Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys

Jenny Tang, *Wellesley College;* Eleanor Birrell, *Pomona College;*
Ada Lerner, *Northeastern University*

https://www.usenix.org/conference/soups2022/presentation/tang

This paper is included in the Proceedings of the Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022).

August 8–9, 2022 • Boston, MA, USA

978-1-939133-30-4

Open access to the Proceedings of the Eighteenth Symposium on Usable Privacy and Security is sponsored by USENIX.

Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys

Jenny Tang
Wellesley College

Eleanor Birrell
Pomona College

Ada Lerner*
Northeastern University

Abstract

Privacy and security researchers often rely on data collected through online crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) and Prolific. Prior work—which used data collected in the United States between 2013 and 2017—found that MTurk responses regarding security and privacy were generally representative for people under 50 or with some college education. However, the landscape of online crowdsourcing has changed significantly over the last five years, with the rise of Prolific as a major platform and the increasing presence of bots. This work attempts to replicate the prior results about the external validity of online privacy and security surveys. We conduct an online survey on MTurk ($n = 800$), a gender-balanced survey on Prolific ($n = 800$), and a representative survey on Prolific ($n = 800$) and compare the responses to a probabilistic survey conducted by the Pew Research Center ($n = 4272$). We find that MTurk response quality has degraded over the last five years, and our results do not replicate the earlier finding about the generalizability of MTurk responses. By contrast, we find that data collected through Prolific is generally representative for questions about user perceptions and experiences, but not for questions about security and privacy knowledge. We also evaluate the impact of Prolific settings, attention check questions, and statistical methods on the external validity of online surveys, and we develop recommendations about best practices for conducting online privacy and security surveys.

*Work was done while Lerner was at Wellesley College.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

1 Introduction

Over the last fifteen years, online surveys conducted through crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) [2] and Prolific [47] have become increasingly critical tools for conducting quantitative usable privacy and security research. Researchers often use these platforms to recruit participants for user studies. However, the external validity of these user studies depends on the extent to which the results of these online studies generalize to the overall population.

Prior work has investigated the validity of online surveys in various domains—such as social sciences [10, 11, 61], health behavior [53], and privacy [32, 52]—with somewhat mixed results. However, work by Redmiles et al.—based on surveys conducted between 2013 and 2017 [49]—made strong, positive claims about the external validity of privacy and security surveys conducted on MTurk. It found that (1) MTurk responses regarding privacy and security experiences, advice sources, and knowledge were more representative of the U.S. population compared to responses from a census-representative web panel and (2) MTurk responses regarding privacy and security experiences, advice sources, and knowledge were generally representative of the U.S. population for respondents who are younger than 50 or who have some college education.

However, the landscape of crowdsourcing platforms has changed significantly in the last five years. One key change is the rise of Prolific as a major crowdsourcing platform. Founded in 2014 specifically as a platform for conducting online user studies, Prolific was only rarely used to recruit participants in 2017. By contrast, we find that by 2021, Prolific was approximately twice as common as MTurk as a recruitment platform for usable privacy and security studies. A second key change is the increasing presence of sophisticated bots on MTurk, which can degrade data quality. While bots do not appear to have been a significant problem on MTurk in 2017, more recent work has estimated that 20-50% of MTurk accounts are actually bots, with significant bot levels dating back to approximately March 2018 [5, 39].

In light of those changes, this work attempts to replicate the key findings of Redmiles et al. [49]. We ask:

(1) Are MTurk responses to privacy and security survey questions still representative of the U.S. population for respondents under 50 or with some college education?

(2) To what extent do various classes of attention check question—reading-based attention checks, open text-response questions, and CAPTCHAs—and/or raking (i.e. demographic weighting) improve the generalizability of MTurk responses?

(3) How well do Prolific responses to privacy and security questions generalize to the general U.S. population?

(4) What are the current best practices for conducting and analyzing online user surveys in the domain of privacy and security?

Additionally, we investigate the limitations of online survey methods for surveying underrepresented demographic groups, reporting on ways that specific groups differ from the general population and how specific populations might be misrepresented by a focus on a representative sample.

To answer these research questions, we conduct an online survey on MTurk ($n = 800$), a gender-balanced survey on Prolific ($n = 800$), and a representative survey on Prolific ($n = 800$) and compare the responses to a probabilistic survey conducted through the Pew Research Center ($n = 4,272$). We find that MTurk response quality has degraded over the last five years, and our results do not replicate the finding that MTurk responses are representative of certain subsets of the U.S. population, even when we exclude the 39% of MTurk responses that fail attention checks and apply raking. We find that data collected through both representative and gender-balanced Prolific samples is generally representative for questions about user experiences, perceptions, and beliefs; however, responses to questions about knowledge of privacy and security concepts and about social media use differ heavily from the overall U.S. population. We also find that racial, age, and education subgroups from our Prolific representative sample are generally moderately representative of their respective subgroups in the American population.

Based on our results, we recommend that privacy and security researchers prefer Prolific to MTurk when recruiting participants for online user studies. Our results show that Prolific provides good quality, generalizable data for certain types of user studies about privacy and security (those that focus on experiences, perceptions, and beliefs), but that Prolific users are generally more technical than the overall population, resulting in different responses about knowledge and behavior. We do not recommend using attention check questions or CAPTCHAs on Prolific, as they lengthen surveys unnecessarily without improving external validity.

2 Related Works

Given the widespread use of crowdsourcing platforms as recruiting tools for user studies, the question of the data quality and external validity of online survey data has been extensively studied from a variety of different angles.

2.1 Generalizability of Online Platforms

Prior work has investigated the generalizability of online user studies conducted through MTurk and Prolific in a variety of different domains.

Amazon Mechanical Turk. Amazon Mechanical Turk has long been a platform favored by researchers across disciplines such as computer science and the social sciences to conduct user studies [11, 41, 42, 61], and thus the external validity of MTurk data has been investigated in various different research contexts [8–10, 25, 53] with varying results. Conclusions about the external validity of MTurk surveys about privacy and security have also been mixed: multiple studies [32, 52] have found significant differences between an MTurk study and a U.S.-representative survey, with MTurk users reporting more concerns about privacy and information use and higher levels of social media use, while Redmiles et al. [49] found that that for participants under 50 years of age or with at least some college education, responses to questions regarding privacy and security were similar to the general population within these demographics, and that MTurk appeared to be more representative overall than a census-representative web panel.

However, there have been noted concerns about demographic differences between the MTurk population and the over U.S. population. In particular, the MTurk population has been found to be younger and with higher levels of education than the overall U.S. population [32, 43, 49, 50]. Concerns have also arisen over the population on MTurk, particularly as highly active MTurk workers tend to complete many of the available tasks before others are able to, making the effective sample population on MTurk only 7000 [44, 56]. Furthermore, while a study conducted in 2014 found that MTurk workers with over an 95% approval rating provide high quality data and do not require attention checks [46], more recent research has shown that data quality on MTurk has decreased dramatically to be less reliable than that on Prolific, even when quality filters (at least 95% approval rating and 100 submitted tasks) were used [45].

Prolific. Prolific was launched in 2014, and was primarily designed for use by researchers [47]. In the past few years, we have seen an increase in the use of Prolific as an alternative to conducting surveys on Amazon Mechanical Turk. Both the number of users and the number of researchers on the platform have increased dramatically in recent years [41, 45], and studies have shown that it is a viable alternative to MTurk [44].

Prolific mandates a minimum hourly payment for studies, and compensation may be adjusted by researchers if the survey takes longer than originally intended. Furthermore, users on Prolific have an option to return their submissions, indicating they no longer wish to participate or that they do not wish their data to be used, making it easy for participants to withdraw consent at any time.

Although a study in 2017 found Prolific to be slower in gathering responses than MTurk and CrowdFlower (another online survey site) [44], we did not find such differences in our sample, perhaps due to the expansion of the Prolific worker pool over the last five years.

2.2 Data Quality and Attention Check Questions (ACQs)

Some prior work has found that MTurk workers performed well on attention check questions [30], but other work found that MTurk workers were less attentive than convenience-sampled college students [25]. Prior work comparing differing survey platforms in 2017 have found that almost half of MTurk and Prolific participants failed at least one attention check question, with MTurk users failing on average fewer attention checks than Prolific [44]. More recent work in 2021 saw Prolific users outperforming MTurk users on completing ACQs [45]. Excluding those based on passing attention checks had little effect for MTurk, and a small effect on Prolific [44].

Prolific specifically allows for Instructional Manipulation Checks (IMCs), which are questions that "explicitly instruct a participant to complete a task in a certain way" such as clicking a specific answer [47]. IMCs and other attention checks have been shown to increase the reliability of data, and have become relatively widely used [16, 27, 29, 40, 45]. However, IMCs might also influence participants to change interpretation and assessment of subsequent questions [29].

Some research has also investigated comprehension, which involves checking that participants are able to understand instructions and explain them back to the researchers. This can be conducted in formats such as IMCs, or through textboxes asking users to summarize the instructions. However, these might not function exactly the same as attention checks, as prior work suggests those who fail attention checks may not be the same as those who do not comprehend instructions [10]. Prior work has also found that Instructional Manipulation Checks making sure participants understood instructions improved data quality [20]. Prolific users also tend to outperform MTurk users on comprehension checks, and there appears to be a positive correlation between correctly passing ACQs and comprehension questions [45].

CAPTCHAs are commonly discussed as a mechanism for improving data quality by eliminating bots from a dataset, however prior work has found that bot accounts are able to reliably pass CAPTCHAs [39].

