proportion of single workers was significantly lower than average at 10am ($\beta = -.03$ Wald $\chi^2 = 16.69$) and significantly higher at 10pm compared to the average ($\beta = .03$ Wald $\chi^2 = 16.68$, ps $< .001$, $d = .08$). This effect remained significant after controlling for time zone.

The proportion of Asian American respondents also increased over the course of the day, growing from 5.9% at 10am to 7.6% at 3pm to 9% at 10pm. The proportion of Asian Americans was significantly lower than average at 10am ($\beta = -.015$ Wald $\chi^2 = 16.09$) and significantly higher than average at 10pm ($\beta = .016$ Wald $\chi^2 = 16.13$, ps $< .001$, $d = .08$). This effect was no longer significant, however, when controlling for time zone, suggesting that it was a function of the fact that Asian Americans are more likely to live on the west coast.

In addition, workers recruited at 10pm reported being less conscientious than those completing the survey earlier in the day ($\beta = .06$ Wald $\chi^2 = 12.57$, $p < .001$, $d = .07$), as the mean conscientiousness reported at 10pm was 5.27 (SD = 1.25) as compared to a grand mean of 5.18

(SD = 1.31). After controlling for time zone, the difference 10pm and the grand mean was virtually unchanged (β = .06 Wald χ^2 = 11.18, p < .001, d = .07).

Finally, relative to the grand mean, more experienced workers were more likely to complete HITs in the morning (β = .51 Wald χ^2 = 37.62, p < .0001, d = .12). Experience was also negatively related to responding to the survey at night (β = -.50 Wald χ^2 = 56.93, p < .0001, d = .15). For example, a typical worker completing the survey at 10am had completed 3.6 HITs (SD = 2.12) while a typical respondent completing the survey at 10pm had completed only 2.71 HITs (SD = 1.74). These effects remained significant after controlling for time zone.

In sum, MTurk workers who completed our study at 10am are more likely to come from Eastern time zones, less likely to be single, less likely to be Asian American, and are more experienced MTurk workers. Conversely, those completing the survey at 10pm are more likely to live in Western time zones, more likely to be single, more likely to be Asian American, and have less MTurk work experience. Late night workers are also more likely to complete HITs using a phone and report lower levels of conscientiousness than workers recruited throughout the rest of the day. Some of these results (particularly differences in the proportion of Asian Americans) are driven by the proportions of workers from different time zones at different times of day.

4.2.b. Day of Week Effects. Of our 210 day-of-week comparisons, we found five instances in which the attributes of subjects recruited on a particular day of the week significantly differed from the sample as a whole.[4]

The average age of respondents varied as a function of the day of the week. The mean age of respondents was 33.51 (SD = 11.31). Participants on Wednesday (M = 32.4, SD = 10.78) and Thursday (M = 32.46 SD = 10.67), (β = -1.14 Wald χ^2 = 15.33, p < .001, d = .08 and β = -

[4] Seven days by thirty variables produces 210 comparisons.

1.67 Wald $\chi^2 = 47.28$, $p < .0001$, $d = .14$, respectively) were somewhat younger than those responding on other days. Respondents completing the survey on Saturday were somewhat older than those doing so during the rest of the week (M = 35.87, SD = 12.47) ($\beta = 2.05$ Wald $\chi^2 = 42.51$, $p < .0001$, $d = .13$).

People completing HITs on Sundays were more likely to be employed full time than those completing the survey on other days ($\beta = .10$ Wald $\chi^2 = 14.24$ $p < .001$, $d = .08$). Workers with full time jobs were more likely to complete HITs posted on Sundays (52% as compared to an average of 48.5% across the total sample), with a corresponding decrease in the proportion of individuals without any formal employment (31.2% as compared to 35.7%). The proportion of workers employed part time remained roughly the same across all days of the week. Asian Americans also made up a substantially larger proportion of the sample on Wednesdays (10.5%) relative to their average for all other days (7.7%) ($\beta = .03$ Wald $\chi^2 = 12.57$, $p < .001$, $d = .07$).

In sum, we find few day of week effects. Younger workers are more likely to complete the survey on Wednesdays and Thursdays, while those doing so on Saturdays are more likely to be older. Asian American workers are more likely to have completed the survey on Wednesdays than on other days of the week, while people employed full-time are more likely to have completed the survey on Sundays.

