Exploring the Quality, Efficiency, and Representative Nature of Responses Across Multiple Survey Panels

Frank Bentley, Kathleen O'Neill, Katie Quehl
Yahoo/Verizon Media
Sunnyvale, CA, USA
[fbentley, kathleen.oneill,
katiequehl]@verizonmedia.com

Danielle Lottridge
University of Auckland
Auckland, NZ
d.lottridge@auckland.ac.nz

ABSTRACT
A common practice in HCI research is to conduct a survey to understand the generalizability of findings from smaller-scale qualitative research. These surveys are typically deployed to convenience samples, on low-cost platforms such as Amazon's Mechanical Turk or Survey Monkey, or to more expensive market research panels offered by a variety of premium firms. Costs can vary widely, from hundreds of dollars to tens of thousands of dollars depending on the platform used. We set out to understand the accuracy of ten different survey platforms/panels compared to ground truth data for a total of 6,007 respondents on 80 different aspects of demographic and behavioral questions. We found several panels that performed significantly better than others on certain topics, while different panels provided longer and more relevant open-ended responses. Based on this data, we highlight the benefits and pitfalls of using a variety of survey distribution options in terms of the quality, efficiency, and representative nature of the respondents and the types of responses that can be obtained.

Author Keywords
Survey; MTurk; SurveyMonkey; Representative.

CCS Concepts
•Human-centered computing → Empirical studies in HCI; Human computer interaction (HCI); •General and reference → Empirical studies;

INTRODUCTION
Surveys are a common method in HCI research that allow researchers to understand the generalizability of a finding to the broader population [37]. Often, surveys are used in conjunction with qualitative research methods so that researchers can first understand a range of behaviors or responses to a research question and then execute a survey to see how common these behaviors or attitudes are in the broader population. Sizing behaviors via surveys is critical for researchers who want their solutions to be beneficial for the broadest possible user base and to prioritize scarce design and development resources in corporate contexts.

When conducting these types of surveys, it is important to reach a representative sample of the population in order to correctly scope the opportunity for a solution to impact the largest number of people. For different types of products, different aspects of user demographics or attitudes are important. If one is building a general news site, ensuring that the respondents match the broader political landscape of the country is important for results to accurately reflect the potential users. If one is building an online video platform, making sure that the users surveyed subscribe to sources at the same rates and watch as much television content as the broader population is important. For most surveys, having representative ages and genders is critical to ensure that responses reflect the behaviors and attitudes of a broad section of the population.

In addition, surveys are also sometimes used to gather a large number of open-ended responses to a particular question [37]. These responses are then analyzed to find themes. For these types of questions, it's important to choose a panel that will provide not only a representative sample of respondents, but quality answers (e.g. not just Yes/No or "blah"). Due to the different ways participants are paid on different panels and their motivations for participating, these different incentives might impact how they answer open-ended questions and the overall usefulness of their responses.

Many papers and case studies in the broader HCI literature have deployed surveys using a variety of survey platforms and panels. Some have used samples of convenience from social media or university/corporate mailing lists [12, 17, 27, 31]. Others have deployed surveys to broader sets of users via Amazon Mechanical Turk (AMT) or the SurveyMonkey Audience panel [8, 21, 47, 43, 34]. Still others have paid market research firms for large, representative samples of respondents [6, 10, 42].

Each of these techniques incurs vastly different costs and time between when the questions are fully drafted and the data is back in the researcher's hand. Each method of surveying also might target very different user bases. Samples of convenience are free, but likely to be made up of respondents who are very similar (in age, education, geographical region, etc.) to the

CHI'20, April 25–30, 2020, Honolulu, HI, USA
 ISBN 978-1-4503-6708-0/20/04…$15.00
DOI: https://doi.org/10.1145/3313831.3376671

researchers. AMT and SurveyMonkey Audience panels are fairly inexpensive (typically $1 per participant for short surveys), but might have samples biased to types of people who would engage in crowd-work or complete surveys for charitable causes. Market research panels often cost thousands or tens of thousands of dollars per survey, but claim to guarantee a more representative panel of respondents.

Given the vast differences in cost and time, we wanted to better understand what that extra money was paying for, and if lower-cost survey platforms could be as accurate as more expensive options for conducting sizing studies on the broader United States population. Given the proliferation and maturity of these panels in the United States, we leave replicating this work in other markets to future work.

We examine the following research questions:

1. How do the audiences from major survey providers vary on aspects related to demographics, technology use, education/income, political affiliation, and other dimensions?
2. How do responses to open-ended questions vary among survey providers in terms of amount written and relevance of answers?
3. What are the limitations and advantages (e.g. speed) of lower cost survey providers, and in which situations might these limitations be acceptable?

We will explore answers to these questions below, discussing results from a single survey that was deployed to ten survey platforms/panels in the summer of 2019. We will explore differences in responses from these panels compared to external ground truth data for 80 response choices and will discuss the various tradeoffs implicit in choosing a particular deployment strategy for a survey.

RELATED WORK

A variety of papers have explored the efficacy of services for conducting survey-based research in HCI and related fields, including Müller et al.'s guide to best practices for surveys in HCI [37]. Landers et al. [33] discuss a wide variety of the different types of panels that survey researchers can use: from student samples, to Amazon Mechanical Turk (AMT), to other online panels. They discuss that the main question to answer when selecting a panel is its external validity for the topic that one is exploring. However they do not perform any studies to quantify the validity of any particular service or panel. Roulin [45] agrees with Landers that crowdsourced and convenience samples need to be examined in more detail and that there might be times when they are acceptable to use.

Researchers have explored the generalizability of panels on AMT. In exploring demographic diversity, Paolacci et al. [40] found AMT panels as demographically diverse as traditional university subject pools. Buhrmester et al. [9] found that AMT is more diverse than many existing Internet samples. Stewart et al. [46] showed that the diversity within AMT samples is similar to that of the population at large and Ross et al. [44] explored changing demographics in AMT workers over time. Antin et al. [2] explored motivations of AMT workers in the US and India.

Psychology researchers have demonstrated that many traditional studies have similar findings when using AMT panels. This holds for a wide variety of psychology studies including Prisoner's Dilemma, the Asian disease problem [7], and the Linda Problem[8]. Horton et al. replicated the Prisoner's Dilemma experiment on AMT and reproduced the findings from previous laboratory experiments [5]. Goodman et al. [28] conducted a study to compare the responses of 107 AMT participants to those of two samples (one with community members and one of students) and found that AMT responses are consistent with standard decision-making biases. Buhrmester et al. [9] found that a sample of 3,000 AMT participants met the psychometric standards associated with published research. Furthermore, a study by Paolacci et al. [40] compared the responses of three samples of participants: one from AMT, one from a traditional university pool, and one from an online discussion board. All participants completed three classic experimental tasks: the Linda Problem, Asian disease problem, and Physician problem. The results from the AMT workers were similar to those of the other two samples, provide evidence that AMT is a reliable source of judgment and decision-making data.

While standard psychology experiments which aim to evaluate universal behaviors and attitudes show similar results on AMT and other platforms, it's less clear that studies about preferences, attitudes, and real-world behaviors hold when comparing samples from AMT with the general population. It is often these types of questions that are more relevant for HCI researchers seeking to size the market opportunity of a given design idea.

