

# Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample

Jill D. Weinberg<sup>a,b</sup>, Jeremy Freese<sup>a</sup>, David McElhattan<sup>a</sup>


a) Northwestern University; b) American Bar Foundation

**Abstract:** Compared to older kinds of sample surveys, online platforms provide a fast and low-cost platform for factorial surveys, as well as a more demographically diverse alternative to student samples. Two distinct strategies have emerged for recruitment: using panels based on population-based samples versus recruiting people actively seeking to complete online tasks for money. The latter is much cheaper but prompts various concerns about data quality and generalizability. We compare results of three vignette experiments conducted using the leading online panel that uses a population-based paradigm (Knowledge Networks, now GfK) and the leading platform for crowdsourcing recruitment (Amazon Mechanical Turk). Our data show that, while demographic differences exist, most notably in age, the actual results of our experiments are very similar, especially once these demographic differences have been taken into account. Indicators of data quality were actually slightly better among the crowdsourcing subjects. Although more evidence is plainly needed, our results support the accumulating evidence for the promise of crowdsourcing recruitment for online experiments, including factorial surveys.

**Keywords:** factorial survey; Amazon Mechanical Turk; crowdsourcing, vignette experiment; online experiment; knowledge networks

**Editor(s):** Jesper Sørensen, Delia Baldassarri; **Received:** April 23, 2014; **Accepted:** May 22, 2014; **Published:** August 4, 2014

**Citation:** Weinberg, Jill D., Jeremy Freese, and David McElhattan 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsourcing-Recruited Sample." *Sociological Science* 1: 292-310. DOI: 10.15195/v1.a19

**Copyright:** © 2014 Weinberg, Freese, and McElhattan. This open-access article has been published and distributed under a Creative Commons Attribution License, which allows unrestricted use, distribution and reproduction, in any form, as long as the original author and source have been credited. 

FACTORIAL surveys, a method devised by Peter Rossi, are an effective tool to study attitudes, opinions, and normative judgments because they combine the internal validity that random assignment affords to experiments with the external validity that population sampling affords (e.g., Rossi 1979; Rossi and Berk, 1982; Rossi and Nock 1992; Mutz 2011). For many years, factorial surveys on larger, more diverse populations were underutilized because they were time-consuming and expensive to field. Online survey platforms have been a boon in this respect. Even though online survey platforms based on population sampling have very low cumulative response rates, comparisons to stronger benchmarks have so far been promising (Baker et al. 2010). Yet even while online panels are much cheaper and permit more flexible access to population-based samples

than telephone or face-to-face surveys, the cost is still high enough to limit their use.

More recently, a different approach to online experiments has arisen in which subjects are recruited from "crowdsourcing" labor sites where individuals seek to complete small tasks in exchange for compensation. Anyone with Internet access can sign up, and, in practice, these platforms provide a much more diverse sample in terms of self-reported demographic characteristics than traditional undergraduate subject pools. Prevailing wage rates on these sites are so low as to make them far cheaper than population-based online surveys—indeed, often even cheaper than experiments using undergraduates (Horton, Rand, and Zeckhauser 2011).

Crowdsourcing platforms provide new challenges to the conventional rationale for population-based

experiments. A *diverse* sample is not the same as a *representative* sample, especially from the perspective of the typical focus of survey research on estimating population parameters. For that matter, a convenience sample that represents a population in terms of observed characteristics may be nonrepresentative in other ways that compromise the generalizability of experimental results. Among other concerns, considering how cheaply crowdsourcing platform participants will work, there is reason to wonder whether they are “weird” subjects who act atypically in experiments. Initial results in this respect have been auspicious (Berinsky, Huber, and Lenz 2012; Horton, Rand, and Zeckhauser 2011; Mullinix, Druckman, and Freese n.d.), but the research involves stimuli that are shorter and tasks that focus much less on interpretation than is characteristic of many sociology experiments.

In this study, we compare results from three vignette-style factorial surveys conducted in parallel, using leading population-based and crowdsourcing survey platforms. The population-based experiments were conducted using the web panel recruited by Knowledge Networks (hereafter KN; the company is now known as GfK Custom research).<sup>1</sup> The crowdsourcing-based experiments were conducted using Amazon Mechanical Turk (hereafter MT), the dominant crowdsourcing site and platform analyzed in previous studies (e.g., Berinsky, Huber, and Lenz 2010; Buhrmester, Kwang, and Gosling 2011; Chandler, Mueller, and Paolacci 2014; Iperiotis 2010; Shapiro, Chandler and Mueller 2013; Mullinix, Druckman, and Freese n.d.).

If crowdsourcing platforms yield little difference in demographic representation and data quality from population-based Internet panel experiments, the justification for the additional expense of population-based Internet panels for factorial surveys is weakened. If differences exist but are accounted for by demographic differences between the samples, then the justification for population-based Internet panels for factorial surveys may also be considered weaker because these differences could in principle be accounted for by reweighting. The strongest case for the extra ex-

pense of population-based Internet panels would be either if the data quality is superior to that of a crowdsourcing sample or if there are divergent results that cannot be accounted for by observable variables.

We consider four questions. First, do the two samples differ in terms of basic sociodemographic characteristics? Second, do they differ in terms of data quality? Third, do they differ in terms of the outcome variables from the experiments, and, if so, can any differences be accounted for by socioeconomic differences in the samples? Finally, do the actual findings from the experiments differ between the two samples, and, if there are differences, can these be explained by differences in observed characteristics? Our findings overall are quite congenial to crowdsourcing sampling for experimental research, although, of course, a broader evidence base is needed before stronger conclusions can be drawn.

## Background

The unique leverage of random assignment to test causal hypotheses is broadly appreciated (e.g., Shadish, Cook, and Campbell 2002). Nevertheless, experiments are often considered hopelessly artificial compared to actual situations of behavior and choice (e.g., Webster and Sell 2007:51). Behavioral science experiments also often rely entirely upon undergraduate subjects (Sears 1986). The rhetorical disadvantage of findings based exclusively on responses from college students in presenting results to non-experimentalist audiences is considerable, especially regarding the issues relevant to social policy for which sociologists often turn to experimentation.

Population-based survey experiments combine the strong internal validity of random assignment with the external validity afforded from the greater population representativeness that surveys can attain (e.g., Mutz 2011). For sociologists, vignette studies have been particularly appealing because they allow responses to concrete, if hypothetical, cases while having the same standardization advantages as surveys (e.g., Rossi and Nock 1982). Vignette experiments present respondents with a short scenario in which certain aspects vary randomly. For example, in one

<sup>1</sup>We refer to it as KN here both because it is accurate in terms of our time of fielding and because, despite the brand change, KN still seems more widely recognized as the name of the platform.

of our studies, we present respondents a scenario of sexualized behavior in the workplace and ask questions concerning whether the scenario constitutes sexual harassment. Different vignettes vary randomly both the gender of the perpetrator and the victim. While not field experiments, vignettes compare favorably in terms of realism to standard survey items, and, to the extent the purpose of the study is less obvious, reduce potential problems from social desirability biases (Sniderman and Piazza 2002).

Vignette-based studies have been particularly helped by the rise of online surveys, because vignettes are a poor fit for telephone surveys, unless kept extremely short. Meanwhile, printed vignettes, whether used in person or via mail, can introduce issues with tracking randomizations; they also often use only a limited number of all the vignettes that would be possible if conditions were assigned independently, which can introduce confounding across conditions (Atzmüller and Steiner 2010). Conducting vignette experiments online does amplify the potential problem of respondents possibly not reading vignettes carefully enough. Among other techniques, researchers routinely embed “comprehension checks” in surveys that directly test respondents about the content of material they previously read or “catch trials” that ask questions for which the answer given by an attentive reader is different from what might be provided by someone who is skimming (Downs et al. 2010; Paolacci et al. 2010).

