



Are Relational Inferences from Crowdsourced and Opt-in Samples Generalizable? Comparing Criminal Justice Attitudes in the GSS and Five Online Samples

Andrew J. Thompson¹ · Justin T. Pickett¹

Published online: 13 November 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

Abstract

Objectives Similar to researchers in other disciplines, criminologists increasingly are using online crowdsourcing and opt-in panels for sampling, because of their low cost and convenience. However, online non-probability samples’ “fitness for use” will depend on the inference type and outcome variables of interest. Many studies use these samples to analyze relationships between variables. We explain how selection bias—when selection is a collider variable—and effect heterogeneity may undermine, respectively, the internal and external validity of relational inferences from crowdsourced and opt-in samples. We then examine whether such samples yield generalizable inferences about the correlates of criminal justice attitudes specifically.

Methods We compare multivariate regression results from five online non-probability samples drawn either from Amazon Mechanical Turk or an opt-in panel to those from the General Social Survey (GSS). The online samples include more than 4500 respondents nationally and four outcome variables measuring criminal justice attitudes. We estimate identical models for the online non-probability and GSS samples.

Results Regression coefficients in the online samples are normally in the same direction as the GSS coefficients, especially when they are statistically significant, but they differ considerably in magnitude; more than half (54%) fall outside the GSS’s 95% confidence interval.

Conclusions Online non-probability samples appear useful for estimating the direction but not the magnitude of relationships between variables, at least absent effective model-based adjustments. However, adjusting only for demographics, either through weighting or statistical control, is insufficient. We recommend that researchers conduct both a provisional generalizability check and a model-specification test before using these samples to make relational inferences.

Keywords Web survey · Selection bias · Collider variable · Amazon Mechanical Turk · Opt-in panel

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10940-019-09436-7>) contains supplementary material, which is available to authorized users.

✉ Andrew J. Thompson
ajthompson@albany.edu

Extended author information available on the last page of the article

Introduction

Nonprobability samples recruited online via crowdsourcing and opt-in panels have begun to appear in leading criminology journals (e.g., Enns and Ramirez 2018; Denver et al. 2017; Dum et al. 2017; Gottlieb 2017; Pickett et al. 2013; Vaughan et al. 2019). These samples tend to overrepresent certain population groups—whites, liberals, females, the college educated, and the young (Levay et al. 2016; Ross et al. 2010; Weinberg et al. 2014). Depending on how these selection variables are related to the specific outcome variables of interest (Elwert and Winship 2014; Morgan and Winship 2015), this may undermine the internal and/or external validity of findings (Pasek 2016). Both selection bias and effect heterogeneity are threats to the generalizability of relational inferences in online non-probability samples (Mercer et al. 2017). Many of the variables that impact online selection, like race and political ideology (Keeter et al. 2015; Tourangeau et al. 2013), are also key predictors of criminal justice attitudes (Brown and Socia 2017; Unnever et al. 2008; Silver and Silver 2017). There is thus a need to assess whether these samples yield generalizable inferences for the types of outcome variables that are the common focus in the discipline—namely, public attitudes toward crime and criminal justice.

Towards this end, the current study compares multivariate findings from the General Social Survey (GSS), which uses a national probability sample, with those obtained from five surveys using two common types of online non-probability sampling: crowdsourcing and recruitment from opt-in panels. The five online non-probability samples include more than 4500 respondents. The same measures are available in the online non-probability samples and in the GSS, and all of the samples are national in scope, providing a unique opportunity for comparison. The evidence indicates that online non-probability samples yield multivariate relationships that are similar in direction but not in magnitude to those in the GSS.

Online Sampling: Methodological Issues and Potential Sources of Bias

Two forms of online non-probability sampling—crowdsourcing through websites like Amazon Mechanical Turk (MTurk) and recruiting panelists from opt-in panels—have emerged as common sources of survey data for academic research over the past decade (Baker et al. 2010; Sheehan and Pittman 2016). Crowdsourcing a survey on MTurk involves posting it on the website where members (or “workers”), who self-recruited to the platform, take it for money. At any time, there are thousands of jobs (or “human intelligence tasks” [HITs]) available for workers and they can see a list and description of those for which they are qualified. Surveys are one type of HIT, but there are many others (e.g., transcribing audio, tagging images, and editing text). Workers choose which HITs to complete, and may base their decision on the characteristics of the HIT or requestor (researcher) (Chandler and Shapiro 2016; Sheehan and Pittman 2016). Opt-in panels are different in that vendors recruit the panelists from other websites, profile them when they join the panel, and then invite them later to specific surveys commissioned by researchers (Brandon et al. 2013; Rivers 2007). However, both types of sampling have the same benefits: they are fast, convenient, and cheap (Callegaro et al. 2015).¹