Bentley et al. [7] compared results from the SurveyMonkey Audience panel, AMT, and a larger market research survey on a variety of questions. Limitations of this work included a lack of ground truth for comparing many of the responses and that not all surveys were fielded at the same time. This inspired us to perform a more rigorous study using additional panels and additional behavioral and attitudinal questions for which we could find a more accurate ground truth.

Researchers have found behavioral differences between participants brought to a company or lab and AMT participants. Findlater et al. [22] compared performance of 30 lab-based and age-matched AMT participants for speed accuracy tradeoffs in Fitts' law related mouse and touchscreen tasks and found those from AMT were significantly faster and less accurate. Locascio et al. [35] explored using employees versus external participants for usability studies. While not focusing on surveys, they demonstrated that employees demonstrated different behaviors and gave different System Usability Score (SUS) ratings to products when compared with external participants. This motivated us to include an internal survey panel in our research, to see if we also saw differences in demographics, behaviors, and attitudes when surveying tech employees.

Given all of this related work, there was still a need to broadly examine a variety of panels in a systematic way. While some researchers have examined one panel or another for specific demographic or behavioral trends, we were unable to find any work that systematically compared a variety of profes-

	Omni-bus	Omnibus Raw	SBA	SM 1000	Users 2	SM 250	Users 1	AMT	UXR Panel	Employ-ees
Completed N	997	1106	893	1059	300	305	596	250	416	85
Time to Results	41 hours	41 hours	6 days	15 hours	6 days	9 hours	6 days	2 hours	48 hours	24 hours
Completion Rate	80%	95%	56%	92%	86%	97%	67%	92%	87%	76%
Completion Time	5:19	5:27	11:00	5:00	9:30	5:02	9:30	4:54	6:00	5:53
Total Cost	$6,000	$5,800	$9,800	$2,000	$1,800	$500	$1,800	$350	$0	$0
Median Error	3.9%	4.4%	4.4%	4.9%	5.3%	5.4%	6.6%	7.0%	10.4%	16.4%

Table 1. The 10 survey panels that were used to deploy an identical survey in Summer 2019. Columns are arranged from least to most median error.

sional, crowdsourced, and convenience panels on a wide rage of demographic, technology use, and behavioral questions. This analysis is important as HCI researchers and industry practitioners often choose panels without considering the representative nature of the sample beyond basics of age, gender, and income/education.

METHODS

After exploring the literature, we crafted a 21-question survey to capture diverse aspects of demographics and technology use. The survey asked users about devices that they owned (e.g. smartphones, laptops, e-readers, etc.), services that they subscribed to (from local newspaper delivery to video streaming sites), types of financial products that they used, and a variety of questions about technology and media use. It also asked for detailed demographic data including age, education, household income, and approval of the US President. The survey was designed to be completed in about five minutes.

We obtained ground truth data for 80 response options from a variety of sources. We favored sources based on actual collected data, such as from the US Census or a company's earnings report stating the number of users that they had within the United States. Other data came from respected industry sources such as Comscore, eMarketer, or Nielsen which base their data on behavioral trackers embedded in websites or from very large panels of users who have installed these trackers on their own devices to capture actual product use. Political opinion data was taken from the FiveThirtyEight weighted average of polls [23]. In the tables that follow, ground truth sources will be noted as a reference for each response section.

We chose a wide array of platforms for deploying the survey, and all surveys were deployed during the same two-day period in the summer of 2019. On the lowest end, we chose a widely subscribed email list internal to our organization (Employees in Table 1). This was free to send to, and is similar to surveys deployed in many HCI papers [12, 17, 27, 31]. We received 85 complete responses from this request.

We also deployed the survey to three lists of external users who have signed up to provide regular feedback to our organization on new features and marketing campaigns. We received 416 responses from a nationwide panel of users who have volunteered to participate in UX Research studies (UXR Panel), 596 from the second panel (Users 1, those who have volunteered to provide feedback for products from one of our brands), and 300 from the third panel (Users 2, those who have volunteered to provide feedback for another of our brands). Both of these panels are paid for responding to surveys that we distribute through a third party vendor.

Other low-cost options included Amazon Mechanical Turk (AMT). We paid respondents $1.00 per completion, an hourly rate above $12, which is standard for surveys from our institution (and higher than many published papers [5, 9]). The survey was fielded from the account of one of the authors, which has been used for several dozen AMT projects over the past five years. We used standard settings of participants who had completed at least 100 HITs with at least a 97% approval rate and were located in the United States. Existing research [7], and our own experience, have indicated that these filters limit scammers and still yield diverse samples of participants. We purchased 250 responses from this method.

The survey was also deployed to a paid SurveyMonkey Audience panel, at a rate of $2 per completed response. This platform provides basic balancing of respondents based on age, gender, income, and US region. We ran two deployments on this platform, one with 250 paid respondents (delivered 305) and another with 1,000 (delivered 1,059) for comparison. Frequently, researchers would like to run cheaper, smaller samples, and we were curious how the smaller deployment would compare to the larger panel at one quarter the cost.

On the high end, we deployed the survey to a number of paid panels from Market Research firms. We used the Omnibus panel from Dynata (formerly Research Now),[1] which cost $6 per participant for a total of 997 cleaned responses and SpringBoard America,[2] at a cost of $11 per participant, with a total of 893 complete responses.

Results from each panel were then analyzed to find significant differences between responses and the ground truths for each question. We will explore each set of questions in the Findings section below and then will discuss the implications of the differences that we observed for researchers who are aiming to pick the best panel to use to deploy their own surveys.

In addition to quantifying differences between these panels, we included two open-ended responses in our survey, one asking participants to "Tell us about the last time you received a notification that annoyed you. What was the notification and why did you not like it? (Please write at least 2 sentences)."

[1] https://www.dynata.com/market-researcher-solutions/research-services/

[2] https://www.springboardamerica.com/

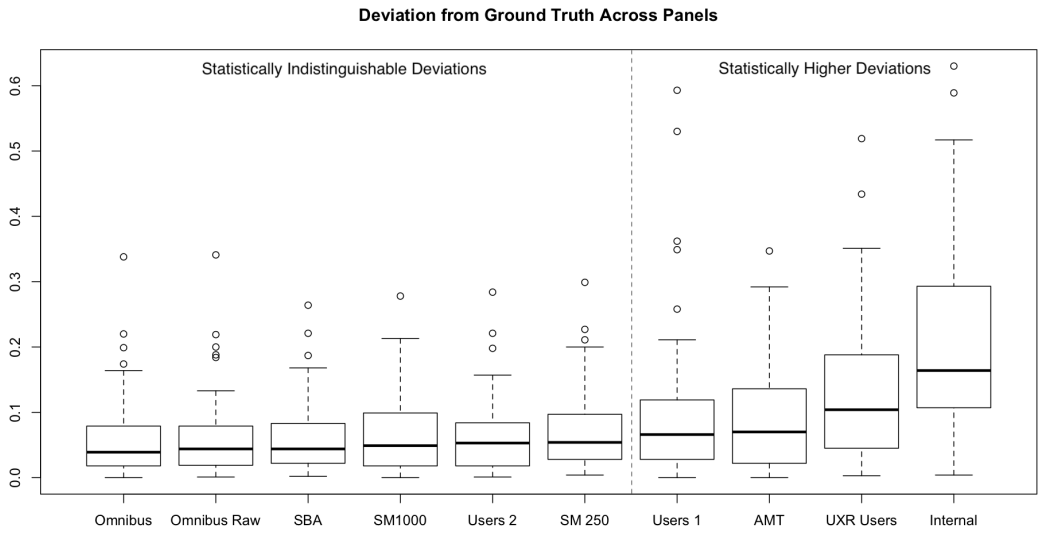

Figure 1. Distribution of deviations for each of the panels for the 80 responses that contained a ground truth. Panels from User 1 to the right all had statistically significant higher deviations from ground truth than the Omnibus panel.