Vignette studies can be conducted an order of magnitude more cheaply using a population-based Internet survey platform than through a major face-to-face survey like the General Social Survey (GSS). Researchers also receive data far more quickly, going from items to data in a matter of weeks, as opposed to the two or more years between developing questions and receiving data from GSS. In turn, crowdsourcing recruitment reduces costs by yet another order of magnitude, and one receives data in days (or hours) instead of weeks. However, while the resulting subjects are more diverse than a study of undergraduates, the crowdsourcing platform makes no pretense of offering researchers a representative sample.

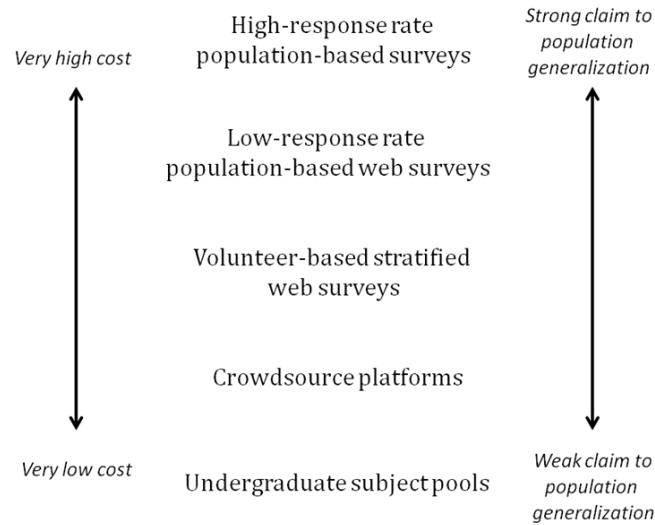
In terms of the tradeoff between the defensibility of our estimates and the cost-feasibility of implementation, Figure 1 depicts a hierarchy

for a platform like GSS (highest defensibility, highest cost) to a convenience sample of undergraduates (lowest defensibility, low cost). In this study, we seek to compare results from parallel studies employing the second and fourth entries of this hierarchy.<sup>2</sup> We use KN for our Internet survey panel using population sampling, and their platform has been found more accurate than volunteer-based web panels in estimating univariate characteristics (the third level in Figure 1; see Baker et al. 2010).

The conventional justification for population-based sampling for surveys is that it permits unbiased estimation of population parameters. For survey experiments, differences between randomly assigned groups can be interpreted as average causal effects in the population and effect size differences between population subgroups can be compared. At the same time, a nonrepresentative subject pool for an experiment does not imply the results would be different if one had conducted the experiment with a broader population. Indeed, we would expect results to differ systematically only if treatment effects differed for the same variables for which the participant pool was a nonrepresentative sample (see, e.g., Druckman and Kam 2011). For example, undergraduate pools may provide bad population estimates of a treatment effect strongly moderated by age, but better estimates of a treatment effect solely moderated by gender.

If a sample of subjects is diverse without being representative, then differences in treatment effects among members of the population due to moderators can be accounted for if those moderators are observed. For example, MT respondents are considerably younger than the general population. Provided we have enough age variation in the MT sample, differences caused by the MT sample not being age-representative could be addressed, at least in principle, by reweighting. By contrast, if MT respondents are unusual in unmeasured respects that also moderate the treatment effects of the experiment, bias introduced

<sup>2</sup>Most web survey platforms occupy the middle of the figure, in which panelists are recruited through various “opt-in” or “volunteer” means that do not involve requesting the participation of individuals sampled from the general population. These platforms are generally cheaper than KN but do not represent anything remotely like the dramatic cost savings that a crowdsourcing platform like Mechanical Turk represents.



**Figure 1:** Hierarchy for Survey Platform by Cost and Generalizability

in population estimates cannot be addressed by reweighting.

The key question is whether our data provide reason to suppose that differences between MT results and KN results are large enough to justify the extra expense of the KN sample. We do not address the separate but important question of how accurately KN results themselves estimate true population parameters, about which we wish there were more evidence (see Baker et al. 2010). Using KN as the basis of comparison reflects the pragmatic reality of the budget possibilities for most factorial survey projects; this should not be mistaken for any claim that KN represents a “gold standard.”

Regarding MT, Berinsky, Huber, and Lenz (2012) have conducted the most comprehensive examination of the comparative properties of MT data to date and report that “the [MT] sample does not perfectly match the demographic and attitudinal characteristics of the U.S. population, but does not present a wildly distorted view of the U.S. population either” (for similar assertions, see Buhrmester, Kwang, and Gosling 2011; Chandler, Mueller, and Paolacci 2014; Ipeirotis 2010; Mason and Suri 2012; Paolacci, Chandler, and Ipeirotis 2010). Specifically, they find their MT sample is more female and better educated than the US population. MT respondents also report lower incomes and are less likely to be

black, married, or homeowners. The most pronounced demographic difference is in terms of age, with MT respondents being younger. On attitudinal measures, the MT sample leans toward Democratic Party affiliation and is more politically interested than respondents in a nationally representative sample.

MT samples have been shown to vary based on the complexity of the task (Kazai, Kamps, and Milic-Frayling 2012), the time of day the sampling occurred (Komarov, Reinecke, and Gajos 2013) and on whether or not workers mention the task on discussion boards (Chandler, Mueller, and Paolacci 2014). Some evidence suggests that the representation of men among the MT pool may be increasing (Mullinix, Druckman, and Freese n.d.). Researchers can use screening questions to quota-sample MT subjects (Chandler, Mueller, and Paolacci 2014), but asking screening questions itself requires paying participants, so this will substantially increase data collection costs. Researchers can also advertise for workers that fulfill specific characteristics (e.g., politically conservative Christians), but abundant concerns have been raised that many workers may lie in order to earn money by participating.

For the simple item-wording and framing experiments they consider, Berinsky, Huber, and Lenz (2012) found similar results comparing MT to nationally representative samples. These ex-

periments involved simpler stimuli and, for the most part, very robust effects. While encouraging, this leaves open the question of whether randomized vignette studies would yield a similarly promising comparison. Compared to the studies in Berinsky, Huber, and Lenz (2012), vignette experiments demand more engaged reading, focus on more complex interpretations of stimuli, and also involve multiple independent conditions manipulated within a single vignette. For that matter, sociologists are often motivated to use vignette experiments and other factorial surveys specifically because of hypotheses about moderators, for which we might expect the implications of nonrepresentative sampling to be strongest. In broad terms, then, the published base of comparisons between MT and other samples is encouraging but largely excludes the sort of experiments in which nonrepresentative samples cause sociologists and other social policy researchers particular concern.

Several threats to validity in crowdsourcing experiments merit discussion. First, while issues of respondent identity can be raised for any online study (Reips 2000), MT investigators are explicitly prohibited (by Amazon) from obtaining identifying information of subjects (Berinsky, Huber, and Lenz 2012; Horton, Rand, and Zeckhauser 2011; Mason and Suri 2012). Different individuals can share the same account, complete different sections of a survey, work together on tasks, or provide false demographic information. Although Amazon provides no check on what MT workers report to researchers, MT workers' self-reported demographics are, for the most part, consistent across studies (Mason and Suri 2012), with 81 to 98 percent agreement on a range of characteristics (Rand 2012).

Second, existing work has evaluated data quality, which includes comprehension and dropout rates (Berinsky, Huber, and Lenz 2012; Buhrmester, et al. 2011; Paolacci et al. 2010; Reips 2000). Payments to MT workers must be approved by requesters, and workers can be excluded from jobs if contractors give them low scores. Paolacci et al. (2010) tested comprehension across participants in MT, a traditional lab setting, and an online forum. They found no significant differences in attentiveness among the three settings although Turkers scored the best with respect to catch trials.