4.2.c. Serial Position Effects. Of our thirty positional comparisons, we found eight instances in which the attributes of subjects recruited earlier during our fielding period significantly differed from their grand means.[5] Workers who completed HITs earlier in the data collection process reported higher levels of emotional stability ($\beta = .003$ Wald $\chi^2 = 37.44$ $p < .001$, $d = .12$), higher levels of conscientiousness ($\beta = .002$ Wald $\chi^2 = 20.66$ $p < .001$, $d = .09$), and higher levels of agreeability ($\beta = .002$ Wald $\chi^2 = 18.75$ $p < .001$, $\underline{d} = .09$). Participants who

[5] Thirty demographic variables, treating time as a linear effect by batch number.

completed earlier batches of HITs also tended to be older (β = -.018 Wald χ^2 = 15.33, p < .001, d = .08), were more likely to have a full time job (β = .003 Wald χ^2 = 15.55, p < .001, d = .08), came from smaller households (β = .002 Wald χ^2 = 13.32, p < .001, d = .08), and were more likely to be male (β = .003 Wald χ^2 = 12.43, p < .001, d = .07). Workers who completed HITs sooner were substantially more experienced than workers recruited later in the study (β = -.009 Wald χ^2 = 405.94, p < .0001, d = .41). The practical implications of serial position effects are illustrated in Table 6, which includes point estimates of these variables at different points in the data collection run.

4.2.d. Pay Effects. Pay effects were included primarily to control for a change in design part way through data collection. Of the thirty payment comparisons, we found evidence of only two characteristics that changed once we offered to pay more. Following the increase in pay, average levels of reported emotional stability increased (β = .187 Wald χ^2 = 11.73, p < .001, d = .07), as did average worker experience (β = .31 Wald χ^2 = 53.17, p < .0001, d = .15).

5. Discussion

In this paper we have described demographic characteristics of a large sample of MTurk workers and examined differences across time, day, payment amount, and serial position. Of our 360 demographic comparisons, we found 25 differences (6.9% of tested effects), with effect sizes ranging from d = 0.07 to d = 0.41 (only a small number of which – 4 – are likely be due to chance). These findings provide evidence that MTurk samples do vary intertemporally. 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.

5.1 Demographic Differences by Day and Time

Day of the week influenced few (2%) demographic characteristics, and these effects were small (d = 0.07-0.13). 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 somewhat larger (d = 0.07-0.17) deviations on a larger proportion (10%) of measured variables. In almost all cases, these differences represented linear trends in sample composition across the day, and differences become much larger when comparing data collected from samples recruited at 10am as compared to 10 pm. While some of the time differences are intuitive (e.g. differences in worker location), others are less so (e.g. variations in conscientiousness) or contradict prior research (e.g. increased use of mobile devices at night; see also [15]). The large proportion observed differences and suggest that time of day effects might be a fruitful area of future research, especially on variables that are correlated with time zone differences.

The findings regarding time of day are particularly relevant for researchers interested in ensuring geographical diversity in their sample and they suggest several strategies for doing so. This diversity can be increased by posting studies at different times of day, for example, or by targeting recruitment to workers in specific states rather than to US workers as a whole. Other demographic differences are a result of differences in worker activity across the day within time zones (e.g. more single workers at night) and thus also require consideration of when HITs are posted.

Contrary to previous research [15], we found that workers were more likely to use mobile devices late at night (5.8% of HITs posted at 10pm were submitted from mobile phones, compared to 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 [37-39] and shorter and fewer open-ended responses [37,40]. As a result, researchers might consider adjusting the time of day at which they post research studies or during which workers are allowed to complete them if they hope to optimize mobile completion.

5.2 Demographic Differences by Serial Position

The effects of serial position were small but far more extensive than time-of-day and day-of-week effects. Almost 25% of the observed variables exhibited serial position effects ranging in size from $d = .07$ to $d = .12$, suggesting that the composition of samples changes as sample size increases. Again, some of these findings are intuitive and replicate as we might expect from previous research, such as work that shows that respondents who report higher levels of conscientiousness respond earlier in a study [22]. Other findings are less intuitive (e.g., early responders were more likely to come from smaller households) or even counterintuitive (e.g. early responders were more likely to be men, in contrast to literature showing women are more likely to respond to surveys first [17-19]). Taken together, these findings provide compelling evidence that early responders to web surveys differ from late responders and suggest that serial position effects are also a fruitful area of additional inquiry.

In contrast to other studies that find that women are more likely to respond to requests to complete both mail surveys [18] and web surveys [17] quickly [20,21], we find that the proportion of women increases as data collection progresses. Other findings are more consistent

with differences between early and late responders observed in other modes of data collection. Early responders in general population samples tended to be older [17,41]. Studies of intertemporal differences in subject pool populations typically find that early responders report higher levels of conscientiousness [20,22]. Thus these latter findings converge nicely with observations in other samples and other modes of data collection.