¹ There may also be fewer errors of observation in online web surveys because of the elimination of interviewer effects, less potential for social desirability bias, and higher quality responding (Chang and Krosnick 2009; Weinberg et al. 2014; Yeager et al. 2011). However, issues such as respondent nonnaïveté may lead

The main drawback to using either type of online non-probability sample is potential bias in estimates due to errors of non-observation at every stage of the sampling process: frame construction, sample selection, and survey response (Callegaro et al. 2015; Tourangeau, et al. 2013). It starts with undercoverage in the frame construction stage. Elliott and Valliant (2017) distinguish between potential and actual (or realized) coverage in online surveys. Potential coverage is all Internet users, which represents incomplete coverage of the total population because about 10% Americans do not use the Internet (Keeter et al. 2015). This 10% of the population is systematically different than the other 90%—it is older, poorer, less educated, less white, more rural, and more Southern (Keeter et al. 2015; Tourangeau et al. 2013). Internet users may also diverge from non-users on attitudes and substantive issues, and these differences may persist after demographic discrepancies are taken into account (Mercer et al. 2018).

What creates greater potential for bias is that it is impossible to randomly sample Internet users. A database of all email addresses does not exist, nor does every Internet user have an email address (Callegaro et al. 2015; Tourangeau et al. 2013). This means that the actual (or realized) coverage includes only that small percentage of Internet users who are members of the crowdsourcing website or opt-in panel that serves as the actual sampling frame for the survey (Elliott and Valliant 2017; Valliant and Dever 2011). For example, the effective size of the MTurk sampling frame (worker pool) appears to be about 7300 workers, based on findings from a large-scale, capture-recapture analysis (Stewart et al. 2017). Individuals with internet access are not equally likely to join a crowdsourcing platform or opt-in panel, or, after joining, to respond to a solicitation to complete a survey. Platform or panel membership may, for instance, depend on familiarity and experience with the internet (Nicolaas et al. 2014). Additionally, many opt-in panelists are members of multiple panels—so-called “panel overlap” (Baker et al. 2013). Compared to the population of interest (U.S. adults), then, the population actually covered (e.g., current MTurk workers or panel members) is likely to differ even more than the population potentially covered (Internet users) (Elliott and Valliant 2017).

The next stage of the sampling process entails selecting respondents from the sampling frame, and in online nonprobability surveys this is normally not done randomly. In crowdsourced surveys on MTurk the workers decide which HITs to complete, which means that the resultant samples are “nonprobability samples of the MTurk population as a whole” (Chandler and Shapiro 2016, p. 65). The payment amount, the timing of the survey, the survey topic, the stated length of the questionnaire, and the requestor’s reputation—some workers follow certain requestors—all may influence self-selection into HITs, as may other factors (Casey et al. 2017; Sheehan and Pittman 2016). There is more control over sampling with opt-in panels, because vendors assign panelists to surveys (Brandon et al. 2013; Callegaro et al. 2015). Assignment may be random, but often it is not, at least not completely. Most vendors use routers to dynamically assign panelists to surveys, meaning that they send out generic email invitations to panelists on a regular basis and then wait until the panelists click an included web link to assign them to an open survey for which they are qualified or are needed (in the case of quotas and surveys targeting particular groups). As a consequence, “the sample for any one survey depends on what other surveys are in the field simultaneously” (Mercer et al. 2017, p. 262).