The other asked participants: "What motivates you to complete surveys such as this one? Why are you a part of this panel? (please provide 2 or more sentences)." We were interested in exploring how participants responded to these questions, both in terms of the length of response on different survey platforms, but also to understand their motivations for taking surveys. We coded the open-ended question data using a grounded theory approach which leveraged both inductive and deductive coding schema. First, we categorized each of the responses along two dimensions: Length of the response and if they followed the directions for the amount of response we wanted, and second if they wrote a valid and relevant response to the question. We also coded for any responses which were obviously poor quality or non-responses such as the random typing of letters (e.g. "iaoisjdo"). During this first pass, we also conducted a close reading of the responses in order to iteratively develop a codebook for why people participated in taking surveys. This codebook was used in our second round of coding in which we coded all responses into 34 categories which fell into 10 major themes.

FINDINGS

In this section we will explore the data received from the ten different survey panels on the 80 responses for which we have ground truth data. We will begin by discussing the different demographics that each panel provided in their samples, and the skews that exist in samples from various panels. Next, we will explore the behavioral questions: which devices people owned, their travel habits, and which devices they use. We will then move on the attitudinal questions, such as the approval of the American president, or if they prefer to turn off their email notifications. Finally, we will explore the responses to two open-ended questions that we placed in the survey to better understand which panels are more suitable for gathering useful and complete qualitative data. This will also include analysis of the open-ended question on why participants were taking surveys on each platform.

Overall Deviations from Ground Truth

We collected ground truths, typically data from governmental sources or behavioral data captured from companies such as ComScore, for 80 of the possible responses in our survey (added as references to Tables 2, 3, and 4). When comparing the responses from each panel to these ground truths, we identified an error score for each question/panel pair of the absolute value of the difference between the value from the survey and the ground truth (in percentage points). For example, 49.2% of the US population is male, [25] but the SurveyMonkey 1000 panel was 47.4% male. Therefore, it had an error of 1.8 percentage points on this demographic attribute.

Figure 1 shows the distribution of errors for each survey source, for the 80 responses where there was a ground truth. As is clear from the figure, there was not a wide difference in errors observed from many of our panels. The Omnibus panel was the most accurate overall with a median error of 4.0% and a mean error of 5.4%. The highest error came from our internal company mailing list, with a median of 16% and a mean error of 20% (in other words, the average response was 20 percentage points off of the ground truth). For example, 88% of participants on the corporate mailing list had an iPhone, whereas only 42% of the U.S. population does, leading to an error of 36 percentage points for this question.

It is interesting to note that panels that cost many dollars per participant were statistically no more accurate than panels that were very low cost. The error in the Omnibus panel was not significantly different from Omnibus Raw, Springboard America, SurveyMonkey 1000, Users 2, and SurveyMonkey 250. However it was significantly less than the error in Users 1 (Welch's t=-2.09, $p < 0.05$), Mechanical Turk (t=-

Response	Ground Truth	Omni-bus	Omnibus Raw	SBA	SM 1000	Users 2	SM 250	Users 1	AMT	UXR Panel	Employ-ees
Gender (Gender Distribution: [15])											
Male	49.2%	49.0%	51.0%	51.0%	47.4%	43.0%	43.0%	64.0%	61.6%	61.1%	62.4%
Female	50.8%	51.1%	49.0%	49.0%	52.6%	57.0%	57.1%	35.0%	38.4%	39.2%	37.7%
Age (Age Ranges: [15])											
18-24	13.9%	13.6%	15.7%	2.0%	15.2%	1.0%	16.7%	2.0%	9.6%	4.3%	1.2%
25-34	18.8%	18.9%	19.9%	26.0%	23.8%	10.0%	26.6%	8.0%	46.4%	21.4%	31.8%
35-44	17.3%	17.4%	17.2%	18.0%	15.0%	16.0%	16.1%	23.0%	27.2%	25.2%	35.3%
45-54	17.3%	17.5%	16.2%	17.0%	21.4%	27.0%	20.7%	24.0%	10.0%	24.0%	18.8%
55-64	18.8%	18.8%	17.5%	19.0%	14.2%	29.0%	9.8%	22.0%	5.2%	21.6%	10.6%
65+	13.8%	13.9%	13.5%	18.0%	9.5%	17.0%	8.2%	20.0%	1.6%	3.4%	2.4%
Living Status (Household composition: [26], Pets: [25])											
Live Alone	9.7%	21.2%	21.9%	17.0%	13.5%	20.0%	11.8%	19.0%	30.8%	19.2%	8.2%
Live Partner	58.0%	51.8%	49.2%	61.0%	51.2%	63.0%	56.1%	61.0%	43.6%	52.9%	62.4%
Live < 18	31.0%	22.5%	22.7%	34.0%	24.8%	25.0%	31.8%	25.0%	28.0%	29.3%	45.9%
Live Pet	48.4%	14.6%	14.3%	22.0%	35.9%	20.0%	36.1%	33.0%	19.2%	20.9%	32.9%
Education (Educational Attainment: [15])											
< High School	2.9%	2.4%	2.6%	1.0%	2.9%	1.0%	2.0%	1.0%	0.0%	0.5%	0.0%
High School	19.7%	17.9%	17.3%	16.0%	17.7%	12.0%	19.3%	5.0%	16.8%	6.7%	1.2%
Some College	31.1%	24.3%	23.6%	22.0%	29.4%	24.0%	26.6%	28.0%	15.6%	22.4%	12.9%
College Grad	30.1%	30.5%	30.5%	35.0%	24.7%	37.0%	25.3%	30.0%	49.2%	38.2%	42.4%
Post-Graduate	17.5%	19.2%	20.0%	21.0%	16.9%	20.0%	20.3%	29.0%	10.0%	24.5%	42.4%
Ethnicity (Ethnic Backgrounds: [15])											
African Amer.	12.3%	12.7%	13.5%	10.0%	7.4%	7.0%	5.3%	10.0%	13.2%	13.5%	0.0%
Asian/Pacific	5.4%	7.2%	7.9%	7.0%	11.6%	5.0%	10.2%	7.0%	7.2%	18.5%	27.1%
Caucasian	78.0%	74.1%	72.4%	75.0%	66.3%	85.0%	69.2%	71.0%	78.4%	53.9%	61.2%
Hispanic	14.6%	7.8%	8.1%	12.0%	12.1%	5.0%	13.8%	7.0%	8.0%	14.9%	7.1%
Native Amer.	0.8%	2.7%	2.9%	2.0%	3.0%	2.0%	2.6%	3.0%	0.8%	1.7%	1.2%
Household Income (Income Distributions: [15])											
$0-25k	12.60%	16.6%	17.1%	12.0%	22.9%	13.0%	21.0%	9.0%	18.4%	9.6%	0.0%
$24-50k	35.20%	37.3%	36.0%	43.0%	37.7%	34.0%	40.3%	29.0%	56.4%	28.1%	5.9%
$75-100k	14.8%	13.9%	14.6%	17.0%	13.2%	15.0%	11.2%	12.0%	12.4%	17.6%	2.4%
$100k+	37.4%	26.6%	26.8%	23.0%	17.8%	34.0%	17.4%	29.0%	11.6%	37.0%	75.3%
City Type (Living Areas: [15])											
City >500k	31.0%	31.2%	33.2%	39.0%	39.2%	27.0%	40.7%	36.0%	40.0%	59.9%	51.8%
Suburb	55.0%	58.1%	56.4%	52.0%	50.1%	63.0%	51.8%	55.0%	51.2%	37.3%	44.7%
Rural Area	14.0%	10.7%	10.4%	9.0%	10.6%	9.0%	7.5%	9.0%	8.8%	2.9%	3.5%
United States Region (Population Distribution: [11])											
Northeast	17.0%	21.4%	20.3%	20.0%	14.4%	25.0%	12.8%	18.0%	17.2%	22.1%	5.9%
Midwest	19.2%	21.3%	20.9%	29.0%	18.5%	21.0%	21.6%	19.0%	20.8%	14.7%	0.0%
South	37.0%	36.5%	37.3%	24.0%	27.1%	29.0%	26.9%	37.0%	30.0%	15.9%	14.1%
West	23.6%	20.9%	21.5%	27.0%	32.5%	24.0%	32.1%	26.0%	26.8%	44.2%	75.3%