On a practical level, variations resulting across serial position are 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 psychological variables including political behaviors and attitudes and consequently might bias samples (for an excellent review, see Gerber et al [42]).

Of general interest to all researchers is the possibility that changes in self-reported personality characteristics are accompanied by behavioral changes that influence data quality. Conscientiousness, for example, may be associated with worker care and diligence when completing questionnaires and may therefore also be related to individual completion rates. Agreeableness may be associated with respondents' propensity toward giving socially desirable answers.

Future research could fruitfully examine differences in experimental effect sizes between early and late responders. Serial position-based variation may also be relevant to researchers who use MTurk to recruit members of specific populations. Many of the known differences between Mechanical Turk workers and the general population are exacerbated as sample sizes increase,

suggesting a potential tradeoff between a larger sample and a more representative sample.[6]

5.3 Differences in Worker Experience

Time of day and serial position were also related to how much MTurk experience respondents had. Although we did not vary pay rates experimentally, we did find that when we increased pay, there was a concomitant increase in the experience of survey participants. Together, we thus observed two separate patterns: more experienced workers completed the survey earlier in data collection, but at the same time, we saw an increase in more experienced workers when we increased the pay rate after the first three weeks of data collection. And although it might be desirable under some circumstances to have more experienced respondents, they may compromise findings as greater exposure to survey tactics, experimental manipulations, or research questions can lead to practice effects [14], to smaller effect sizes on commonly used experimental paradigms [25], and to more extreme and less malleable attitudes towards topics they are frequently asked about [43]. If researchers are concerned that worker savviness might affect their findings [7], they should be attentive to these possibilities when they post their studies, and might also test whether time of day and serial position are correlated with worker experience and with any other variable of interest.

6. Conclusion

This study is the largest and most comprehensive description of MTurk demographics that we are aware of. Data from our study of approximately 10,000 MTurk workers has allowed us to examine three key possible sources of temporal variation in MTurk sample composition:

[6] Levay et al [44] provide guidance on important demographic variables and weighting methods to further improve MTurk sampling and inferences.

(1) time of day, (2) day of week, and (3) serial position in which a survey is posted. Taken as a whole, our results should serve as a source of both comfort and caution to scholars who use MTurk to recruit subjects for their research. On the one hand, we found only minimal day-of-week differences. However, we showed that 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 [42]). Differences in sample composition can compromise claims to generalizability and might lead to challenges with reproducing research findings as well [45].

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" [46,47]. But researchers interested in other topics should pay attention to relationships such as those between serial position and psychological characteristics and experience completing research studies and might consider including information about when they posted their HIT when reporting results.[7]

As MTurk and other similar online convenience samples become more widely available and 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

[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 [48].

limited to) their generalizability across populations and over time. Additionally, while we focused on MTurk in this paper because it is one of the most widely used online data collection platforms, the fact that we observed intertemporal differences in sample composition suggests this may also occur in other sources of online data. This is an important area for future research to examine, particularly as researchers continue or increase reliance on online data collection.