3 Methodology

To evaluate the generalizability of online privacy and security surveys, we compared survey responses from four sources: (1) responses to a U.S. nationally-representative probabilistic sample, (2) people recruited through Prolific using their representative sample option, (3) people recruited through Prolific using their gender-balanced option, and (4) people recruited through Amazon Mechanical Turk (MTurk).

3.1 Survey Questions

To decide what questions to ask on our survey, we started by identifying categories of topics in privacy and security that have been the subject of recent user studies. We identified 28 papers published in the Proceedings on Privacy Enhancing Technologies Symposium (PoPETs) and the Symposium on Usable Privacy and Security (SOUPS) in 2021 that included user surveys. Two papers [21, 23] exclusively surveyed specific technical populations (freelance developers and developers who have used Rust, respectively) about technical topics (security practices when developing code and experience with Rust), so we excluded them from our analysis. For the 26 papers that surveyed the public, we qualitatively coded the categories of questions asked in user surveys; we also determined what platform they used to recruit participants and how they handled attention check questions.

Our qualitative coding identified five classes of questions that characterize the space of recent usable privacy and security surveys:

1. **Behavioral.** Questions about what users do, would do, or have done in relation to technology, social media, and privacy and security tools. These questions refer to active behaviors undertaken by the user. For example, whether they use Twitter or whether they have recently decided not to use a service because of concerns about its data collection practices. 21 papers (80.8%) included behavioral questions in their user survey [1, 6, 7, 13, 17, 19, 22, 27, 28, 31, 35, 36, 48, 54, 57, 59, 60, 62, 64–66].

2. **Experience.** Questions about whether or how often participants had experienced a particular type of event. These questions refer to actions or circumstances that occur to the respondent without active action on the part of that person. For example, how often they had experienced someone taking over their social media or email account without their permission or how often they were asked to agree to a privacy policy. 17 papers (65.4%) included experience questions in their user survey [1, 6, 7, 19, 22, 27, 28, 31, 33–35, 48, 54, 57, 59, 64, 66].

3. **Knowledge.** Factual questions relating to privacy and security topics that test how much participants know about the topic. These questions have factually correct

answers. For example, what it means if a website uses cookies or what a privacy policy is. 11 papers (42.3%) included knowledge questions in their user survey [1, 6, 7, 12, 27, 31, 33, 34, 57, 58, 60].

4. **Perceptions.** Opinion questions about user perceptions of and attitudes towards practices and behaviors. These questions—which focus on what respondents believe a principal would do or the reasons why they believe the respondent would do something—include questions about trust, comfort, and mental models. For example, how confident they were that a company would follow what the privacy policy says it will do or how comfortable they are with companies using their data to help develop new products. 19 papers (73.1%) included perception questions in their user survey [1, 7, 13, 17, 22, 26–28, 31, 33–35, 54, 57–60, 64, 66].

5. **Beliefs.** Opinion questions about what security options or privacy rights people ought to have. Beliefs questions focus on what the respondent thinks should be true rather than asking about perceptions of the current world. For example, whether people should have the right to remove potentially embarrassing photos or criminal history from publicly-available search records. 9 papers (34.6%) included belief questions in their user survey [1, 7, 19, 26, 31, 34, 35, 54, 66].

For each of the five categories of questions, we selected 4-8 questions from a database of questions used in a past Pew Research Center survey [15] (a total of 30 questions). Drawing our questions from this source had two key advantages: (1) Pew questions are extensively validated before being deployed and (2) responses from a large-scale ($n = 4,272$), nationally-representative survey conducted by Pew in June 2019 are publically available [15], precluding the need to deploy our own nationally-representative panel survey. To enable intercomparison, our online surveys closely followed Pew's methods: the phrasing of the questions were the same, the set of possible responses were the same, the order of the questions were the same—with randomization of question order or answer choices matching the Pew questionnaire—and there were no forced responses.

Since the Pew dataset includes demographic information for each participant, we also included basic demographic questions at the end of our survey. To facilitate comparisons with the Pew survey, we used demographic questions that matched the demographics released in the Pew dataset.[1]

Finally, we identified three common techniques for excluding bots from online survey populations: reading-based attention check questions (i.e., questions that require participants to select a particular answer, also known as IMCs) [27, 40], free-response text questions (survey responses are rejected if the answer is nonsensical, irrelevant, or clearly copy-pasted from the Internet), and CAPTCHAs. To allow us to evaluate the effect of these techniques on external validity, we added two additional questions to our survey: one reading-based attention check question and one free-response text question. We also required half of our participants (randomly selected) to successfully complete a CAPTCHA in order to submit the survey.

The full text of the survey can be found in Appendix A.

3.2 Datasets

We use four datasets: (1) a probabilistic dataset from the Pew Research Center panel [15], (2) a representative sample from Prolific (accurate to the US Census on age, sex, and race), (3) a gender-balanced sample from Prolific, and (4) a sample from Amazon Mechanical Turk (MTurk). We compensated online study participants $1.50 for completing the survey, which we estimated to take 6 minutes. This was approved by the Institutional Review Boards of the authors' institutions.

We additionally collected two filtered samples of underrepresented populations that do not appear in the Pew demographic categories: Indigenous people and transgender people. We deployed surveys on Prolific using the platform's prescreening filters to only allow participants in these demographics to take the study. Over a period of 8 days, we received responses from 79 Indigenous users on Prolific, and 197 transgender users.

1. **Pew American Trends Panel Wave 49.** This dataset ($n = 4,272$) was collected by Ipsos Public Affairs between June 3-17, 2019 on behalf of the Pew Research Center [15]. The weighted estimates for this sample are believed to accurate to ± 1.87 percentage points of the US population aged 18 and over. Pew Research Center typically makes survey data publicly available on their website two years after the data collection, so this dataset became publicly available in 2021.

   Participants in this survey were a subset of Pew Research Center's American Trends Panel (ATP) [14], a panel of more than 10,000 U.S. adults recruited and maintained by the Pew Research Center using state-of-the-art techniques.[2] This subset of the panel was chosen to be generally representative of the broader U.S. population; as this was a probabilistic survey, the resulting data was weighted to balance demographics to match

[1] Note that these questions do not reflect current best-practices for asking about gender [55] or race [63]. Nonetheless, we believed that matching the Pew phrasing was critical in order to enable direct comparisons with responses to the Pew survey.

[2] Prior to 2018, panel participants were recruited at the end of a large, national, landline and cellphone random-digit-dial survey that was conducted in both English and Spanish. After 2018, ATP has relied on address-based recruitment to avoid the response-bias that has developed in telephone-based recruiting. It supports non-Internet connected participants by providing tablets that enable those people to take surveys.

the U.S. population (to compensate for any biases due to sampling and non-response).

Our analyses treat this dataset as the gold standard for U.S. responses to our survey questions.

2. **Prolific Representative Sample.** We sampled U.S. participants ($n = 800$) using Prolific's representative sample feature. This sample is stratified on age, sex, and ethnicity based on the simplified U.S. census [47]. The median time to complete the survey was 6.1 minutes.

3. **Prolific Gender-Balanced Sample.** We sampled U.S. participants ($n = 800$) on Prolific, balanced on gender (50% male and 50% female). Prolific has been noted to have demographics that skew younger and more female [18]: currently within the U.S. sample space, there are over twice as many women on the platform as men. This survey took participants a median time of 5.3 minutes to complete. No participants are in both the Prolific representative and gender-balanced samples.

4. **MTurk Sample.** We collected a sample ($n = 800$) from Amazon Mechanical Turk, with participation restricted to people located in the U.S. who have completed over 50 HITs and have over 95% approval rate. We chose these filters as they are common practice for studies of this type deployed on the MTurk platform and believed to produce higher quality data [46, 49]. Participants took a median time of 5.2 minutes to complete the survey.

3.3 Analysis

We used chi-square proportion tests (χ^2) to compare response distributions. For each question, we ran χ^2 tests to compare the distribution of answers for each sample (Prolific representative, Prolific gender-Balanced, MTurk) pairwise against the Pew data. We also used Total Variation Distance (TVD) to quantify the distance between answer distributions in our surveys and the Pew data.

Total Variation Distance. Total Variation Distance (TVD), defined as $\text{TVD}(P,Q) = 1/2 \cdot \Sigma_i(P_i, Q_i)$, is a standard metric for quantifying the distance between two distributions [24]. Intuitively, it corresponds to the fraction of respondents who answer differently between the two samples. A TVD of 0 indicates that two distributions are identical; as the distributions become increasingly disjoint the TVD approaches 1.

To illustrate the concept of TVD, we show how to calculate the TVD between the distribution of responses to the first knowledge question for the Pew sample ($know1_{Pew}$) and responses for the Prolific gender-balanced sample ($know1_{Bal}$). In the Pew survey, .626 of the respondents answered correctly, .093 incorrectly, and .282 with "Not sure"; in our representative Prolific sample, the proportions for correct, incorrect, and not sure were .866, .028, and .106 respectively. Therefore,

$$\begin{aligned} & TVD(know1_{Pew}, know1_{Bal}) \\ = \; & \frac{1}{2}(|.626-.866|+|.093-0.028|+|.282-.106|) \\ = \; & .2405 \end{aligned}$$

This TVD of .2405 shows that approximately one quarter of the responses were distributed differently between the Pew sample and the Prolific representative sample.