Table 2. Differences in the demographics of survey respondents across panels. Highlighted cells have greater than a 10 percentage point deviation from the ground truth (ground truth references are in the section headers). Green cells are over-represented and red cells are under-represented.

3.35, p=0.01), UXR Users (t=-5.6, p<0.001), and the Internal mailing list (t=-8.7, p<0.001). Our SurveyMonkey 250 study was executed for a total cost of $500, compared to $6,000 for the Omnibus.

Audience Demographics

The first section of survey questions focused on the demographics of the participants in each panel. We wanted to know if particular panels skewed towards certain types of people, and if some panels were better than others for reaching otherwise hard to reach audiences (e.g. 18-24 year-olds, higher earners, etc.). Getting demographics that match the broader US population is important when trying to understand audiences for new concepts. If some large segments of the population are not properly represented in a survey, significant behaviors and needs may not be seen.

Table 2 highlights the results from these demographic questions compared to ground truths (mostly from the US Census Bureau and ComScore US Online Population data as shown in the references for each section). Cells that are greater than 10% off from the ground truth are colored (green for 10+%

Response	Ground Truth	Omni-bus	Omnibus Raw	SBA	SM 1000	Users 2	SM 250	Users 1	AMT	UXR Panel	Employ-ees
Device Ownership (Mobile Devices: [15], eReaders [41], Computers/TV/Gaming: [18], Smart Speakers [39])											
iPhone	42.4%	47.3%	46.8%	43.0%	55.9%	42.0%	62.0%	50.0%	39.6%	51.4%	88.2%
iPad	20.7%	32.7%	32.4%	36.0%	39.7%	33.0%	39.0%	41.0%	29.2%	45.0%	68.2%
Android Phone	45.9%	47.3%	45.6%	52.0%	47.9%	47.0%	44.9%	50.0%	61.2%	55.3%	32.9%
Android Tablet	24.7%	24.5%	24.3%	32.0%	27.1%	28.0%	26.6%	29.0%	34.8%	33.9%	21.2%
eReader	26.0%	18.7%	18.1%	27.0%	26.2%	25.0%	27.9%	33.0%	28.0%	29.1%	42.4%
Pers. Laptop	55.2%	64.4%	61.5%	72.0%	72.0%	75.0%	70.8%	81.0%	81.2%	81.0%	84.7%
Pers. Desktop	58.8%	41.4%	40.4%	54.0%	42.1%	62.0%	42.3%	56.0%	64.0%	52.6%	40.0%
Streaming TV	42.3%	35.9%	34.4%	44.0%	50.5%	44.0%	52.8%	49.0%	53.2%	62.3%	74.1%
Smart TV	54.7%	44.2%	42.2%	48.0%	53.8%	55.0%	53.8%	51.0%	48.0%	56.0%	65.9%
Game Console	37.1%	35.2%	34.3%	41.0%	45.0%	38.0%	50.5%	39.0%	60.0%	48.3%	47.1%
Smart Speaker	21.0%	23.6%	23.6%	32.0%	32.2%	33.0%	37.7%	31.0%	34.0%	42.3%	56.5%
Subscriptions (Streaming Services: [15], Cable and Prime: [18], Newspapers [3])											
Spotify	10.8%	14.1%	15.8%	16.0%	21.6%	6.0%	22.3%	10.0%	27.6%	25.0%	41.2%
Pandora	5.1%	8.1%	9.1%	11.0%	8.8%	5.0%	8.2%	7.0%	6.4%	10.8%	7.1%
Hulu	18.7%	25.5%	25.6%	27.0%	35.1%	18.0%	36.4%	24.0%	34.8%	32.5%	28.2%
Netflix	41.9%	53.5%	53.3%	56.0%	63.0%	47.0%	64.6%	53.0%	66.4%	66.1%	83.5%
HBO Now	6.6%	15.4%	16.1%	22.0%	16.8%	20.0%	16.1%	24.0%	20.0%	32.9%	34.1%
Cable	68.3%	51.9%	49.5%	60.0%	40.6%	71.0%	38.4%	58.0%	33.6%	54.8%	36.5%
Amazon Prime	53.6%	42.1%	40.3%	50.0%	62.4%	52.0%	60.3%	58.0%	68.8%	68.0%	83.5%
Newspaper	16.8%	14.7%	14.9%	19.0%	8.1%	8.2%	22%	15.0%	5.2%	14.9%	11.8%
Settings Preferences [internal data]											
Email Notif	57.0%	59.6%	59.9%	52.0%	54.8%	43.0%	56.4%	44.0%	58.4%	58.4%	34.1%
Internet Use (Site Visitation: [15])											
aol.com	13.8%	8.6%	8.9%	12.0%	5.9%	10.0%	4.6%	7.0%	3.6%	16.4%	10.6%
AOL Mail	4.6%	10.8%	11.0%	15.0%	8.8%	13.0%	6.2%	8.0%	6.8%	19.0%	8.2%
Hooli (AC)	0.0%	0.9%	1.6%	2.0%	1.1%	1.0%	1.0%	0.0%	0.8%	0.5%	2.4%
HuffPost	19.0%	8.8%	9.8%	15.0%	14.0%	15.0%	15.4%	26.0%	19.6%	36.6%	48.2%
Tumblr	12.2%	6.3%	7.2%	10.0%	10.6%	6.0%	7.9%	9.0%	9.6%	17.6%	16.5%
yahoo.com	30.8%	22.2%	21.8%	30.0%	20.0%	23.0%	20.7%	67.0%	22.8%	65.9%	72.9%
Yahoo Mail	29.7%	39.0%	39.0%	43.0%	31.0%	28.0%	32.8%	89.0%	34.0%	73.1%	65.9%
Yahoo Finance	13.3%	10.5%	11.0%	15.0%	7.4%	11.0%	5.9%	27.0%	11.2%	40.6%	54.1%

Table 3. Differences in the technology use of survey respondents across panels. Highlighted cells have greater than a 10 percentage point deviation from the ground truth (ground truth references are in the section headers). Note Hooli, which was an attention check as this is a fictional company.

above; red for 10+% below) to easily see where certain panels are not performing well.