REFERENCES

1. Chandler J, Shapiro D. Conducting clinical research using crowdsourced convenience samples. Annual Review of Clinical Psychology. 2016;12: 9.1-9.29. DOI: 10.1146/annurev-clinpsy-021815-093623.2
2. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on Amazon Mechanical Turk. Judgment and Decision Making. 2010;5: 411-419.
3. Paolacci G, Chandler J. Inside the Turk: Understanding Mechanical Turk as a Participant Pool. Current Directions in Psychological Science. 2014; 23(3):184–88.
4. Berinsky AJ, Huber GA, Lenz GS. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. Political Analysis. 2012; 20:351-368.
5. Mullinix K, Leeper T, Druckman J, Freese J. The Generalizability of Survey Experiments. Journal of Experimental Political Science. 2015;2(2):109-138. DOI: 10.1017/XPS.2015.19.
6. Weinberg JD, Freese J, McElhattan D. Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample. Sociological Science 2014;1:292-310.
7. Krupnikov Y, Levine AS. Cross-Sample Comparisons and External Validity. Journal of Experimental Political Science. 2014;1(1): 59-80.
8. Casler K, Bickel L, Hackett E. 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. 2013; 29:2156-2160. DOI:10.1016/j.chb.2013.05.009.
9. Hillygus DS, Jackson N, Young M. Professional respondents in non-probability online samples. In: Callegaro M, Baker R, Bethlehem J, Göritz AS, Krosnick JA, Lavrakas PJ, editors. Online Panel Research: A Data Quality Perspective. John Wiley & Sons, Inc. 2014. p. 219-237.
10. Shapiro DN, Chandler J, Mueller PA. Using Mechanical Turk to study clinical populations. Clinical Psychological Science. 2013;1(2):213-220. DOI: 10.1177/2167702612469015.
11. Corrigan PW, Bink AB, Fokuo JK, Schmidt A. The public stigma of mental illness means a difference between you and me. Psychiatry Research. 2015;226(1):186-91. DOI: 10.1016/j.psychres.2014.12.047.
12. Reidy DE, Berke DS, Gentile B, Zeichner A. Man enough? Masculine discrepancy stress and intimate partner violence. Personality and Individual Differences. 2014;68:160-164. DOI: 10.1016/j.paid.2014.04.021.
13. Couper, MP. Review: Web surveys: A review of issues and approaches. The Public Opinion Quarterly. 2000; 64(4):464-494.
14. Chandler J, Mueller P, Paolacci G. Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. Behavioral Research Methods. 2014;46:112-130.
15. Komarov S, Reinecke K, Gajos KZ. Crowdsourcing performance evaluations of user interfaces. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 31st. 2013, Paris. ACM: New York. 207-216. DOI: 10.1145/2470654.2470684.
16. Lakkaraju K. A Study of Daily Sample Composition on Amazon Mechanical Turk. In Agarwal N, Xu K, Osgood N, editors. Social Computing, Behavioral-Cultural Modeling, and Prediction. New York: Springer International Publishing; 2015. p. 333-338.

17. Sigman R, Lewis T, Yount ND, Lee K. Does the Length of Fielding Period Matter? Examining Response Scores of Early Versus Late Responders. Journal of Official Statistics. 2014;30(4): 651-674. DOI: 10.2478/jos-2014-0042.
18. Gannon MJ, Nothern JC, Carroll SJ. Characteristics of Nonrespondents Among Workers. Journal of Applied Psychology. 1971;55: 586-588. DOI: http://dx.doi.org/10.1037/h0031907.
19. Voigt LF, Koepsell TD, Daling JR. Characteristics of telephone survey respondents according to willingness to participate. American Journal of Epidemiology. 2003;157: 66–73.
20. Ebersole CR., Atherton OE, Belanger AL, Skulborstad HM, Adams RB, Allen J, Nosek BA. "Many Labs 3: Evaluating participant pool quality across the academic semester via replication." 2015. Retrieved from osf.io/ct89g.
21. Cooper H, Baumgardner AH, Strathman A. Do students with different characteristics take part in psychology experiments at different times of the semester? Journal of Personality 1991;59: 109–127.
22. Aviv AL, Zelenski JM, Rallo L, Larsen RJ. Who Comes When: Personality Differences in Early and Later Participation in a University Subject Pool. Personality and Individual Differences 2002; 33: 487-496.
23. Fraley RC, Vazire S. The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. PLoS One. 2014;9(10): e109019. DOI: 10.1371/journal.pone.0109019
24. Stewart N, Ungemach C, Harris AJL, Bartels DM, Newell BR, Paolacci G, Chandler J. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. Judgment and Decision Making. 2015;10(5): 479-491.
25. Chandler J, Paolacci G, Peer E, Mueller P, Ratliff KA. Using Nonnaive Participants Can Reduce Effect Sizes. Psychological Science. 2015; 26(7): 1131–39.
26. Peer E, Vosgerau J, Acquisti A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. Behavior Research Methods. 2014;46(4), 1023-1031.
27. Howe LD, Hargreaves JR, Ploubidis GB, De Stavola BL, Huttly SL. Subjective Measures of Socio-Economic Position and the Wealth Index: a Comparative Analysis. Health Policy and Planning. 2011;26(3): 223–32. DOI: 10.1093/heapol/czq043.
28. Ravallion M, Lokshin M. Subjective economic welfare. World Bank Policy Research Paper 2106. Washington, DC: World Bank. 1999.
29. John OP, Srivastava S. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research. 1999; 2:102-138.
30. Gosling SD, Rentfrow PJ, Swann WB. A very brief measure of the Big-Five personality domains. Journal of Research in personality. 2003;37(6): 504-528.
31. Gates G, Newport F. 2012. "Special Report: 3.4% of U.S. Adults Identify as LGBT." Gallup. http://www.gallup.com/poll/158066/special-report-adults-identify-lgbt.aspx
32. Moore P. "A third of young Americans say they aren't 100% heterosexual." 2015 Aug 20 [cited 2016 Mar 2]. In: YouGov. Available from: https://today.yougov.com/news/2015/08/20/third-young-americans-exclusively-heterosexual/
33. Gosling SD, Rentfrow PJ, Potter J. Norms for the Ten Item Personality Inventory. 2014. Unpublished Data.
34. De Bruijne M, Wijnant A. Mobile response in web panels. Social Science Computer