To show how we use TVD, consider a comparison between the survey questions `know1` and `exp5`. In both cases, a χ^2 test indicates a significant difference in answers between the Pew sample and the Gender-Balanced Prolific sample. To contextualize this result, we calculate TVD values for both pairs of distributions. Using the same definition as above, we find that $TVD(exp5_{Pew}, exp5_{Bal}) = .111$. The lower TVD for `exp5` provides evidence that the balanced Prolific sample may be closer to the Pew sample for the experience question (TVD = .111) than for the knowledge question (TVD = .2405).

We chose to use these two measures (χ^2 tests and TVD) as both have strengths and weaknesses in their ability to provide insights into the representativeness of these platforms' participants. χ^2 tests with p-values provide a thresholded measure of sameness or difference, while TVD provides us with a limited but valuable continuous measure of distance. For example, when χ^2 tests show that answer distributions are statistically distinct for two question categories, TVD augments this analysis by providing a method of estimating whether the non-representativeness of one question category may be of larger magnitude than the other.

As prior work has found that online surveys were representative for populations under 50 years old or with at least a college level education [49], we further explored whether online survey platforms were more representative of certain demographic subsets in the general US population. In particular, we separately analyzed populations aged 18-29, 30-49, and 50+ from each of our samples against the corresponding demographic groups in the Pew dataset. We also conducted an analysis divided by education level, classified into one of three categories: high school graduate or less, some college, college graduate+.

Since the responses ranged from binary to multiple choice to Likert-scale, we did not attempt to code answers into binary variables. For most questions, we kept the response codings as presented to the user. For knowledge questions—which had one correct answer, 3-5 incorrect answers, and a "Not sure" option—we coded answers into three categories: Correct, Incorrect, Not sure. In the studies, participants were able to skip any question. As no more than 2.5% of any question had blank answers, we chose to impute the answers for non-response. Any skipped attention check was coded as a failed attention check. For knowledge questions, any skipped question was classified as "Not sure". For all other questions, non-response

was classified into the most negative category (e.g. "Not at all confident", "No, do not use this"). To validate this choice of imputation, we ran an analysis as well on imputation where non-response was classified as the most positive (e.g. "Very confident", "Yes, use this") and found that TVD remained ± 0.01 both within each category and overall.

To understand whether attention checks are effective in improving data quality, we ran an analysis wherein only responses of those who passed each of the three attention check questions were included. To determine efficacy of reading attention checks, we compared only those who passed the reading attention check (selected "Strongly agree") against the Pew data. For the textbox attention check which asked users to define "digital privacy" in their own words, a researcher from our team coded all responses into accept, reject, and copy-paste. "Reject" responses referred to answers that either were nonsensical, unrelated to the question, or merely repeated the words "digital privacy". "Copy-paste" indicated responses that were plausible definitions of "digital privacy", but appeared verbatim 5 times or more throughout the sample or contained large chunks of text that were copied verbatim from these phrases. Under this coding, some other phrases were indeed often repeated, but were accepted if they appeared less than 5 times. A second researcher resolved cases where it was uncertain whether a response should be rejected. We removed all responses either coded as "Reject" or "Copy-paste" when analyzing samples that are said to have passed the textbox attention check. For the CAPTCHA analysis, for each of our samples, we ran analyses comparing those in the sample who saw and passed a CAPTCHA (around 50% for each sample) against the Pew dataset.

We conducted demographic raking using the R `anesrake` package, weighted by age, sex, education, and race to see if it would improve the generalizability of our samples. We used proportions from the 2017 American Community Survey [3] to match the demographic weighting used for the Pew dataset.

We were further interested in whether underrepresented groups had significantly different responses than the general population. We compared demographic groups from our Prolific representative sample to the same group from the Pew sample to investigate whether demographic groups on Prolific are representative of their respective group in the broader population. With our filtered samples of rare demographic subpopulations (Indigenous and transgender people), we compared our filtered sample to the Prolific representative sample.

3.4 Methodological Limitations

Due to the statistical constraints surrounding sample size and power, smaller sample sizes necessarily have less statistical power. Thus, for smaller samples (e.g., CAPTCHAs, underrepresented groups), we expect to find fewer instances of *statistical* significance (p-values< 0.05), implying that these samples more closely match the Pew dataset. However, this does not mean differences do not exist, but rather that they might be too slight to detect at lower sample sizes. Thus, we examine TVDs in conjunction with p-values in order to obtain a clearer picture rather than simply defaulting to p-values.

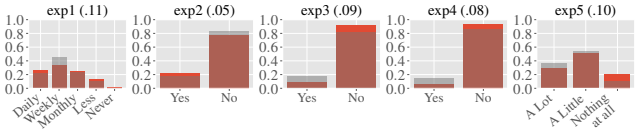
Dem	Response	Pew (%)		Online Samples (%)		
		Raw	Wgt	Repr	Bal	MT
Age	18-29	16	20	23	45	22
	30-49	31	33	34	43	66
	50-64	31	26	30	9	10
	65+	23	20	12	2	2
Edu	HS or less	35	38	13	13	8
	Some College	28	31	34	34	26
	College grad+	38	30	53	52	66
Race	White	78	74	75	75	82
	Black	11	12	13	4	13
	Asian	3	4	7	12	3
	Mixed	4	5	4	7	2
	Other	4	5	2	3	1
Sex	Male	44	48	49	50	68
	Female	56	52	51	50	32

Table 1: Demographic characteristics of the four datasets. Since the Pew dataset was a probabilistic sample, the weighted dataset (Wgt) was analyzed. For the online samples—Prolific representative (Repr), Prolific gender-balanced (Bal) and MTurk (MT)—raw data was analyzed except where we explicitly state that raking was applied.

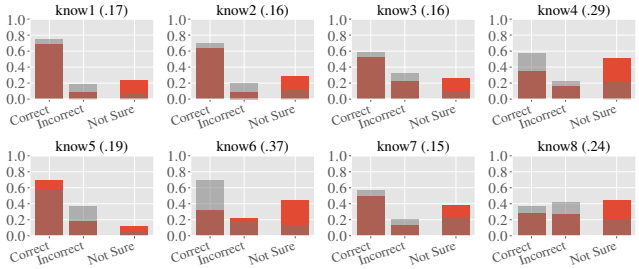
TVD is one of many measures that could be used to summarize our data and quantify distances between distributions. A primary limitation of TVD is that it does not account for whether the underlying data is categorical or ordinal, and thus on Likert-scale style questions, treats participant answers which differ by a small "amount" (e.g., from 1 to 2) identically from those that differ by a large "amount" (e.g., from 1 to 5). Similarly, like χ^2 tests, it cannot distinguish between different specific ways that categorical answer distributions differ. For example, TVDs for knowledge questions are large for both MTurk and Prolific representative $(.30, .23)$, and χ^2 tests find that responses to all 8 knowledge questions are significantly different from Pew responses for both, yet the *direction* of these differences is opposite: MTurkers are more likely to be *incorrect* in their knowledge while Prolific respondents are more likely to be *correct*.

Other options for contextualizing results from surveys are possible, such as visualizations and tables of raw answer proportions, which are provided in our figures and in Appendix B. Qualitative work examining the reasons for and nuances of the differences we observe could provide another avenue of understanding how and why results between probabilistic surveys and online platforms differ and what implications these differences have for the validity of insights provided by usable privacy and security studies.

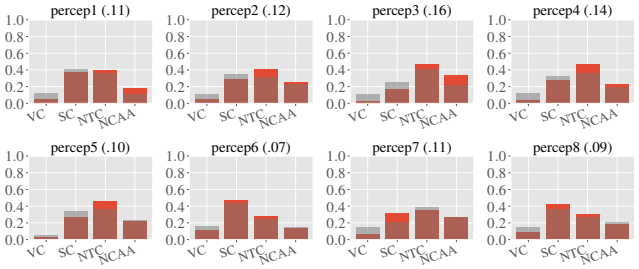
(a) Behavior Questions. Mean TVD = .30



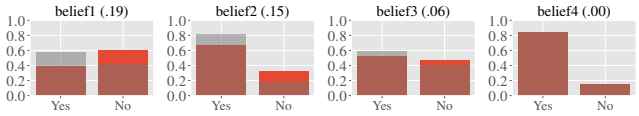
(b) Experiences Questions. Mean TVD = .09



(c) Knowledge Questions. Mean TVD = .22



(d) Perception Questions. Mean TVD = .11



(e) Beliefs Questions. Mean TVD = .10

(f) Legend. Brown is overlap of Pew and MTurk.

Figure 1: Distributions of responses to all questions for the Pew sample (weighted) and the MTurk sample (u50SC/Rak/freeAC). TVDs between the Pew sample and the MTurk sample are given in the captions.

We do not use corrections (e.g., Benjamini-Hochberg, Bonferroni) to analyze our results. These corrections control for Type I errors (false discovery rate) by limiting the number of erroneously statistically significant results (p-values < 0.05). However, we are attempting to find results that *do not* significantly differ between samples (i.e., $p > 0.05$), so these corrections could in fact overstate our results.

4 Results

We deployed three copies of our survey—Prolific representative sample, Prolific gender-balanced sample, and MTurk sample—in February 2022. We found that the Prolific representative survey took significantly longer to run; it took 49 hours to complete, compared to 2.5 hours for the Prolific gender-balanced survey and 2 hours for the MTurk survey.[3] The Prolific representative survey also cost significantly more to deploy: collecting that sample cost $2,784, compared to $1,600.00 for the Prolific gender-balanced sample and $1,682 for the MTurk sample. We then analyzed the responses to evaluate the external validity and data quality of the resulting samples. A summary of the demographics for each of the samples is provided in Table 1, and complete results are summarized in Appendix B.