Most panels were fairly well representative of gender (we included an opened-ended response for self-described genders, following best practices from [29], but only received one response for this, in the Users 1 panel). Four panels (Users 1, Mechanical Turk, the UXR Panel, and the Employees panel) skewed higher towards males.

When exploring ages of respondents, Springboard America, the Users 1 panel, the Users 2 panel, and the Employees list were significantly underrepresented in the 18-24 demographic. The Mechanical Turk panel was underrepresented in adults above 55 years-old.

For household income, most panels underrepresented for $100k+ households. Interestingly, the user panels (Users 1, Users 2, and the UXR Panel), performed best across all income ranges, while the Employees panel overindexed on high earners and underrepresented on all other incomes.

Technology Use

Next, we will examine the representative nature of each panel in terms of how respondents use various technology in their lives, as shown in Table 3. In general, most panels overindexed on various aspects of technology ownership and use.

All panels overindexed on iPad ownership by 9% or more, with the UXR Panel and the Employee panel overindexing the most. All also overindexed on Laptop ownership and most underindexed on Desktop use. However, the Users 2 and AMT panels overindexed on Desktop ownership.

Almost all panels underindexed on subscriptions to Cable/Satellite television services. It seems those who are answering surveys online, in general, are less likely to subscribe to cable. However, the Users 2 panel was within 3% of the ground truth. As mentioned in the previous section, this group tended to be older, and older adults watch much more television that younger adults. [32]

Response	Ground Truth	Omni-bus	Omnibus Raw	SBA	SM 1000	Users 2	SM 250	Users 1	AMT	UXR Panel	Employ-ees
Travel (Travel Habits: [20])											
No Travel	19.0%	15.4%	15.8%	22.0%	11.4%	26.0%	11.8%	14.0%	17.6%	9.4%	2.4%
Sports (Viewership: [38])											
View Superbowl	41.3%	61.2%	61.3%	60.0%	48.7%	57.0%	50.8%	60.0%	53.2%	69.5%	57.7%
Finance (Banking/Investing/Student Loans: [15], Credit Cards: [14], Vehicle Loans: [36], Mortgage: [30], Medical Debt: [13])											
Checking Acc	88.3%	83.8%	79.1%	90.0%	79.6%	91.0%	82.6%	91.0%	86.4%	92.3%	97.7%
Savings Acc	67.6%	69.7%	65.6%	76.0%	66.0%	73.0%	70.5%	76.0%	66.0%	78.6%	83.5%
Money Market	14.4%	17.2%	17.6%	20.0%	12.9%	20.0%	11.5%	21.0%	10.0%	27.4%	29.4%
IRA/Roth IRA	25.0%	31.3%	30.0%	32.0%	23.6%	38.0%	21.6%	38.0%	18.0%	37.5%	65.9%
401k/403b Plan	29.9%	37.8%	36.0%	37.0%	34.7%	43.0%	33.4%	45.0%	28.8%	48.6%	92.9%
Brokerage/Invest.	24.5%	23.2%	22.4%	30.0%	24.6%	32.0%	21.6%	39.0%	18.8%	43.3%	64.7%
Student Loans	7.5%	13.7%	13.9%	17.0%	22.3%	16.0%	24.9%	18.0%	31.2%	21.2%	12.9%
Credit Card	70.2%	73.0%	69.7%	78.0%	68.6%	80.0%	69.5%	81.0%	70.0%	84.1%	95.3%
Vehicle Loan	33.2%	30.3%	28.8%	31.0%	32.7%	28.0%	32.8%	34.0%	25.6%	30.3%	45.9%
Mortgage	40.3%	31.3%	30.2%	37.0%	28.0%	40.0%	26.9%	38.0%	24.0%	32.7%	58.8%
Medical debt	13.9%	7.8%	8.0%	15.0%	14.6%	7.0%	16.4%	10.0%	13.6%	10.8%	3.5%
Politics (Presidential Approval: [23])											
Trump - approve	41.4%	36.5%	38.1%	37.0%	27.8%	32.0%	20.3%	26.0%	27.6%	20.7%	3.5%

Table 4. Differences in the lifestyle or attitudes of survey respondents across panels. Highlighted cells have greater than a 10 percentage point deviation from the ground truth (ground truth references are in the section headers).

	Omni-bus	Omnibus Raw	SBA	SM 1000	Users 2	SM 250	Users 1	AMT	UXR Panel	Employ-ees
Mean Characters	53	51	72	65	65	62	88	103	94	75
2+ Sentences	26.2%	24.2%	42.8%	49.3%	39.7%	52.5%	55.4%	84.4%	48.6%	35.3%
<2 sentences	72.6%	73.9%	57.2%	51.1%	59.7%	46.6%	44.6%	15.6%	51.4%	64.7%
Relevant	97.3%	90.2%	98.8%	92.9%	99.7%	94.1%	99.7%	92.0%	98.6%	96.5%
Non-relevant	2.7%	9.0%	1.2%	7.5%	0.3%	5.9%	0.3%	7.2%	1.4%	3.5%
"Junk"	0.0%	2.7%	0.0%	5.3%	0.0%	5.3%	0.0%	4.4%	0.5%	0.0%

Table 5. Length and quality of open-ended responses in each survey panel sample.

Despite higher deviations from the ground truth on device ownership and subscriptions questions, participants were much closer to the actual usage for a variety of different Internet sites. Users 1, the UXR Panel, and the Employee panel were overindexed on Yahoo properties, while Omnibus underrepresented HuffPost users and Springboard America had over three times the percent of AOL Mail users as the ground truth from ComScore.

Note in the list of Internet sites that we listed Hooli, a fictional company from the HBO show Silicon Valley. This was an attention check question, to make sure users were not just blindly clicking responses to continue in the survey. Less than 2.5% of participants picked Hooli in each panel, with the highest percent in the Employee panel (who were likely aware of the reference and having fun with us by choosing it), and Springboard America, at 2.0%. The fact that all of these are so low gives us additional confidence in these panels and the accuracy of responses to the other questions in the survey.

Lifestyle and Attitudes

Our final set of questions centered on various lifestyle and attitudinal questions, as shown in Table 4.

The panels all performed well at capturing the share of Americans who do not travel in a given year, however the Employee panel significantly underindexed here. All of our panels were more likely than the ground truth data to have watched the Superbowl in 2019, with the SurveyMonkey panel being the closest to actual viewership data.

Participants in Users 1, the UXR Panel, and the Employee panel were more likely to use a variety of financial products such as Savings Accounts, IRAs, 401(k) plans, Investment Accounts, Credit Cards, and various types of loans. However, all panels overindexed on the use of Student Loans, with Mechanical Turk being the highest at 21.2% compared to a ground truth of 7.5% of Americans.