Third, one might worry about possible subject interaction about the experiment, which has long been a concern in undergraduate experiments that involve deception (Horton, Rand, and Zeckhauser 2011). Online discussion boards allow MT workers to discuss the content of tasks, but actual mentions of experimental treatments are relatively infrequent, and researchers can easily monitor the forums (Mason and Suri 2012; Reips 2000). Chandler (2011) finds that only 28 percent of workers follow forums and blogs about MT, and 13 percent of users reported seeing the content of experiments discussed in these settings. Given this evidence and the relatively non-provocative character of vignettes, we do not view this as a likely source of concern for our study but note that researchers have suggested ways this possibility may be further addressed (Chandler, Mueller, and Paolacci 2014).

Finally, MT participants need to be approached as *de facto* "professional respondents." MT samples contain many individuals who have done *hundreds* of other MT studies (Chandler, Mueller and Paolacci 2014). Studies that use standardized knowledge questions or provide performance tasks with a learning curve should not be conducted with the premise of naive subjects. Vignette studies are potentially a good fit for MT in this respect, as they do not involve deception, knowledge questions, or any specific presupposition that subjects have not participated in many surveys before. Even so, the notion of "professional respondents" may raise concerns about how unusual the people who select into these studies are and how this may affect generalizability of results. With KN, even though relatively few of the sampled persons agree to participate in the panel, assenting to a personalized request may still be a far less selective behavior than seeking out opportunities to do surveys online for money.

## Data

Data come from three vignette experiments about employment discrimination fielded as one instrument on both KN and MT. By agreement with KN, we were able to use their platform to administer the survey to subjects recruited via MT. In other words, the screen presentation and other interface aspects of the study were exactly the

same across the two panels, so such differences cannot account for any observed differences in results between the two recruited populations.

The experiments are all between-subjects, mixed-factorial designs intended to examine how ordinary people define *employment discrimination*.<sup>3</sup> They concerned race discrimination, sexual harassment, and reasonable accommodation. We are writing separate substantive papers on each of the experiments, in which we describe the hypotheses in more detail and give more elaborate consideration of the results. For the purposes of this article, we focus on comparing the parallel fieldings of the instrument on KN and MT, and we describe the experiments only briefly and only when germane to the specific purpose of this study. We do note, however, that part of the motivation of the experiments is that these vignettes have also been fielded on a sample of state and federal trial judges. Thus, the question of trying to estimate accurately the average treatment effect for the population in these experiments is of particular interest for our purposes because doing so will provide the baseline for our comparison to the estimated treatment effect for our judge sample.

In the race discrimination experiment, we describe a situation in which an employee asks for a pay raise and is denied, but an employee with similar qualifications and of different race receives a raise. We varied the employees' races, the workplace environment, employee tenure, and employee performance. In the sexual harassment experiment, we describe an employee who has an encounter with another employee and the alleged victim quits shortly afterwards. We vary the gender of the perpetrator and victim, the organizational status of the perpetrator, the conduct in question, and company policies about sexual harassment. Finally, in the reasonable accommodation experiment, we describe a full-time, female

<sup>3</sup>In total, respondents received eight vignettes presenting various workplace disputes. The vignettes featured different company descriptions to avoid learning and exhaustion effects (Auspurg, Hinz and Liebig 2009). Vignettes for the three experiments were interspersed so that consecutive vignettes were parts of different experiments. Within each vignette, vignette conditions were assigned fully independently of one another, so that all possible vignettes had an equal probability of administration. Assignments to each condition were also balanced within persons, to maximize the diversity of conditions presented across vignettes to each respondent.

employee who seeks a workplace accommodation and is denied. We vary the job of the employee, employee identity, and the cost of the accommodation request. After each vignette, respondents were asked several related questions about the scenario, which we describe briefly later. Table A1 in the online supplement presents a more detailed description of the experimental conditions.

## Platform Selection

We selected market leaders among Internet-based platforms that use population-based or crowdsource recruitment samples: Knowledge Networks (KN) and Amazon Mechanical Turk (MT), respectively. KN has been central among Internet-based platforms and was the first to use a population-based approach. Virtually all of the Internet-based examples in Mutz's (2011) *Population-Based Survey Experiments* use KN data. MT has emerged as the dominant crowdsourcing recruitment platform for researchers. It is cost effective and fielding data is extremely fast—taking a few hours to recruit more than 1000 respondents.

## Population-Based Sample (Knowledge Networks, now GfK Custom Research)

KN, now GfK, maintains a panel of 40,000 U.S. households for its surveys. KN has switched from using random-digit dialing for recruitment to address-based methods for better handling cell phoneonly households. The panel at the time of our study comprised individuals recruited through both frames. KN provides sampled respondents who do not have a computer or Internet access at the time of recruitment in which case KN provided these as part of their participation. KN estimates that their sampling methods provide 97 percent coverage, meaning that 97 percent of the intended population falls within the contact methods of recruitment.

From this panel, KN draws a subsample based on a client's specifications for a particular study. Responses to one's study are then combined with information from a core profile completed by panelists earlier, along with items from other previously administered profile surveys that can be added for a surcharge. About 16 percent of those

originally sampled by KN panel were successfully recruited to the panel, and of those, 64.0 percent had completed the core demographic data necessary to be eligible for our study.

For our study, KN selected 4,990 people from its panel. All were non-institutionalized U.S. citizens over the age of 18. Selected individuals receive an e-mail notifying them of a new survey available for them to take, including a link to the survey questionnaire. After three days, a second e-mail reminder is sent out to all nonrespondents. While the field period for KN studies varies, ours was five days.

In our factorial survey, 2,665 members were “screened”—that is, of those sampled, 58.3 percent clicked on the link to proceed with the survey. After reading the informed consent prompt, 2,222 members consented to participate. Ultimately, we received 2,087 completed interviews from the web-enabled panel, implying a 93.9 percent completion rate for those who started the survey, and a 54.7 percent completion rate for those asked to participate.

While none of the aforementioned rates are unusual for contemporary survey research of this type, the combination (.160 x .640 x .583 x .939) indicates a cumulative response rate of 5.6 percent. This figure is not at all low by contemporary polling standards but certainly is for people whose reference point might be the General Social Survey or American National Election Studies.<sup>4</sup> Again, KN has had positive descriptive results attributed to its use of population sampling (Baker et al. 2010), and KN sampling has a more plausible case for better representing the population than MT, if estimates differ. Nevertheless, as said earlier, one should not confuse this advantage with KN results being a gold standard of population description. Also, within sociology, many studies using KN present only the completion rate (e.g., Phelan, Link, and Feldman 2013; Regnerus 2012); readers should be aware

<sup>4</sup>Low response rates are the way of the world in contemporary survey research for work that can be done by individual investigators. Even large-scale, high-response platforms are experiencing a steadily declining response rate (e.g., the ANES 2012 response rate was 49 percent). If a high response rate such as GSS' is the only credible way to field surveys, few researchers would have the budget to do credible research and even more studies would be disallowed from being “good”

that this figure is not a response rate in any population-based sense.

For most investigators wanting to conduct their own factorial surveys, KN likely represents the upper tier of available funding. Adding factorial surveys of comparable length to those reported here to a flagship national study would require considerably more investment, if it were logistically possible at all. Overall pricing of a KN study is based on sample size, number of profile variables, time to complete the survey, and programming complexity. Our survey took about 15 minutes to complete, and KN pricing included an initial pretest of 25 respondents and the addition of two more profile variables to their standard data delivery. The total cost was roughly \$50,000.

### Crowdsource sample (Mechanical Turk)

Amazon.com (2014) reports 500,000 MT worker accounts. In 2011, we advertised the survey on Amazon as taking about one hour and paid respondents \$3 for participation. This sum is actually relatively high by MT standards; the median hourly wage for MT tasks has been estimated at \$1.38 (Horton and Chilton 2010; also Paolacci et al. 2010). We restricted the survey to MT workers classified as 18 or older and living in the United States. We also excluded individuals with approval ratings below 95 percent on previous tasks (which MT calls *Human Intelligence Tasks* or HITs). We posted our HIT to be available for seven days or until we reached our desired number of respondents to complete the task. We obtained our desired number of respondents in less than four days. We recruited 1,349 participants.