[3]However, we note that since a significant percentage of our MTurk responses (39.1%) failed the free-response attention check, rejecting those responses and re-releasing those HITs would significantly increase the total deployment time; since less than 1% of responses failed that check for either Prolific sample, no extra time or effort would be required for those surveys.

4.1 The External Validity of MTurk

Our MTurk sample was heavily weighted toward younger participants (703/800 participants were under 50) and those with higher education (737/800 participants had at least some college education); this finding replicates prior work [32, 49]. However, while Redmiles et. al. [49] found that for the well-represented demographics (people under 50 or with at least some college) MTurker responses to privacy and security questions (about behavior, experiences, and knowledge) had high external validity, we were unable to replicate that result.

When we analyzed the raw MTurk sample, we found that responses collected through MTurk were extremely different from Pew (Table 2). We found statistically significant differences in responses for 29 of the 30 questions, and the overall average Total Variation Distance (TVD) was .29 (intuitively indicating that more than a quarter of the sample answered differently). We attempted to replicate prior work by also analyzing the sample that contained only people with under 50 or some college education and applied raking (i.e., demographic weighting). However, we still found statistically significant differences in responses for 29 questions, and the average TVD dropped only slightly (to .28).

Unlike the earlier work, we found that both raking and filtering out responses that failed a free-response text attention check question had significant effects on data quality. Combining these data quality measures with the demographic restrictions from prior work—i.e., restricting to people under 50 or with some college who passed the text attention check

Category	Raw Sample		u50SC/Rak/freeAC	
	TVD	$p < 0.05$	TVD	$p < 0.05$
Behavior	0.41	5/5	0.30	5/5
Experiences	0.27	5/5	0.09	5/5
Knowledge	0.30	8/8	0.22	8/8
Perceptions	0.30	8/8	0.11	8/8
Beliefs	0.15	3/4	0.10	3/4
Overall	0.29	29/30	0.17	29/30

Table 2: Measures of external validity for the MTurk sample. TVD indicates distance from the (weighted) probabilistic Pew sample. $p < .05$ shows the fraction of questions in each category for which the responses were statistically significantly different from the Pew sample. For both, lower is better.

and applying filtering, denoted u50SC/Rak/freeAC in Figure 1 and Table 2—produced the best-case results for the MTurk sample.

In this best-case scenario, 29 of the questions still had significantly different responses, but the average TVD went down to .17. In particular, responses about experiences, perceptions, and beliefs are somewhat generalizable: TVDs dropped to around .10, although χ^2 tests still showed that responses for most questions were still significantly different from Pew. However, knowledge questions were still very different: MTurkers were much more confident—and incorrect—on knowledge questions even after data quality measures were applied. Behavior questions—which focused on social media use—also remained very different: MTurkers were more likely to use Facebook (63% vs. 53%), Instagram (74% vs. 44%), Twitter (79% vs. 26%), and other social networks (82% vs 34%). Complete results for this best-case scenario are depicted in Figure 1 and measures of external validity are summarized in Table 2.

4.2 The External Validity of Prolific

Like the MTurk sample, the Prolific gender-balanced sample was heavily weighted towards younger participants and those with higher education; the age skew was more extreme and the education skew less. That sample also included significantly fewer Black participants and more Asian and mixed race participants. The Prolific representative sample was representative of the overall U.S. population for age and race, but showed the same skew towards higher education.

Overall, we found that both Prolific samples generalize better than the MTurk sample and that free-response attention checks were no longer critical for data quality. However, the external validity of the samples varied significantly depending on the type of question. Our results are shown in Figure 2, and measures of external validity are summarized in Table 3.

Behavior. Although responses about behavior from the Prolific representative sample were slightly more generalizable in terms of TVD (and both were more generalizable than the MTurk responses), neither of our Prolific samples demonstrated high external validity for behavior questions. Prolific participants were similarly likely to use Facebook compared to Pew participants but differed on other reported behavior, with the fraction of Prolific participants reporting that they use Instagram, Twitter, and "Other Social Media Sites" being 25%–54% higher compared to the Pew sample.

Samp.	Cat.	Raw Sample		u50SC/Rak/freeAC	
		TVD	$p < 0.05$	TVD	$p < 0.05$
Rep.	Behav.	0.22	4/5	0.19	3/5
Rep.	Exp.	0.07	5/5	0.06	5/5
Rep.	Know.	0.23	8/8	0.17	8/8
Rep.	Percep.	0.05	3/8	0.06	4/8
Rep.	Beliefs	0.07	3/4	0.06	2/4
Rep.	Overall	0.13	23/30	0.11	22/30
Bal.	Behav.	0.27	3/5	0.24	4/5
Bal.	Exp.	0.06	4/5	0.05	4/5
Bal.	Know.	0.24	8/8	0.16	7/8
Bal.	Percep.	0.05	4/8	0.07	7/8
Bal.	Beliefs	0.08	2/4	0.06	3/4
Bal.	Overall	0.14	21/30	0.12	25/30

Table 3: Measures of external validity for the Prolific samples. For both, lower is better. See Table 2 for more details.

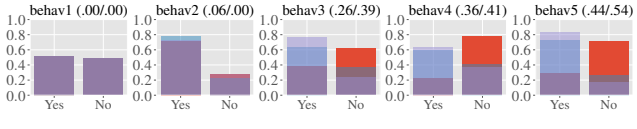
Experiences. Overall, both of our Prolific samples generalized well for questions about prior experiences. While most of those questions showed statistically significant differences, the magnitude of those difference was quite small (TVD = .07 for the representative sample and .06 for the gender-balanced sample).

Knowledge. Knowledge questions did not generalize well for either of our Prolific samples. Prolific respondents were more likely to provide *correct* answers and less likely to answer "Not sure", suggesting that Prolific users are significantly more knowledgeable about privacy and security matters than the overall U.S. population.

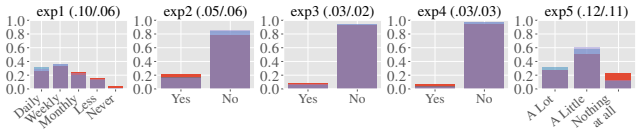
Perceptions. Both Prolific samples had relatively high external validity for questions about perceptions of privacy and security. The Prolific representative sample had statistically different responses compared to the Pew sample for only 3/8 questions (4/8 for the gender-balanced sample), and TVDs between each sample and the Pew sample were small (about .05).

Beliefs. Both Prolific samples also had high external validity for questions about beliefs about privacy and security. While some of the questions were statistically distinguishable, the TVDs were low suggesting that the effect size was small.

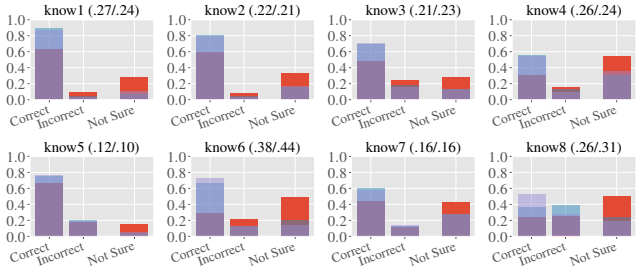
We also analyzed the Prolific samples using the best-case data quality measures from our MTurk analysis—restriction to people under 50 or with some college who passed the text attention check and applying filtering, denoted



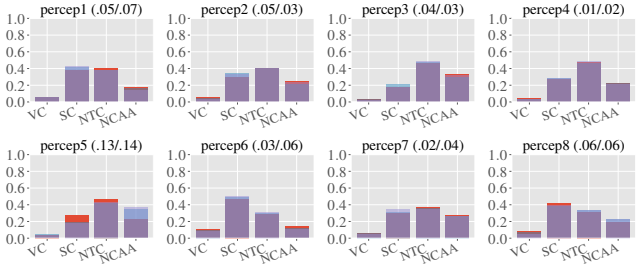
(a) Behavior Questions. Mean TVD (Repr, Bal) = .22, .27



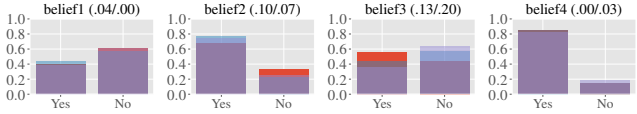
(b) Experiences Questions. Mean TVD (Repr, Bal) = .07, .06



(c) Knowledge Questions. Mean TVD (Repr, Bal) = .23, .24



(d) Perception Questions. Mean TVD (Repr, Bal) = .05, .05



(e) Beliefs Questions. Mean TVD (Repr, Bal) = .07, .08

(f) Legend. Dark purple is overlap of all 3.

Figure 2: Distributions of responses to all questions for the Pew sample (weighted), Prolific representative sample (raw), and Prolific gender-balanced sample (raw). TVDs between the Pew sample and the Prolific samples are given in the captions.

u50SC/Rak/freeAC in Table 3). Overall, TVDs decreased slightly. However, unlike for the MTurk sample, this analysis did not dramatically improve the generalizability of the Prolific samples.