Reaching a politically representative audience is often important for surveys relating to news or public policies. We asked participants "Do you approve or disapprove of the way Donald

	Omni-bus	Omnibus Raw	SBA	SM 1000	Users 2	SM 250	Users 1	AMT	UXR Panel	Employ-ees
Compensation	70.1%	64.8%	70.6%	71.8%	65.7%	79.02%	43.8%	83.6%	41.6%	8.2%
Form of Employment	2.7%	2.4%	2.2%	2.5%	1.7%	0.7%	0.2%	14.8%	0.5%	0.0%
Being Heard/Opinion	20.6%	18.6%	34.6%	19.2%	41.0%	20.3%	35.9%	7.6%	37.0%	8.2%
Having Impact	4.8%	4.3%	8.9%	6.9%	8.7%	4.9%	15.4%	2.4%	21.4%	14.1%
Supporting Research	4.1%	3.9%	3.6%	6.0%	2.3%	5.6%	5.5%	13.6%	7.9%	62.4%
Pastime	22.7%	21.0%	33.2%	25.1%	29.3%	23.0%	25.5%	30.0%	28.4%	8.2%
Passion for topic	4.3%	3.9%	5.4%	2.8%	8.7%	3.0%	18.6%	4.4%	12.3%	10.6%
Research Experience	0.2%	0.2%	0.1%	0.6%	0.3%	0.3%	0.8%	0.4%	0.5%	4.7%
Other	3.5%	3.2%	2.7%	3.6%	2.3%	2.3%	3.9%	3.2%	8.2%	3.5%
Survey Design/Platform	4.3%	4.2%	6.6%	6.6%	5.0%	7.2%	5.9%	2.0%	4.6%	17.7%

Table 6. Percent of respondents who were motivated to take surveys for various reasons.

Trump is handling his job as president?" with options for Approve, Disapprove, and No Opinion. This is the standard form of the question, used in dozens of professional polls. The latest weighted polling from fivethirtyeight.com at the time of the survey had Approve at 41.4% of the American population. The Omnibus and Springboard America panels were relatively close to this value, however all other panels significantly underindexed. For polls where political representation is important, for example, when building a general audience news aggregation platform, it is critical to choose a panel that will represent the opinions and behaviors of the entire American population.

Differences in Open-Ended Responses

Often, survey platforms are used for capturing qualitative data from a larger sample than can be interviewed in a lab setting. We wanted to understand how panels differed in terms of the length and relevance of open-ended responses. We had asked users to write two sentences for each of the two open-ended questions. We calculated the percent of respondents in each panel who followed those directions, the length in characters of the response, and if their responses were relevant in answering the question that we asked. Table 5 shows that AMT workers were the most likely to follow directions and write two or more sentences compared to other panels, but did have irrelevant answers or junk responses. The raw data from the Omnibus panel had the most irrelevant answers, but the cleaned data show vast improvement in individuals who fully answered the question. SBA, Users 1, and Users 2 all had data which was pre-cleaned by the panel companies and had the most relevant answers, although they were shorter in length than others.

In Table 6, we provide an overview of the motivations/reasons why individuals participate in each of the panels we studied.

Compensation and Employment

The most prevalent theme for why individuals took surveys was for the compensation they received. Responses ranged from simple and short phrases such as *"Money"* or *"I do it for the gift cards"* to more descriptive answers such as *"I complete surveys to make a little extra cash on the side. I've used it to pay for groceries for the last 6 months."* or *"I'm disabled and don't work, we're on a fixed income. My hobby is taking surveys and the gift cards I get are like an allowance to spend any way I like."*

Compensation varied by survey platform. Omnibus had the widest range of compensation methods mentioned including: points that could be redeemed, airline or travel/loyalty rewards points, ability to get gift cards to specific stores, or cash. As a "panel of panels" this makes sense since respondents might be entering the survey from several different initiation points compared to other panels which have single incentive programs. SurveyMonkey respondents mentioned being either paid via the ability to redeem gift-cards through their mobile "SurveyMonkey Rewards" or through "SurveyMonkey Contribute" which donates to a non-profit for each survey completed. SurveyMonkey also offers the ability to win random prizes for cash or gift cards. The SBA, Users 1, and Users 2 panels were compensated via giftcards and raffles. AMT workers are paid directly into their bank accounts. Employees who mentioned compensation mentioned other past internal surveys with raffles or that helping the company was part of their job and therefore covered by their salary.

Whereas some responses were just about getting paid and rewarded, others focused on working towards particular goals such as hitting $100 in a particular time period, saving up for a purchase such as a laptop, or working for extra ways to pay for holiday shopping. Other respondents, although a smaller percentage of individuals, considered taking surveys a job or type of supplementary or secondary income rather than just a way *"to make pocket change"* or that it *"helps pay for beer"*. This was highest on AMT where over 14% of respondents saw their work on the platform as either a secondary job or full time job and primary source of income. Retirees, those who are disabled, stay-at-home-parents, or unemployed mentioned their status as being unable to work or wanting to contribute to extra household funds despite not having current employment. Those who scored the highest on compensation as a reason for taking surveys under indexed on higher household incomes.

Some users mentioned that they prefer being compensated for their data rather than companies tracking their data from

browsing history. They are happy to provide their data in hopes that their ads would be more targeted towards them or that companies could use less of their browsing data and more of what they have explicitly decided to tell companies.

Being Heard and Having Impact

The second largest theme included respondents wanting to be heard, share opinions, have impact on products, and support research. This is perhaps unsurprising given that the nature of surveys is often to solicit feedback, get opinions of respondents, and use the data to inform decisions made by companies or to further research.

Most of the panels were heavily used for either marketing or political surveying. In these panels, the category for wanting to make an impact was higher. Users cited wanting to share opinions in order to *"make a difference in helping to shape and develop new advance future products and services."* or to give *"input that may impact something for the better."* Others called out specific groups such as marketers: *"I like knowing my opinions and reviews help marketers making products that benefits consumers."* Other users discussed the joy they feel when they see products or services offered in the market for which they previously completed a survey and felt a sense of pride or accomplishment that they helped inform decisions. AMT workers and the employee panel are less frequently used for such market research surveys. These groups were less likely to be motivated by giving an opinion, but rather to support the researchers running the survey.

There was a small percentage of respondents who mentioned wanting to make sure their voice was heard because they identified as part of a marginalized or under-represented group. Although some of these respondents did not mention what group(s) they identified with, age and political affiliation were both mentioned specifically by multiple participants.

Surveys as a Pastime and About Relevant Topics

For many respondents across all panels, taking surveys was enjoyable, a way to pass time, learn new things, or to engage with topics and companies they cared about. Many found taking surveys to be fun or interesting and mentioned liking to see what types of new products, services, or research agendas were trending. Some reported taking surveys to feel like they can learn something, gain new knowledge, or to make them think. In particular, older users often mentioned that surveys help them *"stay sharp"* or keep them cognitively processing. Others reflected on personal experiences and values while filling out surveys. They say that they learn about themselves at the same time as providing feedback in surveys.

Many users also discussed feeling like taking surveys was a better use of their time than video games or mindlessly browsing the internet. *"Being productive"* or *"doing something worthwhile"* made users feel like they were making good use of their time by taking surveys. Others mentioned multitasking while taking surveys to pass commercial breaks or to do more than one thing while watching TV.

Certain topics were important to many respondents. People mentioned picking-and-choosing surveys that were relevant to their interests or about topics which they passionately followed, such as politics. Users cited skipping or not completing surveys that didn't resonate with their interests. The participants in Users 1 and 2 knew their feedback was focused on a certain company's products, and mentioned their fandom and affinity for learning more about what these companies were working on which motivated them to take their surveys.