We called the HIT “Assessing Workplace Disputes” and described the task as completing an online survey. While the description was vague, we provided keywords to describe the HIT as *website*, *survey*, and *workplace disputes*. When workers opened our HIT, we instructed them to click on the survey URL link, which directed them to the same platform used for the KN study. We also instructed workers that, at the end of the survey, they would receive a code. To be paid, they had to enter this code on the MT webpage.

Amazon MT collects a 10 percent commission on top of the amount the requester pays subjects to complete the HIT. Overall, we paid \$4,500,

roughly 10 percent of the cost to run the same survey using KN.

## Results

### Demographic characteristics of participants

Table 1 shows demographic characteristics of participants in each population. Both weighted and unweighted descriptive statistics for KN are presented. KN weights include post-stratification adjustments based on the Current Population Survey for the corresponding period. Included among the variables used for this weighting are gender, age, race/ethnicity, education, and region, so the weighted KN numbers for these variables presumably directly reflect population distributions.<sup>5</sup>

Overall, our MT and unweighted KN samples differ significantly from one another on all variables except for region. Consistent with prior findings, our MT respondents were younger, more educated, and more likely to be female than their KN counterparts. Only 16.7 percent of MT respondents were over 45 years old, compared to 64.0 percent of the KN sample. As can be seen by looking at the weighted MT results (based on the Current Population Survey), the differences in age distribution between the MT and the unweighted KN samples reflect both an overrepresentation of younger people in the MT sample and an overrepresentation of older adults in the KN sample. Overrepresentation of older adults in population-based surveys is historically common (Tuckel and O'Neill 1995).

Regarding other characteristics, two thirds of the KN sample had at least some college (29.5 percent some college; 37.8 percent college degree or higher), whereas nearly 90 percent of MT participants had at least some (41.3 percent some college; 47.1 percent college degree or higher). More than three in five MT respondents were female. MT respondents were more likely to have never been married (and were not cohabiting), a difference that persists even when the younger age of MT respondents is taken into account.

<sup>5</sup>Note that unweighted characteristics of the KN sample do not necessarily reflect characteristics of the KN panel because KN does do some stratified sampling of its panel.

The KN sample is fairly similar to our MT sample in terms of self-reported race/ethnicity. The major difference between the samples is that MT had nearly three times as many respondents who identified as “other, non-Hispanic” (2.5 percent in KN versus 7.3 percent in MT). This selection presumably reflects a larger number of respondents of East or South Asian descent among MT respondents (Mason and Suri 2012).<sup>6</sup> Because of the greater representation of “other” respondents in the MT sample, KN had more white non-Hispanic (77.8 percent versus 76.1 percent), black non-Hispanic (8.3 percent versus 6.2 percent), and Hispanic respondents (8.7 percent versus 7.4 percent).

### Data quality

Our experiment provided various ways of determining whether respondents provided “quality” data in terms of the care with which they engaged with the stimulus materials and items. Several of these measures provided reason to suppose that problem responses were more common among KN respondents than among MT respondents

First, after having answered questions about a vignette, subjects were given comprehension check items that asked about non-manipulation-related content of the vignette. An example of such a question would be asking about the type of business conducted by the company described in the vignette. All respondents received seven such questions, and we defined problem cases as those in which respondents missed more than one of the manipulation check items. Nearly twice as many KN respondents as MT respondents were problem cases by this criterion (9 percent versus 5.0 percent,  $p < .001$ ). This difference is partly, but not entirely, accounted for by the higher education and higher proportion of females among MT respondents ( $p < .001$ , after controls).

Second, we examined differences in the time respondents took to complete the surveys. The comparability of samples here is limited because KN respondents were able to start and finish the instrument in different sessions, potentially days apart, whereas MT respondents timed out if they did not complete the survey promptly. Conse-

<sup>6</sup>Although we restricted our experiment to respondents in the United States, Amazon MT compensates workers in either U.S. dollars or Indian rupees.



**Table 1:** Comparison of demographic characteristics of Knowledge Networks and Mechanical Turk samples

	Weighted KN	Unweighted KN	MT
Female	51.6	49.5	61.3
Age			
18-30	22.8	14.5	50.9
31-45	24.6	21.6	32.4
46-60	28.2	33.4	14.5
61-95	24.4	30.6	2.2
Race			
White, non-Hispanic	72.8	77.8	76.1
Black, non-Hispanic	11.6	8.3	6.2
Hispanic	4.3	8.7	7.4
Other, non-Hispanic	10.1	2.5	7.3
2+ races reported	1.2	2.7	2.9
Education			
Less than high school	9.8	5.5	1.1
High school	31.8	27.3	10.5
Some college	29.7	29.5	41.3
College degree or higher	28.7	37.8	47.1
Marital status			
Married/living with partner	59.8	67.6	50.8
Never married	22.2	16.8	39.8
Divorced/separated	13.0	11.4	8.8
Widowed	5.1	4.2	0.7
Region			
Northeast	18.4	18.0	18.9
Midwest	22.9	23.5	24.1
South	36.5	34.3	35.2
West	22.2	24.2	21.9
Household head	79.0	84.0	69.4
Household income in thousands	60.2	69.0	49.2
SD		(43.2)	(33.9)
N		2,087	1,349

*Note:* Chi-square test of differences between unweighted KN and MT samples significant for all variables ( $p < .001$ ), except region.

quently, the maximum duration in the MT sample was 3.5 hours, whereas the maximum duration in the KN sample was slightly more than four days. Overall, 13.9 percent of KN respondents took over 45 minutes to complete the survey, compared to 1.3 percent of MT respondents.

We can more fairly compare samples regarding respondents who appear to have finished the study too quickly. Respondent performance on the comprehension check items was weak for respondents who took six minutes or less (68 per-

cent missed two or more comprehension items), modestly affected for respondents taking between seven and ten minutes (12 percent), and not associated with performance thereafter (6 percent). KN respondents were more than three times more likely than MT respondents to finish the survey in six minutes or less (2.4 percent versus 0.7 percent,  $p < .001$ ). These times do not account for the difference in comprehension check performance, however, because KN respondents were still nearly twice as likely to miss more than one

**Table 2:** Means and standard deviations of key outcome variables of each experiment

	Sexual harassment	Racial discrimination	Reasonable accommodation
Knowledge Networks	3.60 (1.99)	5.21 (1.58)	4.16 (1.89)
Mechanical Turk	3.95 (1.97)	5.15 (1.62)	4.02 (1.93)

*Note:* Results for 7-point item asking whether in the respondent's opinion, the scenario in the vignette constitutes (1) sexual harassment, (2) racial discrimination, (3) a reasonable denial of an accommodation request by an employee.

comprehension check item even when respondents with short completion times were excluded (7.0 percent versus 3.8 percent  $p < .001$ ).

Third, we considered item nonresponse as an indicator of response quality. We classified respondents as skipping items if they were nonrespondents for some but not all of the outcome measures. Not surprisingly, skipping items was more common among respondents who finished the survey in six minutes or less (20.0 percent versus 2.3 percent  $p < .001$ ). Even discarding responses from respondents with unusual completion times, however, KN respondents were still more likely to have skipped some of our key outcome items than were MT respondents (2.8 percent versus 1.0 percent,  $p < .01$ ).

Fourth, we considered lack of variation in responses. Our survey included a scale of confidence in the legal system, which has six items in which two are reverse coded. Respondents were coded as problematic if they gave the same response to at least five of the six items. KN participants were more likely to be classified as problematic by this criterion (7.6 percent versus 5.4 percent,  $p < .05$ ), although the difference was no longer significant when the sample was restricted to respondents who took six minutes or less to complete the survey ( $p = .07$ ).