Therefore, both self-selection (in crowdsourced samples) and non-completely-random assignment (in panel-based samples) may introduce additional bias into estimates,

Footnote 1 (continued)

to unique types of errors of observation that are especially problematic for these surveys (Chandler et al. 2014).

beyond any resulting from undercoverage (Chandler and Shapiro 2016; Mercer et al. 2017). Another potential source of bias in online surveys is low response rates (Horton et al. 2011; Zhou and Fishbach 2016). The available evidence suggests that online surveys yield response rates about 11 percentage points lower on average than offline surveys (Tourangeau et al. 2013). Additionally, one form of nonresponse “is particularly pronounced in web surveys”—with meta-analyses reporting median rates “as high as 16–34%”—and is especially likely to stem from questionnaire content: breakoff (Peytchev 2011, p. 34). All forms of nonresponse may introduce bias into estimates if those who decline or drop out of a survey are different than those who complete it. Most evidence suggests that response rates are weakly related to nonresponse bias (Groves et al. 2009). However, the situation may differ in online surveys where a larger proportion of nonresponse is due to breakoffs, which are more closely connected to survey content (Peytchev 2009, 2011). Indeed, Zhou and Fishbach (2016) found that selective attrition (breakoff) was a confounding influence in MTurk experiments, and led to several, sometimes nonsensical, erroneous conclusions.

Inference Type and Fitness for Use

Under certain circumstances, differences between those included and excluded at any stage of the sampling process—frame construction, sample selection, survey response—will bias estimates (Mercer et al. 2017), and the total bias will be a function of the bias at each stage (Groves et al. 2009). The circumstances in which differences translate into bias are the same at each sampling stage. They do, however, differ depending on the inference type (Gelman 2007; Winship and Radbill 1994), such that relational inferences (e.g., regression coefficients) may be more robust than univariate estimates (e.g., means, proportions) to sample quality (Blair et al. 2013; Pasek 2016). Consequently, online non-probability samples may have greater “fitness for use” when the research goal is to analyze relationships between variables (Baker et al. 2013). Before considering the connection between sample-population differences and bias by inference type, it is useful first to layout a general framework for thinking about bias in survey estimates.

The framework’s foundation is Rubin’s (1974) potential-outcome model, which defines a causal effect as the difference in average outcomes under alternative conditions where selection into the conditions is ignorable (independent of the outcome) (Morgan and Winship 2015). Building on this model, Mercer et al. (2017) suggest thinking of probability surveys as experiments where surveyors take measurements only on the treatment group, and selection into the sample is the treatment. If selection is ignorable, the treatment has no effect, allowing for generalization to the unobserved control group (un-sampled members of the sampling frame). The use of random sampling with a frame that has complete (or nearly complete) coverage allows for design-based inference (inference based on probability theory), with model-based adjustments only for nonresponse. The challenge with online non-probability samples is that they are quasi-experiments, due to undercoverage and nonrandom sampling, so inference from them hinges entirely on modeling assumptions and adjustments (Elliott and Valliant 2017).

Two main concerns for model-based inference are exchangeability and positivity (Mercer et al. 2017). Sample members (the treatment group) are exchangeable with unsampled members of the population (the control group) when selection and the outcome variable are unrelated. This might occur at the outset, or after conditioning on a set of covariates; control group members are “missing completely at random” in the first situation, and are “missing at random” in the second (Little and Rubin 2002). Conditional exchangeability

means the probability of selection differs *between* covariate cells (varies across groups), but is equal *within* cells (is the same for all members of a given group) (Gelman 2007). Positivity refers to whether there is enough variance in the sample on relevant covariates to condition on them (e.g., by weighting). This is why ensuring sample diversity is of critical importance for researchers using non-probability samples (Blair et al. 2013)—it increases the probability of positivity on the covariates that determine the likelihood of selection into the sample. Notice, however, that if internet usage is a relevant covariate, any online non-probability sample would lack positivity (Mercer et al. 2017).

Univariate

Unweighted estimates of univariate population characteristics (e.g., means, proportions) will be biased when calculated for any variable (outcome) that is related to selection, causally or spuriously (Groves et al. 2009). The relationship is causal when sampling is endogenous, such that the outcome variable (Y) causes selection (S) into the sample directly ($Y \rightarrow S$) or indirectly through a mediator (M) like topic interest ($Y \rightarrow M \rightarrow S$). The relationship is spurious when a third variable (X) causes both the outcome and selection, either directly ($Y \leftarrow X \rightarrow S$) or indirectly (e.g., $Y \leftarrow X \rightarrow M \rightarrow S$). Confounded sampling equates to conditional exchangeability, meaning that if all confounders are known, are measured without error, and have positivity in the sample, weighting on them would eliminate the bias (Gelman 2007; Mercer et al. 2017).