Review 2014;32(6): 728-742.
35. De Bruijne M, Wijnant A. Improving response rates and questionnaire design for mobile web surveys. Public Opinion Quarterly. 2014; 78(4): 951-962. DOI: 10.1093/poq/nfu046.
36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society 1995; Series B (Methodological): 289-300.
37. Mavletova, A. Data Quality in PC and Mobile Web Surveys. Social Science Computer Review. 2013; 31: 725-743. DOI: 10.1177/0894439313485201.
38. Sommer J, Diedenhofen B, Musch J. Not to Be Considered Harmful: Mobile-Device Users Do Not Spoil Data Quality in Web Surveys. Social Science Computer Review. 2016 DOI: 10.1177/0894439316633452.
39. Wells T, Bailey JT, Link MW. Filling the void: Gaining a better understanding of tablet-based surveys. Survey Practice. 2013; 6(1). ISSN: 2168-0094.
40. Struminskaya B, Weyandt K, Bosnjak M. The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-based General Population Panel. methods, data, analyses. 2015;9(2): 261-292. DOI: 10.12758/mda.2015.014.
41. Filion, F. Estimating Bias Due to Nonresponse in a Mail Survey. Public Opinion Quarterly. 1975;39: 482–492. DOI: http://dx.doi.org/10.1086/268245.
42. Gerber AS, Huber GA, Doherty D, Dowling CM. The Big Five Personality Traits in the Political Arena. Annual Review of Political Science. 2011;14(1): 265–87.
43. Sturgis P, Allum N, Brunton-Smith I. Attitudes over time: The psychology of panel conditioning. Methodology of Longitudinal Surveys. 2009; 113-126.
44. Levay KE, Freese J, Druckman JN. The Demographic and Political Composition of Mechanical Turk Samples. *SAGE Open.* 2016; 6(1):1-17. DOI: 10.1177/2158244016636433.
45. Peterson RA, Merunka DR. Convenience samples of college students and research reproducibility. Journal of Business Research. 2014;67(5): 1035-1041.
46. Aguinis HS, Lawal S. eLancing: A review and research agenda for bridging the science-practice gap. Human Resource Management Review. 2013; 23(1): 6–17.
47. Brawley AM, Pury CLS. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. Computers in Human Behavior. 2016;43:531-546. Doi: 10.1016/j.chb.2015.08.031.
48. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political analysis 2007; 15(3): 199-236.

Table 1 - Demographic Characteristics of Workers

	Total Sample (N = 9770)	First Respondents (N = 438)
Mean Age	33.51[32.3, 33.7]	33.59[32.6. 34.58]
Assigned Female at Birth	51.7%[50.7, 52.7]	46.9%[42.2,51.6]
Transgender	0.5%[0.3,0.6]	0.2%[0.0,0.6]
Gender Queer	0.9%[0.7, 1.1]	0.2%[0.0,0.6]
Mean Worker Experience (Prior HITs completed [24])	3.67[3.5,3.9]	6.94[5.82,8.06]
U.S. Time Zone		
Eastern	52.2%[51.2, 53.2]	56.2%[51.6,60.9]
Central	25.3%[24.4,26.2]	23.3%[19.3,27.3]
Mountain	5.9%[5.4,6.4]	3.9%[2.1,5.7]
Pacific	15.9%[15.2,16.6]	16.4%[12.9,19.9]
Other	0.6%[0.5,0.8]	0.2%[0.0,0.6]
Race & Ethnicity		
White/Caucasian	82.9%[82.2,83.7]	79.5%[75.7,83.3]
African American	8.6%[8.0,9.2]	7.8%[5.3,10.3]
Asian American	7.7%[7.2,8.2]	11.2%[8.3,14.1]
American Indian or Alaskan Native	2.1%[1.8,2.4]	3.2%[1.6,4.9]
Native Hawaiian or Pacific Islander	0.6%[0.4,0.8]	1.6%[0.4,2.8]
Other	1.3%[1.1,1.5]	0.9%[0.0,1.8]
Multi-racial	3.9%[3.5,4.3]	3.9%[2.1,5.7]
Latino	5.5% [5.1,6.0]	6.4%[4.1,8.7]

NOTE: 95% CI indicated in parentheses.