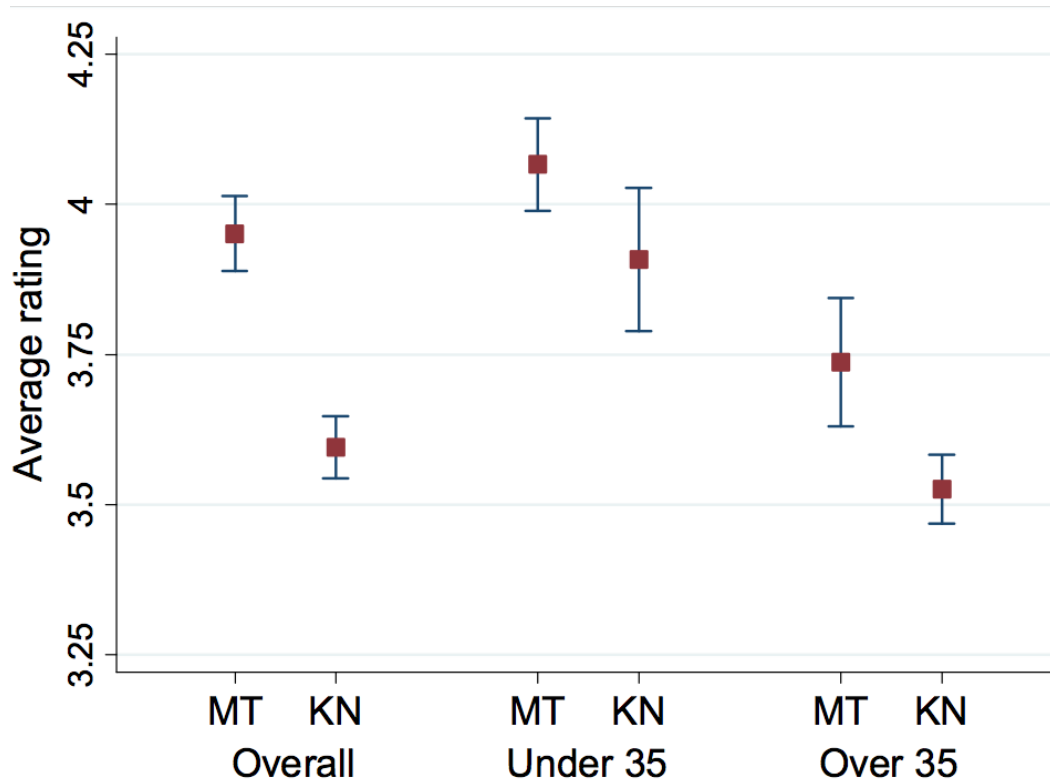
Although neither study had large problems with suspect respondents according to any of the previous criteria, the findings consistently indicate the idea that there are more problems among KN respondents than among MT respondents. If the true denominator of cost per case is considered to be only those cases providing quality data, this difference would suggest even greater value to MT relative to KN. However, the

argument here could be reversed. Given that one complaint about MT is that it has a large number of effectively "professional respondents" who may diverge from an ordinary diverse population, then fewer problem cases in the MT sample than in the KN sample could be an indicator of the nonrepresentativeness of the MT sample. Berinsky, Huber, and Lenz (2012) asked MT participants how many other MT surveys they took, and they found no difference in the effect sizes of their experiments between those who reported taking five or more surveys in a month and those who took fewer.

### Distribution of dependent variable

As noted, the vignettes posed scenarios concerning potential racial discrimination, sexual harassment, and failure to provide a reasonable workplace accommodation. After each vignette, respondents were asked four items. The conclusions of our study do not depend on which item(s) are considered and for simplicity we focus on the questions asking whether, in their opinion, the scenario constituted (depending on the vignette) racial discrimination, sexual harassment, or an unreasonable denial of workplace accommodation.

In terms of the univariate distributions of these measures, no significant differences were found between the KN and MT samples on any of the items for either the racial discrimination or the reasonable accommodation vignettes (Table 2). In contrast, all the outcome measures were significantly different between the two samples for the sexual harassment vignette, with MT respondents more congenial to interpreting the



**Figure 2:** Differences between MT and KN respondents in responses to sexual harassment vignette, both overall and by age.

scenario as harassment. If we consider the “in your opinion” item and average over the three harassment vignettes presented to each respondent, the mean response on the seven-point scale for MT respondents was 3.95, compared to 3.60 for KN respondents ( $p < .001$ ).

The reason for observing differences for sexual harassment and not for racial discrimination and reasonable accommodation can be largely explained by the relationship between these vignettes and the age of respondents. Age was not associated with response to the racial discrimination vignette but was moderately associated with the reasonable accommodation vignette in the direction of older respondents being more supportive of the discrimination interpretation. In response to the sexual harassment vignette, however, the mean rating by respondents 30 years old or younger was 4.05, compared to 3.41 for respondents over age 60 ( $p < .001$ ). Figure 2 shows differences between KN and MT for the sexual

harassment vignette, both overall and separating respondents by whether they were over or under age 35.

If we use an OLS regression model controlling for age, the difference between the unweighted KN and MT samples is reduced from .34 points to .15, or by about 57 percent (see Table 3).<sup>7</sup> Adding further controls for gender, education, race/ethnicity, marital status, and (logged) income only reduces the difference from .15 to .12, although the difference does remain statistically significant even with these controls.<sup>8</sup>

In sum, the only differences in the univariate distributions of the outcome variables of our experiment seem to be largely explained by the

<sup>7</sup>Ordered logistic regression gives similar results, but changes in coefficients across models are more difficult to interpret than in OLS regression.

<sup>8</sup>The difference does become nonsignificant if we estimate a model with fixed effects for each year of age ( $b = .09$ ,  $p = .08$ ), although one may then worry about overfitting.

**Table 3:** OLS coefficients for regression of average response to sexual harassment vignette by sample, age, and selected covariates.

	(1)	(2)	(3)
MT sample	0.343 <sup>†</sup> (0.045)	0.147 <sup>†</sup> (0.052)	0.119* (0.054)
Age (in decades)		-0.112 <sup>†</sup> (0.015)	-0.115 <sup>†</sup> (0.017)
Female			0.191 <sup>†</sup> (0.044)
Less < high school			0.007 (0.133)
Some college			0.100 (0.063)
College degree			0.044 (0.062)
Black			0.298 <sup>†</sup> (0.085)
Latino			0.226 <sup>†</sup> (0.084)
Other race/ethnicity			-0.033 (0.086)
Never married			-0.074 (0.057)
Divorced/separated			-0.034 (0.075)
Widowed			0.118 (0.136)
ln(Family income)			0.032 (0.029)
Intercept	3.597 <sup>†</sup> (0.028)	4.173 <sup>†</sup> (0.083)	4.368 <sup>†</sup> (0.327)
N	3107	3107	3107

*Note:* Standard errors in parentheses.

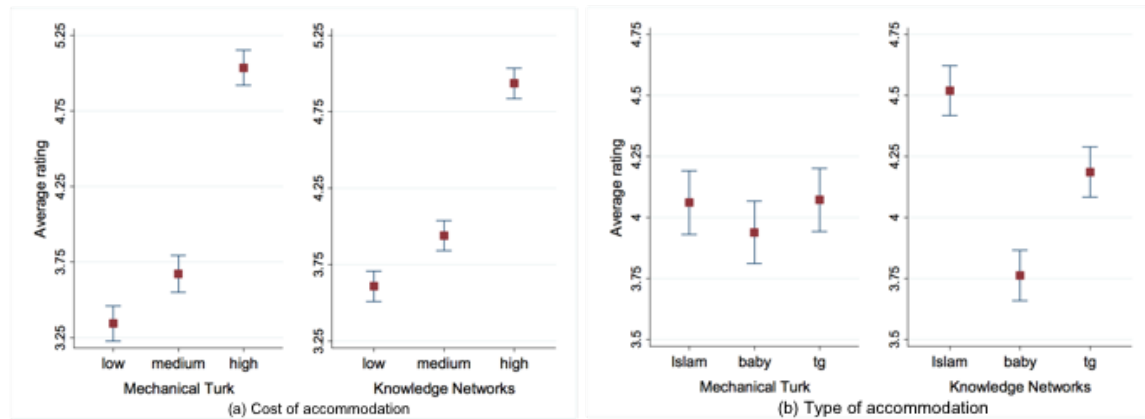
\*  $p < 0.05$ , <sup>†</sup>  $p < 0.01$

differences in the age distributions of the MT and KN samples. Unobservable differences between MT and KN respondents may have also contributed to the difference in response to the sexual harassment vignette, but if so, the combined effect of these differences was still smaller than the consequences of the differences in the age distributions. Although we showed earlier that MT and KN populations differ in other respects, notably gender, educational status, and marital status, these had no more than a negligible consequence for univariate distributions in the experiments here. Of course, they may be

more important in other studies. Overall, our results are quite promising for the comparison of unweighted KN and MT univariate distributions once differences in readily measured demographic characteristics are taken into account.