To illustrate, imagine that *only* race (R) determines selection (S) into an online non-probability survey (an unlikely sampling situation, to be sure) measuring public attitudes (A) toward capital punishment, with Whites and Blacks having different selection probabilities and attitudes ($S \leftarrow R \rightarrow A$). Sampled blacks would be exchangeable with unsampled Blacks, and sampled Whites with unsampled Whites, because selection and attitudes would be unrelated *within* these racial groups. If the sample included both Whites and Blacks (positivity), one could weight on race to render selection ignorable (missing at random). Here the unweighted estimate of the proportion of the public favoring capital punishment would be biased, but the weighted estimate would be unbiased.

Multivariate

The existence of selection-outcome relationship does not necessarily bias unweighted relational inferences, as it does unweighted univariate inferences, because in some circumstances regression estimation by itself results in conditional exchangeability (Pfeffermann 1993; Solon et al. 2015). This is one reason why relational inferences may be more robust to sample quality. With confounded sampling ($Y \leftarrow X \rightarrow S$), regressing the outcome (Y) on the confounder (X) in the treatment group (sample) renders selection ignorable and the unobserved control group (unsampled population) missing at random, so that the slope estimate (b_{yx}) is unbiased (assuming X is well-measured, has positive variance, and is specified correctly [linear or nonlinear]) (Gelman 2007; Pfeffermann 1993). By extension, a spurious selection-outcome relationship will not bias the coefficients from a regression of Y on a set of predictors (X_1, X_2, X_3) that includes the confounder(s) responsible for confounded sampling (e.g., $Y \leftarrow X_2 \rightarrow S$) (Gelman 2007; Gelman and Carlin 2002). Another way to put it is that there is “no need for the sample distribution of the X variables to reflect the population distribution ... a correctly specified model will provide consistent

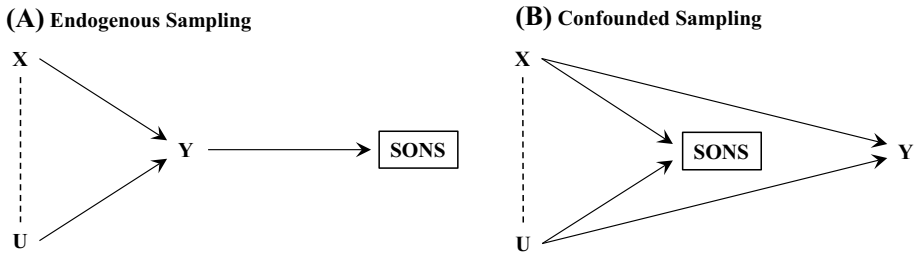


Fig. 1 Examples of when online non-probability sampling would amount to conditioning on a collider variable and bias regression coefficients. *Notes* X, regression predictors; SONS, selection into online non-probability samples; U, unmeasured variables; Y, regression outcome. The box around SONS indicates conditioning through sampling. The dotted line indicates the resultant spurious association

and unbiased parameter estimates regardless of how a sample is drawn with respect to X” (Winship and Radbill 1994, p. 235).²

When does the sampling process bias regression coefficients? The exchangeability condition is met when “sampling probabilities are independent of the error term in the regression equation” (Solon et al. 2015, p. 310), which means there will be bias whenever the sampling process results in a correlation between the predictor variables (X) and the unmeasured causes (U) of the outcome (Y), represented by the error term (Berk 1983; Berk and Ray 1982). This happens when both X and U cause selection, either directly or indirectly, making selection a “collider variable” (Morgan and Winship 2015). Conditioning on the collider by analyzing sampled respondents induces a spurious association between its common causes, X and U, and results in selection bias (Elwert and Winship 2014).

Endogenous sampling—where the outcome causes selection ($Y \rightarrow S$)—amounts to conditioning on a collider and introduces bias in two situations: (1) when X independently causes selection ($X \rightarrow S \leftarrow Y \leftarrow U$), or 2) when X causes Y (Elwert and Winship 2014). An example of how the second situation would apply to selection into online nonprobability samples (SONS) is shown in panel A of Fig. 1. Removing selection bias due to endogenous sampling requires an explicit selection model, like weighted regression, that adjusts for the differences in selection probabilities across values of the outcome (Berk and Ray 1982; Winship and Radbill 1994).