Table 2 - Socioeconomic Characteristics of Workers

	Total Sample (N = 9770)	First Respondents (N = 438)
Household Income		
<14,999	11.7%[11.1,12.3]	11%[8.1,13.9]
15,000-29,999	17.5%[16.8,18.3]	17.4%[13.9,21.0]
30,000-49,999	24.6%[23.8,25.5]	25.8%[21.7,29.9]
50,000-74,999	20.7%[19.9,21.5]	21.5%[17.7,25.3]
75,000-99,999	12.2%[11.6,12.9]	12.6%[9.5,15.7]
>$100,000	12.9%[12.2,13.6]	11.7%[8.7,14.7]
Household Size	2.82[2.79,2.85]	2.76[2.62,2.90]
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]
Working part-time	15.7%[15.0,16.4]	14.2%[10.9,17.5]
Homemaker	8.6%[8.0,9.2]	8%[5.5,10.5]
Unemployed	9.4%[8.8,10.0]	9.1%[6.4,11.8]
Retired	2.2%[1.9,2.5]	1.4%[0.3,2.5]
Student	11.9%[11.3,12.5]	7.5%[5.0,10.0]
Permanent Disability	1.9%[1.6,2.2]	2.3%[0.9,3.7]
Other	1.7%[1.4,2.0]	2.3%[0.9,3.7]
Education		
Less than High School	0.7%[0.5,0.9]	1.4%[0.3,2.5]
High School or GED	10.2%[9.6,10.8]	10.5%[7.6,13.4]
Some college	31.4%[30.5,32.3]	22.4%[18.5,26.3]
2 Year college degree	11.7%[11.1.12.3]	9.6%[6.8,12.4]
4 Year college degree	34.8%[33.9,35.7]	44.5%[39.9,49.2]
Postgraduate Degree	11.1%[10.5,11.7]	11.6%[8.6,14.6]

Table 3 - Relationship Characteristics of Workers

	Total Sample (N = 9770)	First Respondents (N = 438)
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 Non-Monogamous	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]
Married	34.9%[34.0,35.9]	29.2%[24.9,33.5]
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]
Divorced	6%[5.5,6.5]	7.3%[4.9,9.7]
Widowed	0.8%[0.6,1.0]	0.5%[0.0,1.2]
Sexual Orientation		
Lesbian or Gay	3.8%[3.4,4.2]	2.3%[0.9,3.7]
Bisexual	6.9%[6.4,7.4]	6.6%[4.3,8.9]
Straight	86.8%[86.1,87.5]	88.8%[85.9,91.8]
Other	2.2%[1.9,2.5]	2.1%[0.8,3.4]

Table 4 - Attitudinal and Personality Characteristics of Workers

	Total Sample (N = 9770)	First Respondents (N = 438)
Political Affiliation		
Identifies as Republican	17.90%[17.1,18.7]	18.3%[14.7,21.9]
Identifies as Democrat	41.30%[40.3,42.3]	47%[42.3,51.7]
Ideology (1 = Extremely liberal, 7 = Extremely conservative)	3.39[3.36,3.42]	3.31 [3.16,3.46]
Religion		
Christian – Mainline Protestant	16%[15.3,16.7]	13.3%[10.1,16.5]
Christian - Evangelical	8.5%[8.0,9.1]	8.6%[6.0,11.3]
Christian - Catholic	11.4%[10.8,12.0]	14.3%[11.0,17.6]
Christian - Other/not specified	10%[9.4,10.6]	7.3%[4.9,9.7]
Jewish	1.2%[1.0,1.4]	0.5%[0.0,1.2]
Muslim	0.6%[0.5,0.8]	1.4%[0.3,2.5]
Atheist	20.4%[19.6,21.2]	25.5%[21.4,29.6]
Nothing in particular	24.6%[23.8,25.5]	23.6%[19.6,27.6]
Other	7%[6.5,7.5]	5.3%[3.2,7.4]
Religiosity		
Attends at least weekly	9.2%[8.6,9.8]	6.9%[4.5,9.3]
Attends at least monthly	12.1%[11.5,12.8]	13.1%[9.9,16.3]
Attends a few times per year	24.2%[23.4,25.1]	22.6%[18.7,26.5]
Never attends	54.1%[53.1,55.1]	57.4%[52.8,62.0]
Big Five		
Extraversion	3.58 [3.55,3.61]	3.48[3.33,3.63]
Agreeableness	5.11 [5.09,5.13]	5.18[5.06,5.30]
Conscientiousness	5.24 [5.21,5.27]	5.40[5.28,5.52]
Emotional Stability	4.70 [4.67,4.73]	4.90[4.76,5.04]
Openness	5.09 [5.07,5.11]	4.86[4.74,4.98]