## Comparing study results

This article focuses on comparing results from the KN and MT platforms for the three vignette experiments conducted; substantive discussion of these three experiments is presented elsewhere (Weinberg, Nielsen, and Freese 2012; Weinberg



**Figure 3:** Differences in results of reasonable accommodation experiment between MT and KN.

n.d.). We focus here on the “in your opinion” items that were administered in parallel for each vignette. Data were analyzed using hierarchical linear regression in Stata 12.2 to account for the within- and between-subjects comparisons.

In the sexual harassment vignette, we observed above that MT respondents, on average, were more favorable to a harassment interpretation of vignettes than KN respondents, but that this difference was largely accounted for by the difference in the two samples’ age distributions. Regardless, a univariate difference between two samples does not indicate a difference in estimates from experiments on those samples.

In the case of the sexual harassment vignette, such difference is precisely what we observed. Vignettes varied on four different dimensions (the genders of the parties, the organizational relationship between the parties, the type of behavior described, and the strictness of policies). On all four dimensions, there were significant differences across at least two of the conditions. However, in none of the four cases did we observe any significant difference in the omnibus test for whether differences between conditions varied between the two samples. In sum, the experiment yielded positive findings for all its conditions, and these were similar in magnitude across both platforms.

In the reasonable accommodation vignette, we did not observe any difference in the univariate distributions of the outcome between samples. However, with respect to the experimental results, we observed two differences between samples for the three conditions of the experiment.

These differences are shown in Figure 3. First, vignettes differed in how costly it would be for the business to implement the accommodation. We hypothesized respondents would be less likely to perceive discrimination as accommodations became more costly. The hypothesized relationship was observed in both samples but was stronger in magnitude among the KN respondents (a difference of 1.3 points between extreme conditions for KN respondents and 1.7 for MT respondents  $p < .001$ ; see Figure 3a). While we explore this difference further below, we note that the substantive implications for interpreting the study results seem minor.

More consequential is the observed difference according to type of accommodation requested (Figure 3b). The three types are having converted to Islam, having had a baby, and undergoing gender transition. In the KN sample, responses to the three conditions were significantly different from one another, but there were no significant differences in the MT sample. In other words, the same design would have yielded positive, significant findings in the KN sample and null findings in the MT sample.

In both cases, the culprit turns out to be the difference in age distribution between the two samples. Age is a significant moderator of responses concerning both the effect of the cost of the accommodation and the type of accommodation. Specifically, older respondents were more responsive to the distinctions between vignettes. When we estimate a hierarchical model that includes the three-way interaction between

**Table 4:** Hierarchical random-effects OLS models of the interaction of experimental conditions by sample type and age for reasonable accommodation vignette experiment.

	Cost condition		Type of accommodation condition	
	(1)	(2)	(3)	(4)
MT sample	-0.267 <sup>†</sup> (0.078)	-0.271 (0.253)	-0.458 <sup>†</sup> (0.084)	0.008 (0.274)
Low cost condition	0.333 <sup>†</sup> (0.067)	0.365 (0.226)		
High cost condition	1.327 <sup>†</sup> (0.068)	1.962 <sup>†</sup> (0.230)		
MT × Low	-0.003 (0.106)	0.252 (0.339)		
MT × High	0.366 <sup>†</sup> (0.105)	0.130 (0.341)		
Age × Low		-0.007 (0.042)		
Age × High		-0.124 <sup>†</sup> (0.043)		
Baby condition			-0.757 <sup>†</sup> (0.073)	0.295 (0.244)
Transgender condition			-0.333 <sup>†</sup> (0.073)	0.712 <sup>†</sup> (0.248)
MT × Baby			0.635 <sup>†</sup> (0.114)	-0.148 (0.370)
MT × Trans			0.344 <sup>†</sup> (0.115)	-0.309 (0.372)
Age × Baby				-0.205 <sup>†</sup> (0.045)
Age × Trans				-0.203 <sup>†</sup> (0.046)
Age (in decades)		0.113 <sup>†</sup> (0.032)		0.206 <sup>†</sup> (0.033)
Age × MT		0.059 (0.061)		-0.030 (0.068)
Age × MT × Low		-0.075 (0.083)		
Age × MT × High		0.007 (0.083)		
Age × MT × Baby				0.126 (0.091)
Age × MT × Trans				0.086 (0.091)
Intercept	3.609 <sup>†</sup> (0.050)	3.030 <sup>†</sup> (0.171)	4.519 <sup>†</sup> (0.053)	3.461 <sup>†</sup> (0.179)
N	6292	6290	6292	6290

Note: Standard errors in parentheses.

\*  $p < 0.05$  †  $p < 0.01$

age, platform, and experimental conditions, the interactions between age and experimental conditions are significant, and neither the two-way interactions within platform nor any of the three-way interactions are significant (Table 4). That is, there are no remaining significant differences in responses between the KN and MT platforms once age is accounted for, and the magnitude of the moderating effect of age is not significantly different between KN and MT.

In the racial discrimination vignette, all three dimensions produced significant differences in both the KN and MT samples, and only in one case were these differences significantly different in magnitude from one another. This difference is shown in Figure 4. In KN, changing the victim in the vignette from a white employee to a black employee increased the perception of discrimination by .21 points on the seven-point scale. In MT, this difference was .44 points, or more than twice as large ( $p < .001$ ).

Unlike the reasonable accommodation experiment, age is not a significant moderator of the effect of race manipulation in this experiment, and age does not explain the effect size difference between the two samples. None of the demographic variables do, although introducing indicator variables for educational attainment reduces the difference to marginal significance ( $p = .05$ ). On the basis of the information at hand, then, we have reason to think that the difference in the effect of employee race on perceived discrimination cannot be accounted for by observable variables. Although a difference in the magnitude of a positive finding is not as dramatic as a difference in the direction of results, this is, nevertheless, the kind of difference that could be used to justify a more expensive data collection platform over a cheaper one.

Even here, though, it is quite plausible that the failure of observed variables to account for differences between KN and MT simply reflects some sharp limitations in our available measures. In particular, we were not able to collect political ideology from the MT sample because it is not part of KN's core profile variables. Profile variables can be added for additional time and cost, however. We were able to obtain political ideology using other variables in the dataset. As noted, other research indicates that MT respondents may be more liberal than the population at

large (Berinsky et al. 2012). If so, then political differences might readily account for most or all of the difference in the effect of employee race.