4.3 Data Quality Measures

We evaluated four data quality measures: reading-based attention checks, free-response text attention checks, CAPTCHAs, and raking. For the MTurk sample, we found that a free-response attention check (which 39.1% of responses failed) and raking both significantly improved data quality for the MTurk sample. Despite the data quality issues with MTurk, neither reading-based attention checks nor CAPTCHAs (which no respondents failed) significantly improved data quality. Although Prolific respondents did slightly less well than MTurkers on our reading attention check question (7.75%-8.25% failed), none of our data quality measures significantly improved data quality for the Prolific samples.

4.4 Beyond the "Average" User

While the standard metric of external validity is the extent to which results generalize to the overall population, overall generalizability does not necessarily imply that results are valid across all subgroups. We therefore also examine the question of how well our results generalize for various demographics subpopulations. We apply two analysis techniques: (1) we compare demographics slices from our online samples to the corresponding demographic slices of the Pew sample, an approach that parallels the investigations of MTurk generalizability by Redmiles et al. [49] and (2) we compare demographic slices from rarer, traditionally understudied subpopulations to the overall population.

4.4.1 Prolific vs. Pew Demographic Subpopulations

Since Pew is our gold standard, we can perform this analysis only for demographic variables reported by Pew, and only for values of those variables which occur sufficiently frequently in the population to enable meaningful comparison. Based on these limitations, we choose to analyze two racial subpopulations (Black and Asian American), educational attainment, and age. Table 1 presents the numbers of people in each of these categories. Because the sample sizes are inherently smaller for these subpopulations than for the overall population, we focus our analysis exclusively on distance between the distribution of responses provided in the online surveys and the (weighted) distribution of responses to the Pew survey instead of considering p-values or the number of questions with statistically different responses.

Overall, we found that the Prolific representative sample tends to be the best of all three collected samples for each of these demographic brackets (although the Prolific gender-balanced sample is often nearly as good) and that the Prolific samples generalize better for younger and for more highly-educated subpopulations.

Race. Overall, the Prolific representative sample is almost as representative of the Black and Asian subpopulations as it is for the overall population. Average TVDs measuring the representativeness Black and Asian subpopulations are about .01–.02 higher on average than TVDs for the full dataset.

Age. For Prolific, we found that as ages increase, the samples become less generalizable. For people age 18-29, both Prolific samples are fairly representative of the general U.S. population, with particular improvement for knowledge questions (TVD = .15) and behavior questions (TVD = .16). Within both 18-29 and 30-40 age brackets, the Prolific samples actually generalize to the Pew dataset better than comparing the full datasets. By contrast, generalizability for people over 50 is worse, particularly for knowledge questions (TVD: Repr = .28, Bal = .32) and behavior questions as older Prolific users demonstrate significantly more knowledge of privacy and security and significantly higher levels of technology use. For our MTurk sample, both the 18-29 and the over 50 subpopulations were more generalizable than the full dataset, although they still had lower data quality than the corresponding slices of the Prolific samples.

Education. For our Prolific samples, we found that as education increases the samples become more generalizable. For respondents with a high school education or less, Prolific samples are less generalizable for this demographic slice than for the overall population, with participants reporting particularly higher levels of technology use. For respondents who are college graduates, both Prolific samples are reasonably representative of the overall population of U.S. college graduates (TVD: Repr = .10, Bal = .12), with more representative responses to knowledge questions (TVD: Repr = .13, Bal = .14). Conversely, the generalizability of the MTurk sample for people with high school education or less did not decrease (although the data quality remained worse than the Prolific samples) while data quality does decrease slightly for the subpopulation with Bachelor's degrees (TVD = .30).

4.4.2 Rare Demographic Subpopulations

Finally we identified two populations with relatively low representation on Prolific—Indigenous people and transgender people—and explored (1) how effective Prolific's filters are at producing large samples of rare (and frequently understudied) subpopulations and (2) to what extent generalizable results for the overall population apply to these subpopulations.

Indigenous People. Our filtered sample of Indigenous people ran for 8 days and obtained 79 responses during that time, an average of about 10 people per day. For context, at the time that we launched this filtered sample, Prolific reported that there were 294 Indigenous respondents who had been active in the past 90 days.

Comparing the distributions of responses of these 79 respondents to our full Prolific representative sample, we found that variations were relatively small, with Indigenous people on Prolific being generally somewhat similar to other people completing surveys on Prolific. Comparable to the difference between our Prolific representative sample and Pew on the most representative question categories, mean TVDs comparing Indigenous respondents and the general Prolific population were under .10 for all question categories.

We emphasize that given the small size of this sample, we are unable to make conclusive statements about trends among Indigenous people on Prolific. Generally speaking, our data supports the claim that Indigenous people are more similar than different to other Prolific users, with TVDs between .04 and .05 for 3 question categories (Experiences, Perceptions, Beliefs). In terms of behaviors, they are more likely to use all social networks, including especially Facebook (TVD = .16) and other social networks (TVD = .12). Indigenous people in our sample appear to be slightly more likely to answer "Not Sure" to knowledge questions. No other clear trends emerge in how Indigenous respondents on Prolific are different from other respondents on Prolific in terms of experiences, perceptions, or beliefs.

Transgender People. Our filtered sample of transgender people ran for 8 days and obtained 197 responses during that time, an average of about 25 people per day. At the time that we launched this filtered sample, Prolific reported that there were 1231 transgender respondents who had been active in the past 90 days.

Comparing the distributions of responses of these 197 respondents to our full Prolific representative sample, we find that variations are small to moderate, with transgender people on Prolific being generally somewhat similar to other people completing surveys on Prolific. Mean TVDs comparing transgender Prolific respondents to the overall Prolific representative sample were under .12 for all question categories.

Although the low sample size precludes definitive findings, our data for this subpopulation provides preliminary evidence of some potential interesting trends. In terms of behavior, transgender people were less likely to use Facebook, and more likely to use Instagram, Twitter, and other social networks, than the Prolific Representative population. Transgender people were also more knowledgeable than Prolific participants overall, answering 6/8 knowledge questions correctly more often. Notably, transgender people were particularly more likely to understand how private browsing works, with a very large TVD of .26 distinguishing them as much more likely to answer `know8` correctly and much less likely to answer incorrectly or with "Not Sure". This result might be due to the need for transgender people to use private browsing mode to protect themselves and their browsing habits from local adversaries, such as family, while seeking community, gathering information, and engaging in activism online [37].

Transgender people were also consistently more likely (TVD .07-.13) to select "Not Confident At All" in response to perception questions that asked about confidence that companies will follow their privacy policies, promptly notify about

data breaches, publicly admit mistakes that lead to privacy breaches, use personal information in appropriate ways, and be held accountable by the government for privacy missteps. Finally, they were slightly more likely to believe that people should have the right to have various personal information removed from public search results, with a particularly large (TVD = .16) increase compared to the Prolific representative sample in the likelihood that transgender people would say that people should be able to have "Negative media coverage" about themselves removed from public search results. We hypothesize that this might emerge from the likelihood of transgender people to experience media coverage about them as negative, for example if articles misgender them or include out-of-date personal details such as deadnames.

5 Discussion

While our results quantify the external validity of online surveys about privacy and security, they also provide insight into best practices for conducting online studies in this space.

Recommendation 1: *We recommend preferring Prolific to MTurk when recruiting participants for privacy and security surveys.*

Overall, we found major degradation of MTurk data quality and external validity since 2017 [49]. If MTurk samples are used, applying the data quality measures studied in this paper—including demographic raking and a stringent open textbox attention check—is critical to enhance external validity. However, even when applying these data quality measures to MTurk data, Prolific gender-balanced samples provide better validity and their use is recommended at the current time. It is important to remember, however, that both Prolific and MTurk samples better represent younger and more educated populations. Additionally, online samples appear to be differently representative for different types of questions, as we discuss below in the Recommendation 3.

Future work should examine the validity of samples from other platforms, which could be comparable to or better than Prolific. For example, we note that CloudResearch, which uses the MTurk population, has been found to provide similar data quality to Prolific when the default data quality filters are applied [38]. Although our literature review did not find any papers that used CloudResearch, it it provides an alternative platform for recruiting participants in the future. Future work should also continue to monitor the external validity of MTurk and Prolific, as population demographics and data quality may continue to change over time.

Recommendation 2: *Determinations about whether to use Prolific's representative sample feature can make trade-offs between generalizability and logistical constraints without significantly impacting data quality for most studies.*

We find that Prolific's representative sample feature produces data that most closely matches the results from the nationally representative sample from the Pew dataset. However, the representative sample takes much longer (49 hours vs. 2.5 hours for 800 responses) and is significantly more expensive ($2784 vs. $1600) to deploy than collecting a gender-balanced sample of the same size from Prolific. In most cases, a Prolific gender-balanced sample performs nearly as well as a representative sample, with less than .02 difference in average TVD across all question categories. The largest gains for representative over gender-balanced were for behavior questions, for which neither was representative. All other question categories had very small differences (TVD $<$ .01) between representative and gender-balanced samples.

Recommendation 3: *Be cautious when drawing conclusions from online studies about privacy and security knowledge or social media use, as these results might not be representative of the overall U.S. population.*

None of our online samples were representative of the overall population for knowledge questions—which posed factual questions about privacy and security topics—or behavior questions—which were dominated by questions about rates of social media use. We recommend that researchers take care in designing studies and interpreting data which depend on these properties of respondents. Similar to prior work, we do find that the younger and more highly educated the population, the more Prolific is representative, particularly for knowledge questions, which drop from TVDs of .28 for those over 50 to .15 for those 18-29, and from .27 for those with high school degrees or less education to .13 for those with college degrees. Even these TVDs are quite high, however, indicating that 15-13% of responses are different than they would be for that actual age or education range in an census-representative sample, and so we still urge caution in relying on such data.