Table 5 – Significant results by time of day, day of week, serial position, and pay rate

This table includes the 25 comparisons that revealed statistically-significant differences. The entries in the table are arranged in ascending order of p-values. As noted in the text, we used the Benjamini-Hochberg adjustment for multiple comparisons and consider all p-values less than 0.00084 to be statistically significant (this ensures that the false discovery rate across all comparisons is held constant at .01).

DV	IV	Wald	df	p	Interpretation
Superturker	10am	37.62	1	<0.00001	Workers are more experienced at 10am
Superturker	10pm	56.93	1	<0.00001	Workers are less experienced at 10pm
Superturker	Pay	53.17	1	<0.00001	Workers more experienced once pay was increased
Superturker	Batch	405.94	1	<0.00001	Workers more experienced earlier in the data collection
Timezone	10am	74.09	1	<0.00001	Workers come from earlier time zones
Timezone	10pm	70.01	1	<0.00001	Workers come from later time zones
Age	Thurs	47.28	1	<0.00001	Workers were younger on Thursdays
Age	Sat	42.51	1	<0.00001	Workers were older on Saturdays
Emotional Stability	Batch	37.44	1	<0.00001	Workers more emotionally stable earlier in the data collection
Conscientiousness	Batch	20.66	1	0.00001	Workers more conscientious earlier in the data collection
Agreeableness	Batch	18.75	1	0.00001	Workers more agreeable earlier in the data collection
Phone	10pm	18.37	1	0.00002	Workers more likely to be using phones at 10pm
Age	Batch	18.01	1	0.00002	Workers were older earlier in the data collection
Relationship Status	10am	16.69	1	0.00004	Workers less likely to be single at 10am
Relationship Status	10pm	16.68	1	0.00004	Workers more likely to be single at 10pm
Asian American	10pm	16.13	1	0.00006	Workers more likely to be Asian at 10pm
Asian American	10am	16.09	1	0.00006	Workers less likely to be Asian at 10am
Employment Status	Batch	15.55	1	0.00008	Workers more likely to have full-time jobs earlier in the data collection
Age	Wed	15.33	1	0.00009	Workers were younger on Wednesdays
Employment Status	Sun	14.24	1	0.00016	Workers more likely to have full-time jobs; less likely to lack formal employment altogether (no change in part-time status)
Household Size	Batch	13.32	1	0.00026	Workers come from smaller households earlier in the data collection

Asian American	Wed	12.57	1	0.00039	Workers more likely to be Asian on Wednesdays
Conscientiousness	10pm	12.57	1	0.00039	Workers are less conscientious at 10pm
Gender	Batch	12.26	1	0.00046	Workers more likely to be male earlier in the data collection
Emotional Stability	Pay	11.73	1	0.00061	Workers more emotionally stable once pay was increased

Table 6 - Worker Characteristics as a Function of Serial Position

	Serial Position Batch 30 (-1SD)	Serial Position Batch 122 (+1 SD)	Linear Trend
Age	34.66 [34.29,35.03]	32.99[32.42,33.56]	$\beta = .018$ Wald $\chi^2 = 15.33$, $p < .001$, $d = .08$
Female (Gender identity)	50%[48,51]	56%[53,58]	$\beta = .003$ Wald $\chi^2 = 12.43$, $p < .001$, $d = .07$
Household Size	2.73[2.69,2.78]	2.92[2.85,2.99]	$\beta = .002$ Wald $\chi^2 = 13.32$, $p < .001$, $d = .08$
Employed Full Time	50%[48,52]	45%[42,47]	$\beta = .003$ Wald $\chi^2 = 12.26$, $p < .001$, $d = .08$
Conscientiousness	5.33[5.29,5.37]	5.12[5.06,5.19]	$\beta = .002$ Wald $\chi^2 = 20.66$ $p < .001$, $d = .09$
Agreeableness	5.19[5.15,5.23]	5.00[4.94,5.06]	$\beta = .002$ Wald $\chi^2 = 18.75$ $p < .001$, $d = .09$
Emotional Stability	4.82[4.77,4.87]	4.51[4.44,4.58]	$\beta = .003$ Wald $\chi^2 = 37.44$ $p < .001$, $d = .12$
Worker Experience	5.36[5.13,5.59]	1.88[1.72,2.05]	$\beta = -.009$ Wald $\chi^2 = 405.94$, $p < .0001$, $d = .41$

Questionnaire

Respondents first received the following questions to check eligibility:

What is your current age in years?