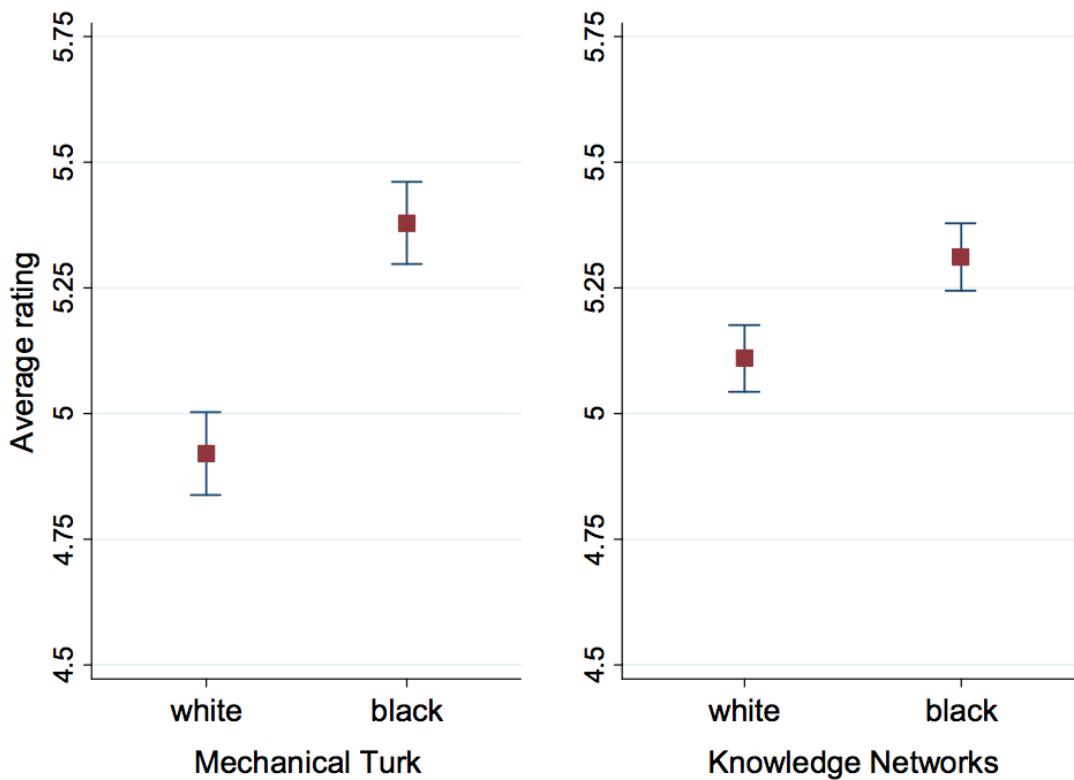
## Conclusion

Outside psychology, research designs based on random assignment in the social sciences have long been inhibited by the logistical and financial difficulties of studying people away from one's campus. This concern may be especially resonant for sociologists, who unlike economists and political scientists have not seen a dramatic increase in the use of experimental methods in recent years (Jackson and Cox 2013). Factorial surveys were invented to reach broad populations, and the method has flourished as the Internet has drastically reduced the costs of recruiting diverse and distant subjects.

The banner of "online experiments" now flies over vastly different methods of recruiting participants, and this study has focused on two extremes. One, led by KN, uses established population-sampling techniques and seeks to provide representative samples of prospective participants. The other, led by MT, involves researchers venturing into an online labor market and, with few restrictions, accepting whatever people select their study from a roster of available paid tasks.

In terms of representing basic descriptive characteristics of the general population, KN unsurprisingly does a much better job than MT. KN respondents are still somewhat disproportionately highly educated, female, and white, but MT is more divergent from the general population on all these variables and is strikingly nonrepresentative of the population in terms of age. Crowdsourcing platforms will surely not be displacing population surveys in the basic task of describing populations any time soon.

Our purpose here, however, is to assess the platforms' comparability over a set of sociological vignette experiments. We find that MT experiments produce potentially better data than do the KN experiments in the narrow sense of having fewer problem respondents. Of course, given that population surveys are intended to measure whole populations accurately and real populations include inattentive and incompetent people, precisely what constitutes the ideal level of prob-



**Figure 4:** Differences in results for racial discrimination experiment between MT and KN. “White” and “black” refer to the race of the employee in the vignette.

lem responses in a factorial survey is debatable. The higher response quality from MT could be because of a number of factors, including the cultivation of “professional respondents;” heightened attention caused by the explicitly transactional nature of the exchange given the online labor market setting; or platform recruitment of younger, more educated respondents with high familiarity with the Internet.

In comparing the results from the parallel experiments, our overall conclusions are quite promising for crowdsource-based recruitment. This adds encouragement to the possibility of its use in factorial surveys on a mainstream sociological topic to what has been a favorable run of comparisons of results on other kinds of experiments between MT and more established platforms (see also Berinsky, Huber, and Lenz 2012; Horton, Rand, and Zeckhauser 2011; Mullinix, Druckman, and Freese n.d.). Our three vignette

experiments encompassed ten different experimental conditions. In only three conditions were there significantly different effect sizes between the MT and KN platforms. Two of these were largely accounted for by age difference in respondents, and it is plausible the remaining difference might be explained by readily measurable differences in political ideology. For the most part, then, we would have observed substantively the same results in our experiments had we used MT instead of KN, and most of the remaining differences could have been addressed by reweighting the samples to match the known population age distribution.

Experimental studies typically recommend careful theoretical justification of moderators examined in analyses. This remains sage advice for the internal validity and replication prospects of such studies, as loose consideration of moderators can lead to a proliferation of false positives due to



underdocumented multiple testing. Nevertheless, when considering the possible generalizability of findings in a diverse but nonrepresentative sample like those produced by MT, it seems advisable to encourage testing of moderation on broad characteristics known to diverge markedly between the sample and population, even when not explicitly theorized in advance. Respondent age, political orientation, and gender appear to be the most obvious suspects in this regard for MT.

In sum, our results further substantiate the optimism for using crowdsourcing platforms to conduct factorial surveys, even for studies motivated by interest in estimating average treatment effects for the US population. Of course, caution is warranted, and more comparisons need to be made across more substantive topics and more types of experiments, as well as a broader range of platforms. However, our results suggest participants recruited through MT are not so unusual that their behavior noticeably diverges from that of KN respondents, especially once differences in basic measurable characteristics are taken into account. Of course, similarities in results between MT and KN could mean that *both* groups are unusual relative to the population at large, a gloomy prospect that remains possible given the low response rates for online surveys in general. Work comparing these platforms to samples with high response rates needs to be done, but this is hindered not just by the much greater expense of the latter but also by differences in the modes by which most such surveys are presently conducted.

A separate possibility is that, regardless of what evidence ultimately indicates about the comparability of results between crowdsourcing-based samples and other samples, problems with the recruitment method's face validity may prove an insuperable rhetorical disadvantage for factorial surveys directed at social policy, not unlike how policy-minded experiments based entirely on undergraduates are currently received by some audiences. The terrain of population research is changing quickly, but we expect that both population- and crowdsourcing-based methods of recruitment have long futures ahead. The challenge for researchers is to determine the most cost effective and valid ways of combining evidence from both.

## References

- Alexander, Cheryl S. and Henry Jay Becker. 1978. "The Use of Vignettes in Survey Research." *Public Opinion Quarterly* 42(1):93-104. <http://dx.doi.org/10.1086/268432>
- Amazon Mechanical Turk. 2014. "Service Summary." Seattle, Washington. Retrieved April 20, 2014 <https://requester.mturk.com/tour>.
- Atzmüller, Christiane; Steiner, Peter M. 2010. "Experimental Vignette Studies in Survey Research." *European Journal of Research Methods for the Behavioral and Social Sciences* 6(3):128-38. <http://dx.doi.org/10.1027/1614-2241/a000014>
- Auspurg, Katrin, Thomas Hinz, and Stefan Liebig. 2009. "Complexity, Learning Effects, and the Plausibility of Vignettes in Factorial Surveys." Paper presented at the American Sociological Association Annual Meeting, August 7-11, San Francisco, California.
- Baker, Reg, Stephen J. Blumberg, J. Michael Brick, Mick P. Couper, Melanie Courtright, J. Michael Dennis, Don Dillman, Martin R. Frankel, Philip Garland, Robert M. Grovers, Courtney Kennedy, Jon Krosnick, and Paul J. Lavrakas. 2010. "Research Synthesis: AAPOR Report on Online Panels." *Public Opinion Quarterly* 74(4):711-81. <http://dx.doi.org/10.1093/poq/nfq048>
- Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3): 351-68. <http://dx.doi.org/10.1093/pan/mpr057>
- Buhrmester, Michael, Tracey 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): 3-5. <http://dx.doi.org/10.1177/1745691610393980>
- Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. "Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions For Behavioral Researchers." *Behavior Research Methods* 46(1):112-30. <http://dx.doi.org/10.3758/s13428-013-0365-7>