Our results show that participants recruited through MTurker and Prolific are more confident about privacy and security knowledge compared to the overall U.S. population, with fewer respondents answering "Not sure" to most questions. We observe that this is a similar phenomenon to past results which found that MTurk workers are more certain about what information is available about them online [32]. This confidence gap also raises questions about the generalizability of non-survey studies that recruit participants through these online platforms, since prior work has found that confidence is a better predictor of security behaviors than actual knowledge [51].

One factor that might have contributed to the drastically higher reported use of social media in our online samples is response bias from participants who worry that they may be excluded from a survey (and thus not be paid) if they don't use certain products, leading them inaccurately claim that

they use social networks which they actually do not. Another possibility is that the population on these platforms have different behaviors regarding social media than the general U.S. population, leading to higher adoption and use of social media platforms. Indeed, prior work on MTurk has also found that U.S. MTurk workers have higher reported social media use than the general US population [32], which is supported by our findings in our Prolific and MTurk samples.

While it is possible that some of the difference in responses to behavior questions might have been due to actual differences in social media use between 2019 and 2022, a 2021 survey [4] conducted by Pew about a year into the pandemic shows that social media adoption has not drastically increased since mid-2019, when the American Trends Panel Wave 49 was conducted. That survey found that in 2021, 69% of Americans used Facebook, 40% used Instagram, and 23% used Twitter, numbers which are very closely compatible with the 71%, 38%, and 23% found in 2019. This suggests that the higher usage numbers we find in our online samples are genuine symptoms that users of online crowdsourcing platforms are not representative of the overall population in terms of their social media use.

Recommendation 4: *Attention Check Questions and CAPTCHAs are not recommended for online surveys conducted on Prolific.*

We do not recommend reading attention checks (Instructional Manipulation Checks), textbox attention checks, and CAPTCHAs when collecting survey responses on Prolific. Our reading attention check was failed by 66/800 (representative sample) and 62/800 (gender-balanced sample) participants, but data quality was not improved by analyzing the data with these responses removed (see Section 4.3). Prolific users almost never fail textbox attention checks (7/1600 failures) or CAPTCHAs (0/1600 failures). Based on our results, using such attention check questions lengthens surveys unnecessarily. Using IMCs might also change participants interpretation of subsequent questions [29].

Recommendation 5: *Raking is not currently necessary when analyzing the results of online privacy and security studies.*

Raking is often used in survey methods that intend to be representative of the general population since perfect response rates from demographic groups cannot be ensured by any sampling approach. Although we found raking had little effect on the representativeness of either of our Prolific samples (Section 4.3), studies in other disciplines have seen success in using raking for MTurk survey data to better generalize to the US population [53], and we would recommend researchers consider it. However, we note that raking also requires decisions on which demographic variables to weight on and might differ for different questions and fields of study.

Recommendation 6: *Special care should be taken to include a diverse population of study participants, particularly for demographics that are rare or underrepresented on crowdsourcing platforms.*

Prior work has noted that online platforms tend to be younger, more highly educated, and more white than the general U.S. population [49]. This demographic imbalance could then lead to a fallacy of the "average" user on such platforms not being at all representative of the general population. Groups that do not make up the majority might also have significantly different preferences than the "general population". For example, participants from racial minorities were more unsure about their security knowledge than the general population, and transgender people had lower confidence in companies taking responsible action regarding privacy issues than the general Prolific population. Therefore, we encourage researchers to consider specifically sampling underrepresented populations to understand possibly divergent privacy and security perceptions and backgrounds to avoid over-general interventions and claims that could contribute to further marginalizing already marginalized populations.

Study Limitations. As we limited our studies to participants located in the United States, we cannot make claims as to whether Prolific is similarly representative of other jurisdictions or of the overall global population. Indeed, prior work has also found differences between privacy attitudes between MTurk workers located in the U.S. and in India [32]. Additionally, while the Pew dataset was weighted on myriad strata of demographics to best represent the adult U.S. population, we still recognize that it is not perfectly representative of the general US population. As with all surveys, there might be non-response bias, even with the most carefully selected probabilistic sample. In other words, just as those who choose to take surveys on online platforms differ from the general population in terms of tech familiarity and use, so too might probabilistic studies vary from exact national preferences. Therefore, though we use Pew as our gold standard, we cannot guarantee that it is a perfect representation of the preferences of all Americans.

6 Conclusion

Online crowdsourced samples are an important source of data for usable privacy and security survey research today. Understanding the external validity of these samples is critical to ensuring that the results from such research generalize and can appropriately guide individuals, technologists, lawmakers, and regulators. Our work evaluates the external validity two popular crowdsourcing sites—MTurk and Prolific—and provides recommendations about best practices for conducting security and privacy surveys on these platforms.

Acknowledgments

We are grateful to Cassandra Pattanayak and Elissa Redmiles for their advice and support on this work. This work was supported by the National Science Foundation (CNS Award #1948344) and Wellesley College.

References

[1] Desiree Abrokwa, Shruti Das, Omer Akgul, and Michelle L. Mazurek. Comparing Security and Privacy Attitudes Among U.S. Users of Different Smartphone and Smart-Speaker Platforms. In *17th Symposium on Usable Privacy and Security*, pages 139–158, 2021.

[2] Amazon mechanical turk. https://www.mturk.com.

[3] American communities survey. https://www.census.gov/programs-surveys/acs/data.html, 2017.

[4] Brooke Auxier and Monica Anderson. Social media use in 2021. *Pew Research Center*, 1:1–4, 2021.

[5] Hui Bai. Evidence that a large amount of low quality responses on MTurk can be detected with repeated GPS coordinates. https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random, 2018.

[6] Daniel V. Bailey, Philipp Markert, and Adam J. Aviv. "I have no idea what they're trying to accomplish:" Enthusiastic and Casual Signal Users' Understanding of Signal PINs. In *17th Symposium on Usable Privacy and Security*, pages 417–436, 2021.

[7] David G. Balash, Dongkun Kim, Darika Shaibekova, Rahel A. Fainchtein, Micah Sherr, and Adam J. Aviv. Examining the Examiners: Students' Privacy and Security Perceptions of Online Proctoring Services. In *17th Symposium on Usable Privacy and Security*, pages 633–652, 2021.

[8] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. Comparing the similarity of responses received from studies in amazon's mechanical turk to studies conducted online and with direct recruitment. *PloS one*, 10(4):e0121595, 2015.

[9] Tara Behrend, David Sharek, Adam Meade, and Eric Wiebe. The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3):800–813, 2011.

[10] Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances. Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, 58(3):739–753, 2014.

[11] Michael D. Buhrmester, Sanaz Talaifar, and Samuel D. Gosling. An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13:149–154, 2018.

[12] Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. Automated Extraction and Presentation of Data Practices in Privacy Policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110, April 2021.

[13] Jason Ceci, Hassan Khan, Urs Hengartner, and Daniel Vogel. Concerned but Ineffective: User Perceptions, Methods, and Challenges when Sanitizing Old Devices for Disposal. In *17th Symposium on Usable Privacy and Security*, pages 455–474, 2021.

[14] Pew Research Center. American trends panel. https://www.pewresearch.org/our-methods/u-s-surveys/the-american-trends-panel/.

[15] Pew Research Center. American trends panel wave 49. https://www.pewresearch.org/internet/dataset/american-trends-panel-wave-49/, June 2019.

[16] Jesse Chandler, Cheskie Rosenzweig, Aaron J. Moss, Jonathan Robinson, and Leib Litman. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods*, 51(5):2022–2038, 2019.

[17] Varun Chandrasekaran, Chuhan Gao, Brian Tang, Kassem Fawaz, Somesh Jha, and Suman Banerjee. Face-Off: Adversarial Face Obfuscation. *Proceedings on Privacy Enhancing Technologies*, 2021(2):369–390, April 2021.

[18] Nick Charalambides. We recently went viral on TikTok - here's what we learned. https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned, August 2021.

[19] Camille Cobb, Sruti Bhagavatula, Kalil Anderson Garrett, Alison Hoffman, Varun Rao, and Lujo Bauer. "I would have to evaluate their objections": Privacy tensions between smart home device owners and incidental users. *Proceedings on Privacy Enhancing Technologies*, 2021(4):54–75, October 2021.

[20] Matthew J.C. Crump, John V. McDonnell, and Todd M. Gureckis. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410, 2013.

[21] Anastasia Danilova, Alena Naiakshina, Anna Rasgauski, and Matthew Smith. Code Reviewing as Methodology for Online Security Studies with Developers - A Case Study with Freelancers on Password Storage. In *17th Symposium on Usable Privacy and Security*, pages 397–416, 2021.

[22] Pardis Emami-Naeini, Tiona Francisco, Tadayoshi Kohno, and Franziska Roesner. Understanding Privacy Attitudes and Concerns Towards Remote Communications During the COVID-19 Pandemic. In *17th Symposium on Usable Privacy and Security*, pages 695–714, 2021.

[23] Kelsey R. Fulton, Anna Chan, Daniel Votipka, Michael Hicks, and Michelle L. Mazurek. Benefits and Drawbacks of Adopting a Secure Programming Language: Rust as a Case Study. In *17th Symposium on Usable Privacy and Security*, pages 597–616, 2021.

[24] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[25] Joseph K Goodman, Cynthia E Cryder, and Amar Cheema. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.