Are you a resident of the United States?

Then, upon signing the consent form, they received the following:

In what state do you currently reside?

What is the highest level of education you have completed?
[8 options, ranging from "Less than high school" to "Professional degree (JD, MD)"]

Which statement best describes your current employment status?
[Working full-time, Working part-time, Homemaker, Temporarily unemployed, Retired, Student - undergraduate, Student - graduate/professional, Permanent disability, Other (please specify)]

Is your current employment status by choice, or would you prefer to be employed full-time?

What is your current occupation? (please specify)

Are you married, partnered, divorced, separated, widowed, or have you never been married?

What best describes your current relationship?
[Single, Casually dating (you are not committed solely to one person), Monogamous (you and your partner have agreed to be sexually/romantically exclusive with one another), Consensually non-monogamous (e.g. polyamory, swinging, open relationship) (please specify), Another term best describes my relationship (please specify)]

How many people (including yourself) are currently living or staying in your primary residence?

You told us your age at the beginning of the survey. What is the age of the second person living or staying in your primary residence?

What is the age of the [third/fourth/fifth/sixth/seventh/eighth/ninth/tenth] person living or staying with you?

What is your present religion, if any?

[Christian, Jewish, Muslim, Buddhist, Atheist, Nothing in particular, Something else (please specify)]
[If not Atheist:] What is your denomination? [open-ended]

Lots of things come up that keep people from attending religious services these days even if they want to. Thinking about your own life these days, how often do you attend religious services, apart from occasional weddings, baptisms, or funerals?
[Every week, Almost every week, Once or twice a month, A few times a year, Never]

What race do you consider yourself to be? Please check all that apply:
[Black or African-American, Asian American, Native Hawaiian or Pacific Islander, American Indian or Alaskan Native, White, Other (please specify)]

Are you Spanish, Hispanic, or Latino?
[No, Yes]

[If yes:] What is your Spanish, Hispanic, or Latino background? This question is about ethnicity, not nationality or citizenship.
[Mexican, Puerto Rican, Chicano, Cuban, Other (please specify)]

Generally speaking, do you usually think of yourself as a Republican, a Democrat, an Independent, or something else?

[If Republican:] Would you call yourself a strong Republican, or a not very strong Republican?
[If Democrat:] Would you call yourself a strong Democrat, or a not very strong Democrat?
[If Independent or Other:] Do you think of yourself as closer to the Republican Party or to the Democratic Party? [Closer to the Republican Party, Neither, Closer to the Democratic Party]

Information about income is important to understand how people are doing financially these days. Your answers are confidential. What is your best guess of the total income of all the members of your family living with you in 2014, before taxes? This figure should include income from all sources, including salaries, wages, pensions, Social Security, dividends, interest, and all other income.
[There were 25 bins, ranging from “None or less than $2,999” to “$150,000 and over”]

When it comes to politics, how would you describe yourself?
[Extremely liberal, Liberal, Slightly liberal, Moderate; middle of the road, Slightly conservative, Conservative, Extremely conservative]

Which of the following words do you use most often to describe your sexual orientation?

[Gay or lesbian, Bisexual, Straight or heterosexual, Asexual, Different identity (please specify)]

How do you describe your current gender or gender identity? (Check all that apply.)
[Female, Male, Transgender female/trans woman, Transgender male/trans man, Genderqueer/gender non-conforming, Different identity (please specify)]

What sex were you assigned at birth, on your original birth certificate?
[Female, male]

Here are a number of personality traits that may or may not apply to you. Please indicate the extent to which you agree or disagree with each statement. You should rate the extent to which each pair of traits applies to you, even if one characteristic applies more strongly than the other. I see myself as...

Extraverted, enthusiastic
Critical, quarrelsome
Dependable, self-disciplined
Anxious, easily-upset
Open to new experiences, complex
Reserved, quiet
Sympathetic, warm
Disorganized, careless
Calm, emotionally stable
Conventional, uncreative

[Disagree strongly, Disagree moderately, Disagree a little, Neither agree nor disagree, Agree a little, Agree moderately, Agree strongly]