- Downs, Julie S., Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. 2010. "Are Your Participants Gaming the System? Screening Mechanical Turk Workers." In *Proceedings of the 28th International Conference on Human Factors in Computing Systems* (2399-2402). New York: ACM.
- Druckman, James D. and Cindy D. Kam. 2011. "Students as Experimental Participants: A Defense for the 'Narrow Data Base.'" Pp. 41-57 in *Cambridge Handbook of Experimental Political Science*, edited by James D. Druckman, Donald P. Green, James H. Kuklinski and Arthur Lupia. New York: Cambridge University Press.
- Gätcher, Simon. 2010. "(Dis)advantages of Student Subjects: What is Your Research Question." *Behavior and Brain Sciences* 33(2-3): 92-93. <http://dx.doi.org/10.1017/S0140525X10000099>
- Goodman, Joseph K., Cynthia E. Cryder, and Amar Cheema. 2012. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making*.
- Gosling, Samuel D., Simine Vazie, Sanjay Srivastava, and Oliver P. John. 2004. "Should We Trust Web-Based Studies?" *American Psychologist* 59(2): 93-104. <http://dx.doi.org/10.1037/0003-066X.59.2.93>
- Horton, John J. and Lydia B. Chilton. 2010. "The Labor Economics of Paid Crowdsourcing." *Proceedings of the 11th ACM Conference on Electronic Commerce 2010*. Retrieved July 18, 2014 from [http://ssrn.com/abstract=\\$1596874](http://ssrn.com/abstract=$1596874).
- Horton, John J., David G. Rand and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14(3):399-425. <http://dx.doi.org/10.1007/s10683-011-9273-9>
- Ipeirotis, Panos. 2010. "Demographics of Mechanical Turk. (CeDER Working Paper-10-01). New York University. Retrieved July 18, 2014 from <http://hdl.handle.net/2451/29585>.
- Jackson, Michelle and D. R. Cox. 2013. "The Principles of Experimental Design and Their Application to Sociology." *Annual Review of Sociology* 39: 27-49. <http://dx.doi.org/10.1146/annurev-soc-071811-145443>
- Kazai, Gabriella, Jaap Kamps, and Natasa Milic-Frayling. 2012. "Worker Types and Personality Traits of Crowdsourcing Relevance Labels" in *Proceedings of 20th International Conference on Information and Knowledge Management (CIKM)*. ACM: New York.
- Komarov, Steven, Katharina Reinecke, and Krzysztof Z. Gajos. 2013 "Crowdsourcing Performance Evaluations of User Interfaces" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM: New York.
- Lynch Jr., John G. 1982. "On the External Validity of Experiments in Consumer Research." *Journal of Consumer Research* 9(3):225-39. <http://dx.doi.org/10.1086/208919>
- Mason, Winter and Siddarth Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44(1): 1-23. <http://dx.doi.org/10.3758/s13428-011-0124-6>
- Mullinix, Kevin J., James N. Druckman, and Jeremy Freese. "When Convenience Samples Yield Generalizable Estimates in Social Science Experiments." Unpublished manuscript.
- Mutz, Diana C. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton University Press.
- Nelson, Thomas E., Rosalee A. Clawson, and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effects on Tolerance." *American Political Science Review* 91(3):567-83. <http://dx.doi.org/10.2307/2952075>
- Paolacci, Gabriele, Jesse Chandler, and Panos Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5(5) 411-19.
- Peterson, Robert A. 2001. "On the Use of College Students in Social Science Research: Insights From a Second Order Meta-Analysis." *Journal of Consumer Research* 28 (3): 450-61. <http://dx.doi.org/10.1086/323732>
- Phelan, Jo C., Bruce G. Link, and Naumi M. Feldman. 2013. "The Genomic Revolution

- and Beliefs about Essential Racial Differences: A Backdoor to Eugenics?" *American Sociological Review* 78(2): 167-91. <http://dx.doi.org/10.1177/0003122413476034>
- Rand, David G. 2012. "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments." *Journal of Theoretical Biology* 299(21): 172-79. <http://dx.doi.org/10.1016/j.jtbi.2011.03.004>
- Regnerus, Mark. 2012. "How different are the adult children of parents who have same-sex relationships? Findings from the New Family Structures Study." *Social Science Research*. 41(4): 752-70. <http://dx.doi.org/10.1016/j.ssresearch.2012.03.009>
- Reips, Ulf-Dietrich. 2002. "Standards for Internet-Based Experimenting." *Experimental Psychology* 49(4): 243-256.
- . 2000. "The Web Experiment Method: Advantages, Disadvantages, and Solutions." *Psychological Experiments on the Internet*. Ed. Birnbaum, Michael H. San Diego: Academic Press. 89-118.
- Rossi, Peter H. and Steven L. Nock. eds. 1982. *Measuring Social Judgments: The Factorial Survey Approach*. Beverly Hills, CA: SAGE Publications.
- Sears, David O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51(3): 515-30. <http://dx.doi.org/10.1037/0022-3514.51.3.515>
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
- Shapiro, Danielle N., Jesse Chandler, and Pam A. Mueller. 2013. "Using Mechanical Turk to Study Clinical Populations." *Clinical Psychological Science* 1(2):213-20. <http://dx.doi.org/10.1177/2167702612469015>
- Sniderman, Paul M. and Thomas Piazza. 2002. *Black Pride and Black Prejudice*. Princeton, NJ: Princeton University Press.
- Tuckel, Peter and Harry O'Neill. 1995. "A Profile of Telephone Answering Machine Owners and Screeners" *Proceedings of the Section on Survey Research Methods of the American Statistical Association*. Retrieved July 18, 2014 from [http://www.amstat.org/sections/srms/proceedings/papers/1995\\_201.pdf](http://www.amstat.org/sections/srms/proceedings/papers/1995_201.pdf).
- Webster, Murray and Jane Sell. 2007. *Laboratory Experiments in the Social Sciences*. London: Academic Press.
- Weinberg, Jill D. n.d. "Identity and Social Acceptability: Public Perceptions of Reasonable Accommodation." Manuscript in Progress (on file with author).
- Weinberg, Jill D., Laura Beth Nielsen, and Jeremy Freese 2012. "Rub My Shoulders But Don't Send Me Emails: Public Perceptions of Sexual Harassment." Paper presented at Annual Meeting of the Law and Society Association, June 5-8, 2012, Honolulu, HI.
- Zelditch Jr, Morris. 1969. "Can You Really Study an Army in the Laboratory." Pp. 528-39 in *A Sociological Reader on Complex Organizations* Edited by Amitai Etzioni. Austin, Texas: Holt, Rinehart, and Winston.

**Acknowledgements:** This research was funded by the American Bar Foundation. The authors would like to thank participants of the Applied Quantitative Methods Work at Northwestern University who provided wonderful feedback and encouragement. Direct correspondence to Jill D. Weinberg, Northwestern University Department of Sociology, 1810 Chicago Avenue, Evanston, Illinois, 60208.

**Jill D. Weinberg:** Northwestern University; American Bar Foundation. E-mail: [jweinberg@abfn.org](mailto:jweinberg@abfn.org).

**Jeremy Freese:** Northwestern University. E-mail: [jfreese@northwestern.edu](mailto:jfreese@northwestern.edu).

**David McElhattan:** Northwestern University. E-mail: [DavidMcElhattan2017@northwestern.edu](mailto:DavidMcElhattan2017@northwestern.edu).