Minor Additional Reasons

A few minor themes also emerged, including a handful of respondents in each panel who identified themselves as researchers and curious about how others asked questions or knew the importance of reliable data. The "other" category mostly included answers such as *"no reason"*, *"I don't know"* or *"a friend recommend I try to do some surveys so I signed up."* The UXR Panel also included many "other" responses around wanting to qualify for in-person studies which pay more.

LIMITATIONS

There are several limitations to our work. First, we have chosen to limit this analysis to United States panels. Given the complexity in fielding a study across many platforms, and the difficulty in finding international ground truth data for many of our questions, we feel this is a reasonable limitation for this work. Replicating this study in other markets will be interesting future work.

Second, we only fielded one survey on each panel. Specific demographic skews may not be representative of all surveys conducted with that panel, should a different set of people choose to respond. With some of the larger surveys (n=1000) there should be enough variation that this would not be a large concern, but for some of the smaller panels, repeatedly running the survey may lead to differences in the overall composition of the panel based on who responds on a given day.

DISCUSSION

As shown in the previous section, there are many tradeoffs to using particular research panels and platforms. There are many constraints to consider when choosing where to deploy a survey. One of the most significant is often money, with another being time. In an industry setting, at times waiting a week to receive results from a professional panel such as Springboard America might be too long to wait. In the academic world, paying thousands of dollars for a sample might not be an option.

Most interesting to us was that some very low cost and fast panels had error margins that were statistically indistinguishable from the $6,000 Omnibus panel. For example a $500 SurveyMonkey panel of 250 participants answering up to 20 questions (they charge $1/participant per 10 questions), was within 1.5 percentage points of the much more expensive survey for the median question response error. For most surveys, either in industry or academia, this is sufficient as long as one does not plan to run complex cross-tabs where more participants are desired from different demographics.

Speed was also an issue, with Springboard America taking six days to return data after the survey questions were delivered to them, while AMT completed in just under two hours—getting

only 2.6 percentage points more accurate on average for each question when waiting six days and paying significantly more for these results. Many design research questions merely seek to understand if something is ubiquitous in the population (>80%), fairly common (around 50%), or uncommon (< 20%) and 2.6 percentage points of error will not matter for questions like these.

Another interesting finding was the extremely biased samples from our internal Employees panel. It is always tempting to use a convenience sample, and to ask folks around you about topics. However, this led to a median error of 16.4% with a mean error of 20% on each question response. Some questions were off from the ground truth by as much as 63%! The research literature [16, 19, 24, 35] has shown convenience samples are biased, but our work has quantified the bias on a wide range of demographic, behavioral, and lifestyle questions.

RECOMMENDATIONS FOR THE CHI COMMUNITY

Given the benefits and limitations discussed above, we would like to present our recommendations to the CHI community based on this study. We hope that this will be a set of best practices that other researchers can follow when trying to choose the most appropriate survey panel for their needs.

Use Different Panels for Different Topics

Different panels had better or worse representation for different topics. For political opinions, professional panels such as Omnibus and Springboard America performed best, while self-service platforms such as SurveyMonkey and Mechanical Turk had much larger deviations from the US population.

However for capturing a variety of different family types and living arrangements, SurveyMonkey performed better than the Omnibus across all types of living arrangements. Any research where family type is important (e.g. services meant to be used within the home) should leverage platforms that are more representative in this area. For age, panels such as SBA and our User panels significantly underrepresented 18-24 year-olds, while panels such as the Omnibus and SurveyMonkey were very close to the ground truth values.

In summary, researchers should check the tables in this paper to find panels that best match the demographic and behavioral attributes that are important for a given research project and ensure that they pick a panel that represents those users well.

Utilize AMT for Open-Ended Responses

When looking for more in-depth qualitative feedback from survey respondents, open-ended questions can provide valuable data. When comparing the length, ability to follow directions, and provide rich feedback, our data shows that AMT respondents provided the most detail and followed the directions more carefully than other panels. Although other panels had slightly higher relevancy rates, AMT responses were overall more complete, thoughtful, and grammatically correct than other panels. In part, we hypothesize this is because AMT workers know their HITs can be rejected if they are not done correctly and fully. Some data cleaning may still be required, but overall, AMT was the best panel for running more complex and writing-intensive survey questions.

Choose Panels Where you Have Control

Our set of ten panels included a mix of vendors who manage the process of programming, launching, and providing data (e.g. Omnibus, SBA, Users 1, and Users 2) and self-service platforms where researchers control all steps of the process (e.g. Survey Monkey, AMT, UXR Panel, and Employees). Although outsourcing tasks to a vendor can save time for researchers and provides a level of quality control of the survey instrument, it also creates additional points of friction such as having less control over when you can schedule surveys and less visibility into the progression of survey results compared to a self-service panel. These added layers of coordination can add days to a task meant to be streamlined by outsourcing the work in the first place. As such, researchers should be mindful of these pros and cons of working with vendors to launch surveys, especially when they are on a tight deadline or wish to check data as it comes in.

Avoid Convenience Samples

As shown above, while convenience samples are convenient, they often do not capture the broader population. Most HCI researchers who publish at CHI are in prestigious research universities or large companies. Results from surveying this type of audience will lead to a very skewed sample on political beliefs, demographics, and technology use. Academics [1] and technology industry professionals [4] are more likely to be liberal in their views and use technology more than the typical American. This will likely greatly skew the results of many types of HCI research survey deployed to this audience. While free, surveying your peers will not lead to data you can trust.

CONCLUSION

We have performed an analysis of 10 different survey panels and compared their accuracy at capturing demographics and behaviors of the broader United States population on a set of 80 different responses. We have shown that some panels are better than others overall at having low amounts of error, but that on certain topics particular panels are better than others at achieving a representative sample. In addition, panels that might not be as representative are better at eliciting longer open-response text.

This work has clear implications and recommendations for the CHI community. We hope that this will improve the discussion around what is truly a "representative" sample in research studies. We encourage researchers to look at various aspects studied in this paper in their own work and report more fully on various demographic and lifestyle attributes of participants when reporting their work. As we have shown, just because a panel is balanced on age and gender does not mean that it accurately captures a representative set of Americans on other demographic, technology use, and lifestyle dimensions.

ACKNOWLEDGMENTS

We would like to thank Brooke White for the original idea for this research as well as the members of the Yahoo/Verizon Media Research and Accessibility team for their feedback on question coverage, wording, and the panels that we should include.

REFERENCES

[1] Samuel J Abrams. 2016. There are conservative professors. Just not in these states. *The New York Times* (2016).

[2] Judd Antin and Aaron Shaw. 2012. Social Desirability Bias and Self-Reports of Motivation: A Study of Amazon Mechanical Turk in the US and India. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI âĂŹ12)*. Association for Computing Machinery, New York, NY, USA, 2925âĂŞ2934. DOI: http://dx.doi.org/10.1145/2207676.2208699

[3] API. 2019. Paying for news: Why people subscribe and what it says about the future of journalism. (2019). https://www.americanpressinstitute.org/publications/reports/survey-research/paying-for-news/

[4] Perry Bacon Jr. 2018. Can Conservatives Ever Trust A Tech Industry Staffed Mostly By Liberals? *fivethirtyeight* (2018).

[5] Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior research methods* 43, 3 (2011), 800.

[6] Frank Bentley, Nediyana Daskalova, and Nazanin Andalibi. 2017a. If a person is emailing you, it just doesn't make sense: Exploring Changing Consumer Behaviors in Email. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 85–95.