There is a third situation where X and U may become correlated due to nonprobability online sampling: when SONS and the outcome variable are related spuriously. Confounded sampling amounts to conditioning on a collider and introduces bias when the confounder is a cause of both SONS and Y. Panel B in Fig. 1 illustrates how this can happen. Confounded sampling is probably the most common way that online non-probability sampling leads to selection bias in regression coefficients.³ Here the researcher, drawing on theory and prior

² Although selection on X does not introduce bias, it does reduce efficiency and statistical power (Berk 1983). It also has different consequences for the bivariate correlation (r_{YX}) and regression coefficient (b_{YX}), because both variables are outcomes for the correlation (Sackett and Yang 2000). The Pearson correlation between two variables, X and Y, is simply the geometric mean of the slopes (b_{YX} and b_{XY}) from regressing Y on X and then X on Y.

³ The reason is that typically there are more possible sources of confounded sampling than of endogenous sampling. In an online study of death penalty support, for example, all common causes of SONS and death penalty support (e.g., race, gender, political ideology) would be confounders, whereas the only potential source of endogenous sampling would be death penalty support (or a variable caused by death penalty support).

research, includes in the regression model known correlates (X) of both SONS and Y . In the case of criminal justice attitudes, the X variables would include race and political ideology, among others (e.g., age). However, there remain other unmeasured variables (U) that affect both SONS and Y (Mercer et al. 2017). SONS is thus a collider variable between the included X variables and the unmeasured U variables, and conditioning on SONS (the box around SONS in the Figure) by analyzing crowdsourced workers or opt-in panelists would induce a spurious correlation (dotted line in the Figure) between X and U , resulting in selection bias (Elwert and Winship 2014; Morgan and Winship 2015). A researcher using an online nonprobability sample can remove selection bias due to confounded sampling by including the confounders (U) in the regression model (assuming they are measurable and have positive variance in the sample) (Gelman and Carlin 2002; Gelman 2007). This would not work for endogenous sampling.

Normally, bias in regression coefficients is a function of a larger number of correlations than bias in univariate estimates (e.g., $X \rightarrow \text{SONS} \leftarrow U \rightarrow Y$ vs. $Y \leftarrow X \rightarrow \text{SONS}$), so it is often smaller. For example, selection bias resulting from confounding bias (e.g., confounded sampling) is typically weaker than the confounding bias itself (Greenland 2003). This is another reason why relational inferences may be more robust to sample quality.

When online non-probability sampling leads to selection bias in regression coefficients, the findings will fail to generalize because they will *wrong* in the sample (lack internal validity) and thus *wrong* in the population (Berk 1983; Berk and Ray 1982). Put differently, selection bias results in relationships that are spurious, fully or partially. There is another circumstance where regression coefficients may be internally valid (non-spurious) but externally invalid—that is, where they may be *correct* in the sample, but *wrong* in the population: unmodeled effect heterogeneity (Pasek 2016). Effect heterogeneity occurs when two predictors (X_1 , X_2) interact to affect Y , such that X_1 's effect on Y varies across population subgroups defined by X_2 . In this situation, the composition of the sample with respect to the X variables matters, because it determines the size (and sometimes direction) of estimates in the sample, but not whether they are spurious (Mercer et al. 2017; Solon et al. 2015).

As an example, imagine that: (1) prejudice (X_1) interacts with race (X_2) to affect punitiveness (Y), with prejudice having a positive effect only among whites, and (2) race affects SONS, resulting in an online sample that over-represents whites. Here, regressing Y on X_1 and X_2 in the sample using an otherwise correctly specified model would return an estimate of X_1 's average partial effect that is accurate for the sample, with its particular racial composition, but that overestimates the *population average effect*, since the population has fewer whites. Weighting the sample on the basis of race would not remedy the problem; the weighted regression model typically would still yield an *externally invalid* estimate of the average partial effect (Solon et al. 2015). To identify population regression coefficients in the presence of effect heterogeneity, the researcher would need to include the interaction term(s) in the model and then poststratify using information about the population distribution of the variables (Gelman 2007).

Effect heterogeneity is a concern with online samples because many variables that affect SONS, like race and political ideology (Keeter et al. 2015; Levay et al. 2016), also moderate the effects of other variables on criminal justice attitudes (King and Wheelock 2007; Peffley and Hurwitz 2007; Roche et al. 2016; Simmons 2017). When this is the case, and a researcher using an online non-probability sample fails to include the necessary interaction term(s), the findings will not generalize to the population, even if they are accurate for the sample.