[26] Thomas Groß. Validity and Reliability of the Scale Internet Users' Information Privacy Concerns (IUIPC). *Proceedings on Privacy Enhancing Technologies*, 2021(2):235–258, April 2021.

[27] David Harborth and Alisa Frik. Evaluating and Redefining Smartphone Permissions with Contextualized Justifications for Mobile Augmented Reality Apps. In *17th Symposium on Usable Privacy and Security*, pages 513–534, 2021.

[28] Ayako A. Hasegawa, Naomi Yamashita, Mitsuaki Akiyama, and Tatsuya Mori. Why they ignore English emails: The challenges of non-native speakers in identifying phishing emails. In *17th Symposium on Usable Privacy and Security*, pages 319–338, 2021.

[29] David J Hauser and Norbert Schwarz. It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *Sage Open*, 5(2), 2015.

[30] David J Hauser and Norbert Schwarz. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 48(1):400–407, 2016.

[31] Maximilian Häring, Eva Gerlitz, Christian Tiefenau, Matthew Smith, Dominik Wermke, Sascha Fahl, and Yasemin Acar. Never ever or no matter what: Investigating adoption intentions and misconceptions about the Corona-Warn-App in Germany. In *17th Symposium on Usable Privacy and Security*, pages 77–98, 2021.

[32] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy attitudes of mechanical turk workers and the U.S. public. In *10th Symposium On Usable Privacy and Security*, pages 37–49, 2014.

[33] Ankit Kariryaa, Gian-Luca Savino, Carolin Stellmacher, and Johannes Schöning. Understanding Users' Knowledge about the Privacy and Security of Browser Extensions. In *17th Symposium on Usable Privacy and Security*, pages 99–118, 2021.

[34] Smirity Kaushik, Yaxing Yao, Pierre Dewitte, and Yang Wang. "How I know for sure": People's perspectives on Solely Automated Decision-Making (SADM). In *17th Symposium on Usable Privacy and Security*, pages 159–180, 2021.

[35] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. Designing toxic content classification for a diversity of perspectives. In *17th Symposium on Usable Privacy and Security*, pages 299–318, 2021.

[36] Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. Digital Inequality Through the Lens of Self-Disclosure. *Proceedings on Privacy Enhancing Technologies*, 2021(3):373–393, July 2021.

[37] Ada Lerner, Helen Yuxun He, Anna Kawakami, Silvia Catherine Zeamer, and Roberto Hoyle. Privacy and activism in the transgender community. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[38] Leib Litman, Aaron Moss, Cheskie Rosenzweig, and Jonathan Robinson. Reply to MTurk, Prolific or panels? Choosing the right audience for online research. https://ssrn.com/abstract=3775075, 2021.

[39] Aaron Moss and Leib Litman. After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it. https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it/, 2018.

[40] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, 45(4):867–872, 2009.

[41] Stefan Palan and Christian Schitter. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, Mar 2018.

[42] Gabriele Paolacci and Jesse Chandler. Inside the turk: Understanding mechanical turk as a participant pool. *Current directions in psychological science*, 23(3):184–188, 2014.

[43] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.

[44] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[45] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, pages 1–20, 2021.

[46] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior research methods*, 46(4):1023–1031, 2014.

[47] Prolific. https://www.prolific.co.

[48] Hirak Ray, Flynn Wolf, Ravi Kuber, and Adam J. Aviv. "Warn Them" or "Just Block Them"?: Investigating Privacy Concerns Among Older and Working Age Adults. *Proceedings on Privacy Enhancing Technologies*, 2021(2):27–47, April 2021.

[49] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *IEEE Symposium on Security and Privacy*, pages 1326–1343, 2019.

[50] Joel Ross, Lilly Irani, M Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872. 2010.

[51] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *CHI Conference on Human Factors in Computing Systems*, pages 2202–2214, 2017.

[52] Sebastian Schnorf and Aaron Sedley. A comparison of six sample providers regarding online privacy benchmarks. In *Symposium on Usable Privacy and Security*, 2014.

[53] Daniel J Simons and Christopher F Chabris. Common (mis) beliefs about memory: A replication and comparison of telephone and mechanical turk survey methods. *PloS one*, 7(12):e51876, 2012.

[54] Daniel Smullen, Yaxing Yao, Yuanyuan Feng, Norman Sadeh, Arthur Edelstein, and Rebecca Weiss. Managing Potentially Intrusive Practices in the Browser: A User-Centered Perspective. *Proceedings on Privacy Enhancing Technologies*, 2021(4):500–527, October 2021.

[55] Katta Spiel, Oliver L Haimson, and Danielle Lottridge. How to do better with gender on surveys: A guide for HCI researchers. *Interactions*, 26(4):62–65, 2019.

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

[57] Peter Story, Daniel Smullen, Yaxing Yao, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. Awareness, Adoption, and Misconceptions of Web Privacy Tools. *Proceedings on Privacy Enhancing Technologies*, 2021(3):308–333, July 2021.

[58] Christian Stransky, Dominik Wermke, Johanna Schrader, Nicolas Huaman, Yasemin Acar, Anna Lena Fehlhaber, Miranda Wei, Blase Ur, and Sascha Fahl. On the Limited Impact of Visualizing Encryption: Perceptions of E2E Messaging Security. In *17th Symposium on Usable Privacy and Security*, pages 437–454, 2021.

[59] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Deciding on Personalized Ads: Nudging Developers About User Privacy. In *17th Symposium on Usable Privacy and Security*, pages 573–596, 2021.

[60] Jenny Tang, Hannah Shoemaker, Ada Lerner, and Eleanor Birrell. Defining Privacy: How Users Interpret Technical Terms in Privacy Policies. *Proceedings on Privacy Enhancing Technologies*, 2021(3):70–94, July 2021.

[61] Andrew J Thompson and Justin T Pickett. Are relational inferences from crowdsourced and opt-in samples generalizable? Comparing criminal justice attitudes in the GSS and five online samples. *Journal of Quantitative Criminology*, 36(4):907–932, 2020.

[62] Jan Tolsdorf, Florian Dehling, Delphine Reinhardt, and Luigi Lo Iacono. Exploring mental models of the right to informational self-determination of office workers in Germany. *Proceedings on Privacy Enhancing Technologies*, 2021(3):5–27, July 2021.

[63] United States Food and Drug Administration. Collection of race and ethnicity data in clinical trials: Guidance for industry and food and drug administration staff. https://www.fda.gov/media/75453/download, 2016.

[64] Rick Wash, Norbert Nthala, and Emilee Rader. Knowledge and capabilities that non-expert users bring to phishing detection. In *17th Symposium on Usable Privacy and Security*, pages 377–396, 2021.

[65] Shikun Zhang, Yuanyuan Feng, Lujo Bauer, Lorrie Faith Cranor, Anupam Das, and Norman Sadeh. "Did you know this camera tracks your mood?": Understanding Privacy Expectations and Preferences in the Age of Video Analytics. *Proceedings on Privacy Enhancing Technologies*, 2021(2):282–304, April 2021.

[66] Leah Zhang-Kennedy and Sonia Chiasson. "Whether it's moral is a whole other story": Consumer perspectives on privacy regulations and corporate data practices. In *17th Symposium on Usable Privacy and Security*, pages 197–216, 2021.

A Survey Questions

This appendix contains the list of questions asked during our user study. These questions are taken from the Pew American Trends Panel run between June 3-17, 2019. Each question could be skipped by the user.

1. Do you use any of the following social media sites?
   The order of the first three of the following questions are randomized
   (a) Facebook *[behav2]*
      - Yes, use this / No, do not use this
   (b) Instagram *[behav3]*
      - Yes, use this / No, do not use this
   (c) Twitter *[behav4]*
      - Yes, use this / No, do not use this
   (d) Any other social media sites *[behav5]*
      - Yes, use this / No, do not use this
2. In your own words, what does "digital privacy" mean to you?
   Participants are given a textbox to type in.
3. How often are you asked to agree to the terms and conditions of a company's privacy policy? *[exp1]*
   - Almost daily / About once a week / About once a month / Less frequently / Never
4. How confident are you, if at all, that companies will do the following things?
   The order of the following questions are randomized
   (a) Follow what their privacy policies say they will do with your personal information *[percep1]*
      - Very confident / Somewhat confident / Not too confident / Not confident at all
   (b) Promptly notify you if your personal data has been misused or compromised *[percep2]*
      - Very confident / Somewhat confident / Not too confident / Not confident at all
   (c) Publicly admit mistakes and take responsibility when they misuse or compromise their users' personal data *[percep3]*
      - Very confident / Somewhat confident / Not too confident / Not confident at all
   (d) Use your personal information in ways you will feel comfortable with *[percep4]*
      - Very confident / Somewhat confident / Not too confident / Not confident at all
   (e) Be held accountable by the government if they misuse or compromise your data *[percep5]*
      - Very confident / Somewhat confident / Not too confident / Not confident at all
5. How comfortable are you, if at all, with companies using your personal data in the following ways?
   The order of the first, second, and last questions are randomized
   (a) To help improve their fraud prevention systems *[percep6]*
      - Very comfortable / Somewhat comfortable / Not too comfortable / Not comfortable at all
   (b) Sharing it with outside groups doing research that might help improve society *[percep7]*
      - Very comfortable / Somewhat comfortable / Not too comfortable / Not comfortable at all
   (c) This question is not part of the survey and just helps us to detect bots and automated scripts. To confirm that you are a human, please choose 'Strongly agree' here.
      - Strongly disagree / Disagree / Somewhat disagree / Neither agree nor disagree / Somewhat agree / Agree / Strongly Agree