[7] Frank R Bentley, Nediyana Daskalova, and Brooke White. 2017b. Comparing the reliability of Amazon Mechanical Turk and Survey Monkey to traditional market research surveys. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1092–1099.

[8] Robert Blaise, Michael Halloran, and Marc Muchnick. 2018. Mobile commerce competitive advantage: A quantitative study of variables that predict m-commerce purchase intentions. *Journal of Internet Commerce* 17, 2 (2018), 96–114.

[9] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[10] Daniel Burda and Frank Teuteberg. 2013. Towards Understanding an EmployeeâĂŹs Retention Behavior: Antecedents and Implications for E-Mail Governance. (2013).

[11] United States Census Bureau. 2018. Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2018. (2018). http://www2.census.gov/programs-surveys/popest/tables/2010-2018/national/totals/nst-est2018-01.xlsx?#

[12] Jonathan J Cadiz, Gina Venolia, Gavin Jancke, and Anoop Gupta. 2002. Designing and deploying an information awareness interface. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 314–323.

[13] CDC. 2018. 2018 National Health Interview Survey (NHIS). (2018). ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2018/personsx_freq.pdf

[14] census.gov. 2010. Table 1188. Usage of General Purpose Credit Cards by Families: 1995 to 200. (2010). https://www2.census.gov/library/publications/2010/compendia/statab/130ed/tables/11s1188.pdf

[15] ComScore. 2019. ComScore Internet Population - July 2019. (2019). http://mymetrix.comscore.com

[16] Sunny Consolvo, Frank R Bentley, Eric B Hekler, and Sayali S Phatak. 2017. Mobile user research: A practical guide. *Synthesis Lectures on Mobile and Pervasive Computing* 9, 1 (2017), i–195.

[17] Tilman Dingler and Martin Pielot. 2015. I'll be there for you: Quantifying Attentiveness towards Mobile Messaging. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 1–5.

[18] eMarketer. 2019. Devices Owned by US Internet Users, Feb 2019. (2019). https://www.emarketer.com/chart/228605/devices-owned-by-us-internet-users-feb-2019-of-respondents

[19] Ilker Etikan, Sulaiman Abubakar Musa, and Rukayya Sunusi Alkassim. 2016. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics* 5, 1 (2016), 1–4.

[20] FatWallet. 2015. 2015 Travel Study: 8 in 10 Americans Take Vacations. (2015). https://www.prnewswire.com/news-releases/2015-travel-study--8-in-10-americans-take-vacations-300065251.html

[21] Adrienne Porter Felt, Serge Egelman, and David Wagner. 2012. I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns. In *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*. ACM, 33–44.

[22] Leah Findlater, Joan Zhang, Jon E Froehlich, and Karyn Moffatt. 2017. Differences in crowdsourced vs. lab-based mobile and desktop input performance data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6813–6824.

[23] fivethirtyeight.com. 2019. How Popular is Donald Trump? (2019). https://projects.fivethirtyeight.com/trump-approval-ratings/?ex_cid=rrpromo

[24] Ronald D Fricker. 2008. Sampling methods for web and e-mail surveys. *The SAGE handbook of online research methods* (2008), 195–216.

[25] Richard Fry. 2013. 2013: Where are the Animal Companions. (2013). https://www.census.gov/programs-surveys/ahs/visualizations/where-are-the-animal-companions-.html

[26] Richard Fry. 2017. The share of Americans living without a partner has increased, especially among young adults. (2017). https://www.pewresearch.org/fact-tank/2017/10/11/the-share-of-americans-living-without-a-partner-has-increased-especially-among-young-adults/

[27] Huiqing Fu, Yulong Yang, Nileema Shingte, Janne Lindqvist, and Marco Gruteser. 2014. A field study of run-time location access disclosures on android smartphones. *Proc. USEC* 14 (2014).

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

[29] Samantha Jaroszewski, Danielle Lottridge, Oliver L Haimson, and Katie Quehl. 2018. Genderfluid or attack helicopter: Responsible HCI research practice with non-binary gender variation in online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 307.

[30] Tendayi Kapfidze. 2018. U.S. Mortgage Market Statistics: 2018. (2018). https://www.magnifymoney.com/blog/mortgage/u-s-mortgage-market-statistics-2018/

[31] Matthew Kay, Dan Morris, Julie A Kientz, and others. 2013. There's no such thing as gaining a pound: Reconsidering the bathroom scale user interface. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 401–410.

[32] Dan Kopf. 2019. Old people in the US are watching a lot more TV. (2019). https://qz.com/1563911/who-watches-the-most-tv/

[33] Richard N Landers and Tara S Behrend. 2015. An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology* 8, 2 (2015), 142–164.

[34] Brooke Fisher Liu, Michele M Wood, Michael Egnoto, Hamilton Bean, Jeannette Sutton, Dennis Mileti, and Stephanie Madden. 2017. Is a picture worth a thousand words? The effects of maps and warning messages on how publics respond to disaster information. *Public Relations Review* 43, 3 (2017), 493–506.

[35] Joanne Locascio, Rushil Khurana, Yan He, and Jofish Kaye. 2016. Utilizing employees as usability participants: exploring when and when not to leverage your coworkers. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 4533–4537.

[36] Niall McCarthy. 2019. The Rise In U.S. Auto Loan Debt Shows No Signs Of Slowing Down. (2019). https://www.forbes.com/sites/niallmccarthy/2019/01/03/the-number-of-americans-holding-auto-loan-debt-shows-no-signs-of-slowing-down-infographic/#1b3c33607ffe

[37] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. *Survey Research in HCI*. Springer New York, New York, NY, 229–266. DOI: http://dx.doi.org/10.1007/978-1-4939-0378-8_10

[38] Nielsen. 2019. Super Bowl LIII Draws 98.2 Million TV Viewers, 32.3 Million Social Media Interactions. (2019). https://www.nielsen.com/us/en/insights/article/2019/super-bowl-liii-draws-98-2-million-tv-viewers-32-3-million-social-media-interactions/

[39] NPR. 2019. NPR Report: Smart Speakers See 78% Increase YOY. (2019). https://www.npr.org/about-npr/682946406/npr-report-smart-speakers-see-78-increase-yoy

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

[41] Andrew Perrin. 2018. One-in-five Americans now listen to audiobooks. (2018). https://www.pewresearch.org/fact-tank/2018/03/08/nearly-one-in-five-americans-now-listen-to-audiobooks/

[42] Zuzanna Pieniak, Wim Verbeke, Federico Perez-Cueto, Karen Brunsø, and Stefaan De Henauw. 2008. Fish consumption and its motives in households with versus without self-reported medical history of CVD: A consumer survey from five European countries. *BMC Public Health* 8, 1 (2008), 306.

[43] Paulo Rita, Ana Brochado, and Lyublena Dimova. 2019. MillennialsâĂŹ travel motivations and desired activities within destinations: A comparative study of the US and the UK. *Current Issues in Tourism* 22, 16 (2019), 2034–2050.

[44] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.

[45] Nicolas Roulin. 2015. Don't throw the baby out with the bathwater: Comparing data quality of crowdsourcing, online panels, and student samples. *Industrial and Organizational Psychology* 8, 2 (2015), 190–196.

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

[47] Jessica AR Williams and Selena E Ortiz. 2017. Examining public knowledge and preferences for adult preventive services coverage. *PloS one* 12, 12 (2017), e0189661.