Both selection bias and unmodeled effect heterogeneity are threats to the generalizability of relational inferences from *observational* studies using online non-probability samples. However, only the second of these two—effect heterogeneity—is also a threat to the generalizability of *experimental* inferences from such samples. Because experimental research with online non-probability samples involves random assignment of treatments *after* selection, pre-treatment selection into the sample is not a collider variable. This may explain the apparent robustness of experimental inferences to sample quality (Mullinix et al. 2015; Weinberg et al. 2014). Of course, if experimenters estimate regression models with experimental data and include non-randomized mediators or post-treatment control variables, selection bias would be a threat to the internal validity of findings, as it is in observational studies (Elwert and Winship 2014; Bullock et al. 2010). Criminological experiments conducted with online nonprobability samples often include such non-randomized variables (e.g., Berryessa 2018; Pickett et al. 2018). The key question, then, is whether observational inferences about relationships between criminal justice outcomes and other variables are similar in online non-probability samples and probability samples. The remainder of this paper addresses that question.

Methodology

Data

The probability sample in our analysis is from the General Social Survey (GSS). Since 1994, the GSS, conducted by the National Opinion Research Center (NORC), has gathered data biennially via face-to-face interviews. The GSS utilizes a multistage cluster sampling design, selecting respondents from primary sampling units (PSUs), in order to garner a representative sample of US adults. Because all of the online non-probability samples we analyze were collected between 2015 and 2017 (see below), we use data from the 2016 GSS to make comparisons while minimizing the confounding influence of time period.⁴ While there was a total of 2867 participants surveyed in 2016 GSS, the full analytical samples range from $N=2589$ to $N=821$. The variability in analytical sample sizes is due to item nonresponse as well as the sampling procedure utilized by the GSS. The GSS employs a dual sample rotation design whereby most of the roughly 200 “core” questions are directed at two-thirds of each sample.⁵

We use five online non-probability samples that were drawn from either a crowdsourcing platform, Amazon Mechanical Turk (MTurk), or an opt-in panel platform, Survey-Monkey Audience. Data from each of the five online samples were collected for separate research projects focused on different research questions, but all of the projects included relevant questions from the GSS with identical wording. Four of the online samples came from MTurk. MTurk workers self-select into HITs for small sums of money, and researchers can customize the characteristics of workers allowed to participate in their posted HITs. In line with best practices for using MTurk (Peer et al. 2014), we restricted participation to workers who had completed at least 50 prior HITs and had approval ratings of at least 90%

⁴ The response rate of the 2016 GSS sample is not yet available, however, response rates have consistently hovered around 70% since the year 2000.

⁵ For more information about how the questionnaire items are administered, see Appendix Q of the General Social Survey (GSS), retrieved from <http://gss.norc.umd.edu/DOCUMENTS/CODEBOOK/Q.pdf>.

or 95%. The only other sampling restrictions applied were that workers had to be adults (18 and over) residing in US. No quotas were applied.

SurveyMonkey has a pool of over 400,000 active volunteer panelists who are mostly recruited into the panel after taking surveys administered by individual users—approximately 30 million people participate in such surveys each month (Brandon et al. 2013). After joining SurveyMonkey's opt-in panel, members are then randomly assigned via a router to open surveys for which they meet the criteria.⁶ (For a review of SurveyMonkey's and MTurk's panel recruitment and maintenance strategies, see Brandon et al. 2013; ESOMAR 28). As with the MTurk samples, we did not apply any quotas, and restricted participation only to adults (18 and over) residing in the US.

Before proceeding, it bears noting that the survey modes are different in the online samples (self-administered) and GSS (interviewer-administered). However, we do not believe mode effects are likely to be a problem in our analysis. First, large mode effects normally reflect social desirability bias (Tourangeau and Yan 2007), but none of our outcomes deal with sensitive topics. Second, most of our predictors are demographic variables, which are resistant to errors of observation, like mode effects (Couper 2011). Third, correlations appear to be relatively robust to mode differences (Hox et al. 2015).