(d) To help them develop new products *[percep8]*

- Very comfortable / Somewhat comfortable / Not too comfortable / Not comfortable at all

6. Have you recently decided NOT to use a product or service because you were worried about how much personal information would be collected about you? *[behav1]*

   - Yes, have done this / No, have not done this

7. Here's a different kind of question. (If you don't know the answer, select "Not sure.") As far as you know… *The order of the following questions is randomized. For each question, the order of the first four options is randomized.*

   (a) If a website uses cookies, it means that the site… *[know1]*

   - Can see the content of all the files on the device you are using
   - Is not a risk to infect your device with a computer virus
   - Will automatically prompt you to update your web browser software if it is out of date
   - Can track your visits and activity on the site *[correct]*
   - Not sure

   (b) Which of the following is the largest source of revenue for most major social media platforms? *[know2]*

   - Exclusive licensing deals with internet service providers and cellphone manufacturers
   - Allowing companies to purchase advertisements on their platforms *[correct]*
   - Hosting conferences for social media influencers
   - Providing consulting services to corporate clients
   - Not sure

   (c) When a website has a privacy policy, it means that the site… *[know3]*

   - Has created a contract between itself and its users about how it will use their data *[correct]*
   - Will not share its users' personal information with third parties
   - Adheres to federal guidelines about deceptive advertising practices
   - Does not retain any personally identifying information about its users
   - Not sure

   (d) What does it mean when a website has "https://" at the beginning of its URL, as opposed to "http://"without the "s"? *[know4]*

   - Information entered into the site is encrypted *[correct]*
   - The content on the site is safe for children
   - The site is only accessible to people in certain countries
   - The site has been verified as trustworthy
   - Not sure

   (e) Where might someone encounter a phishing scam? *[know5]*

   - In an email
   - On social media
   - In a text message
   - On a website
   - All of the above *[correct]*
   - None of the above
   - Not sure

   (f) Which two companies listed below are both owned by Facebook? *[know6]*

   - Twitter and Instagram
   - Snapchat and WhatsApp
   - WhatsApp and Instagram *[correct]*
   - Twitter and Snapchat
   - Not sure

   (g) The term "net neutrality" describes the principle that… *[know7]*

   - Internet service providers should treat all traffic on their networks equally *[correct]*
   - Social media platforms must give equal visibility to conservative and liberal points of view
   - Online advertisers cannot post ads for housing or jobs that are only visible to people of a certain race
   - The government cannot censor online speech
   - Not sure

(h) Many web browsers offer a feature known as "private browsing" or "incognito mode." If someone opens a webpage on their computer at work using incognito mode, which of the following groups will NOT be able to see their online activities? *[know8]*

- The group that runs their company's internal computer network
- Their company's internet service provider
- A coworker who uses the same computer *[correct]*
- The websites they visit while in private browsing mode
- Not sure

8. Do you think that ALL Americans should have the right to have the following information about themselves removed from public online search results?
The order of the following questions is randomized

(a) Data collected by law enforcement, such as criminal records or mugshots *[belief1]*

- Yes, should be able to remove this from online searches / No, should not be able to remove this from online searches

(b) Information about their employment history or work record *[belief2]*

- Yes, should be able to remove this from online searches / No, should not be able to remove this from online searches

(c) Negative media coverage *[belief3]*

- Yes, should be able to remove this from online searches / No, should not be able to remove this from online searches

(d) Potentially embarrassing photos or videos *[belief4]*

- Yes, should be able to remove this from online searches / No, should not be able to remove this from online searches

9. Today it is possible to take personal data about people from many different sources – such as their purchasing and credit histories, their online browsing or search behaviors, or their public voting records – and combine them together to create detailed profiles of people's potential interests and characteristics. Companies and other organizations use these profiles to offer targeted advertisements or special deals, or to assess how risky people might be as customers. Prior to today, how much had you heard or read about this concept? *[exp5]*

- A lot / A little / Nothing at all

10. In the last 12 months, have you had someone...
The order of the following questions is randomized

(a) Put fraudulent charges on your debit or credit card *[exp2]*

- Yes / No

(b) Take over your social media or email account without your permission *[exp3]*

- Yes / No

(c) Attempt to open a line of credit or apply for a loan using your name *[exp4]*

- Yes / No

11. What is your age?

- 18-29
- 30-49
- 50-64
- 65+

12. What is your sex?

- Male
- Female

13. Please indicate your highest level of education

- Less than high school
- High school graduate
- Some college, no degree
- Associate's degree
- College graduate/some post grad
- Postgraduate

14. Choose the race that you consider yourself to be
The first four options are presented in randomized order

- White
- Black or African American
- Asian or Asian American
- Mixed Race
- Some other race

B Survey Response Summaries

Table 4 and Table 5 summarize how our four datasets compare on each of the thirty individual questions. Responses are within $\pm 5\%$ Pew proportions are highlighted in green; responses are $\geq 10\%$ off from Pew proportions ar highlighted in orange.

Q	Ans	Pew	Repr	Bal	MTurk
behav1	Yes	51.6	51.7	51.2	62.6
behav2	Yes	71.9	77.4	72	93.2
behav3	Yes	38.0	63.5	76.5	88.9
behav4	Yes	22.6	58.8	63.6	88.7
behav5	Yes	29.0	73.1	82.9	84.9
exp1	Daily	25.2	31.9	28	33.8
	Weekly	32.1	35.8	35.6	34.2
	Monthly	24.3	19.2	22.5	23.1
	Less	15.4	12.6	13.6	7.8
	Never	3.0	0.5	0.2	1.1
exp2	Yes	21.4	16.2	14.9	45.2
exp3	Yes	8.0	5.4	6.4	47.8
exp4	Yes	6.1	3.4	2.8	48.1
exp5	A lot	27.2	31.8	27.6	40.6
	A little	49.8	57	60.5	53.9
	Nothing	23.0	11.2	11.9	5.5
know1	Correct	62.6	89.7	86.6	48.6
	Incorrect	9.3	3.8	2.8	38
	Not sure	28.2	6.5	10.6	13.4
know2	Correct	58.9	80.5	80	48.9
	Incorrect	8.5	4.1	3.4	36.1
	Not sure	32.6	15.4	16.6	15
know3	Correct	47.8	68.8	71	44.4
	Incorrect	24.6	18.5	15.6	40.9
	Not sure	27.6	12.8	13.4	14.8
know4	Correct	30.3	56.2	54.2	37.8
	Incorrect	15.1	13	9.9	41
	Not sure	54.6	30.8	35.9	21.2
know5	Correct	67.1	75.5	76.2	31.6
	Incorrect	17.6	20.9	18.9	58
	Not sure	15.3	3.6	4.9	10.4
know6	Correct	28.7	67	73.1	64.7
	Incorrect	21.9	12.4	12.6	26
	Not sure	49.4	20.6	14.2	9.2
know7	Correct	44.6	59.9	58.3	43.2
	Incorrect	12.0	12.9	13.9	38.4
	Not sure	43.4	27.2	27.9	18.4
know8	Correct	24.4	37.1	53.4	26.6
	Incorrect	25.5	38.5	27.5	53
	Not sure	50.1	24.4	19.1	20.4

Table 4: Proportions of responses to each question for the full samples. Green responses are within ±5% Pew proportions, orange responses are ≥ 10% of Pew proportions.

Q	Ans	Pew	Repr	Bal	MTurk
percep1	VC	4.8	5.5	5.9	26.9
	SC	37.1	41.2	42.8	44.4
	NTC	40.3	37.6	37.5	21
	NCAA	17.7	15.6	13.9	7.8
percep2	VC	5.1	4	3.5	26
	SC	29.6	34.5	32.6	38.4
	NTC	40.6	40.6	40.5	22.5
	NCAA	24.8	20.9	23.4	13.1
percep3	VC	2.9	2.8	2	22.8
	SC	17.8	21	17	39.1
	NTC	46.4	47	49	25.9
	NCAA	32.9	29.2	32	12.2
percep4	VC	3.6	3.4	3.1	26.8
	SC	27.2	28	27.6	41.6
	NTC	47.1	46	48.5	22.4
	NCAA	22.1	22.6	20.8	9.2
percep5	VC	3.6	4.4	2.9	22.8
	SC	27.2	18.9	18.2	40.4
	NTC	47.1	42.2	42.4	21.5
	NCAA	22.1	34.5	36.5	15.4
percep6	VC	10.4	9.9	8.4	28.9
	SC	47.0	49	49.8	44.5
	NTC	28.5	29.5	31.5	18.1
	NCAA	14.1	11.6	10.4	8.5
percep7	VC	5.7	6.2	4.5	23.5
	SC	30.2	31.4	34.1	40.8
	NTC	36.6	35.9	35.1	23.6
	NCAA	27.4	26.5	26.2	12.1
percep8	VC	8.1	6.6	5	29.8
	SC	42.1	38	39.5	42.8
	NTC	31.3	33.2	32.8	16.6
	NCAA	18.5	22.1	22.8	10.9
belief1	Yes	39.1	43.2	38.8	71
belief2	Yes	67.3	76.9	74.4	82.3
belief3	Yes	56.1	43.5	36.4	67.2
belief4	Yes	84.9	85.1	82	83

Table 5: Proportions of responses to each question for the full samples. Green responses are within ±5% Pew proportions, orange responses are ≥ 10% of Pew proportions.