Dependent Variables

Across the five online non-probability samples, there are four outcome variables that measure attitudes toward crime and justice and are identical in wording to measures in the GSS.⁷ Table 1 lists the samples, their recruitment platforms, and the outcomes contained in each. The outcomes measure global attitudes towards the police (polhitok), death penalty support (cappun), fear of crime (fear), and law enforcement spending preferences (natcrimty). The exact question wording and coding can be found in Table 2.

Independent Variables

Following previous studies of public opinion on criminal justice (e.g., Johnson and Kuhns 2009; Pickett 2016; Silver and Pickett 2015; Unnever and Cullen 2010), independent variables include available demographic and attitudinal predictors. In cases where response

⁶ The analytic samples for the models estimated with the SurveyMonkey sample are much smaller than the full sample for two reasons. First, several hundred cases have item missing data on education. SurveyMonkey measured this variable at the profile stage of panel recruitment and provided it to us. Changes in the profiling process before our survey resulted in several hundred panelists lacking data on this pre-recorded variable. This data appears to be missing at random with respect to both outcomes—neither outcome differs significantly across respondents with versus without item missing data on education. Second, 288 respondents answered “don't know” to the cappun question, and 101 to the fear question, and these responses are treated as missing in the analysis.

⁷ There is one small presentational difference in the law enforcement spending question asked in the MTurk17 and GSS samples. In the GSS respondents are asked, “We are faced with many problems in this country, none of which can be solved easily or inexpensively. I'm going to name some of these problems, and for each one I'd like you to tell me whether you think we're spending too much money on it, too little money, or about the right amount. First (READ ITEM A)... are we spending too much, too little, or about the right amount on (ITEM)?” Respondents are then asked to decide their spending preferences on a variety of issues. In the MTurk17 sample, it is a standalone question with the same introduction (i.e., respondents are not asked about spending on other topics).

Table 1 Sample descriptions

Sample	Platform	Year fielded	Full sample (N)	Outcomes (full analytic sample)
MTurk15	Amazon Mechanical Turk	2015	1025	“Cappun” (960)
MTurk16a	Amazon Mechanical Turk	2016	915	“Cappun” (874)
MTurk16b	Amazon Mechanical Turk	2016	1149	“Polhitok” (960)
SM16	SurveyMonkey	2016	1781	“Fear” (920) “Cappun” (771)
MTurk17	Amazon Mechanical Turk	2017	1009	“Natcrimy” (963) “Polhitok” (966)

Table 2 Sample outcomes and coding

Outcome	Measure	Coding
“Cappun”	Do you favor or oppose the death penalty for persons convicted of murder?	0 = No 1 = Yes
“Polhitok”	Are there any situations you can imagine in which you would approve a policeman striking an adult male citizen?	0 = No 1 = Yes
“Fear”	Is there an any area right around where you live—that is, within a mile—where you would be afraid to walk alone at night?	0 = No 1 = Yes
“Natcrimy”	We are faced with many problems in this country, none of which can be solved easily or inexpensively. Are we spending too much, too little, or about the right amount on law enforcement?	1 = Too much 2 = About right 3 = Too little

options were different (e.g., income categories), responses were coded to match in the GSS and online nonprobability samples. Demographic predictors include race (0 = Non-White, 1 = White), ethnicity (0 = Non-Latino, 1 = Latino), sex (0 = female, 1 = male), age, education, and income. Age is measured continuously in years in the GSS and in all of the MTurk samples.⁸ But the SM16 sample does not contain a continuous measure of age, so for the models estimated with this sample, we construct a matching categorical age variable for the GSS (1 = 18–29, 2 = 30–44, 3 = 45–59, and 4 = 60 or older). Education is measured dichotomously (1 = Bachelor’s degree). The online samples that include measures of income do not have response categories that align perfectly with the GSS; we recode the GSS categories to match the online samples. We constructed two measures of income in order to create comparable measures of income that maximized the number of categories across the three samples. The first measure of income, used for comparing the GSS to MTurk samples, contains three categories (1 = \$0–\$9999, 2 = \$10,000–\$49,999, 3 = \$50,000 or more). The second measure of income, used for comparing the GSS and SM16 sample, is a binary variable (0 = \$0–\$49,999, 1 = \$50,000 or more).

⁸ For presentational purposes, we divided the continuous age variable by 50. This approach, suggested by one reviewer, makes it easier to see the differences across samples by widening the otherwise small confidence intervals.

The attitudinal variables include political ideology and racial resentment. Ideology is measured as 1 = Liberal, 2 = Moderate, and 3 = Conservative. Racial resentment is measured using an item from the GSS that asks respondents to evaluate the statement: “Irish, Italians, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same without special favors.” Responses are measured using a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). This racial resentment measure is available in the MTurk15 and MTurk16b samples. Consistent with prior work, fear of crime is also utilized as an attitudinal measure to predict death penalty support (Holbert et al. 2004; Johnson 2009), in addition to being a separate outcome variable.

Analytic Strategy

For the analysis, univariate and multivariate, we weight the GSS data to account for the multistage sampling design.⁹ Consistent with conventional practice (Mullinix et al. 2015), however, we do not weight the online non-probability samples. The main reason is that the weights would be based solely on the predictor variables already included in the model, and thus if the model specification is correct, their use would only harm the precision of the estimates (increase the standard errors) (Solon et al. 2015; Winship and Radbill 1994). Nonetheless, we do provide weighted results for the non-probability samples in the online supplement and discuss them in the Supplementary Analysis section below.

Before comparing multivariate results, we first compare the descriptive statistics for the unweighted online non-probability samples to the weighted GSS sample. We then compare the univariate statistics (prevalence estimates) for the outcome variables across the samples. We test whether the sample demographics and outcomes in the online nonprobability samples are significantly different from those in the GSS. Next, we examine whether the online non-probability samples and GSS generate similar multivariate relationships. To do this, we estimate models for each online non-probability sample that include all available predictors for the relevant dependent variable, and estimate an identical model using the GSS data. In total, seven sets of comparable models are estimated. “Don’t know” responses are treated as missing in the analysis, for two reasons: (1) most public opinion studies treat them as missing, and (2) they are less comparable than substantive responses across survey houses (or organizations).¹⁰

To interpret the findings, one question that we must address is: what should be compared, the coefficients’ direction, magnitude, or significance levels? Of these, the least comparable are significance levels, because small differences in p-values across samples (e.g., .049 versus .051) would lead to different conclusions. By contrast, comparing the coefficients’ direction is straightforward, although two coefficients may be in the same direction but of vastly different magnitude. Another method of comparison that is used in the replication literature is to examine whether the coefficient in one sample falls within the 95% confidence interval of another (Open Science Collaboration 2015). We adopt this approach. In addition, when a coefficient in an online non-probability samples falls outside

⁹ To weight the GSS data we used the “WTSSALL” variable and adjusted for the geographic clustering of respondents with the “VSTRAT” and “VPSU” variables. We did this in Stata 15 using the following command: `svyset [weight = wtssall], strata(vstrat) psu(vpsu) singleunit(scaled)`.

¹⁰ As Page and Shapiro (1992, p. 422) explain, “the evidence indicates that ‘house effects’ are mostly limited to one specific area: ‘don’t knows’ ... Thus it is generally safe to compare identical questions across survey organizations, so long as one excludes ‘don’t knows’”.

the GSS confidence interval, we test whether the difference between the coefficients in the two samples is statistically significant¹¹

Results

Table 3 displays the descriptive statistics for each sample. Consistent with previous research (Levay et al. 2016; Weinberg et al. 2014), we find that the demographics of the online non-probability samples consistently differ from those of the GSS sample, and many of these differences are large. The online samples over-represent Whites, young people, the highly educated, and liberals. The differences in education and political ideology are dramatic. The online non-probability sample participants are almost twice as likely to have a college degree as their GSS counterparts, and the percent liberal is around 20 percentage points higher in most of the online samples. Three of the five online samples, MTurk16a, MTurk16b, and MTurk17, over-represent males. The proportion of respondents with incomes of \$50,000 or more is significantly lower in two online samples than in the GSS. Finally, in the two online samples where we measured racial resentment, MTurk15 and MTurk16b, the levels of prejudice are significantly lower than in the GSS sample.

The univariate statistics for the outcome variables measuring attitudes toward criminal justice are also included in Table 3. The prevalence estimates from the online samples not only differ from the GSS estimates, but they also differ from each other. Whereas 61% of GSS respondents support the death penalty, the level of support is consistently lower in the online samples, ranging from a low of 47% (MTurk16a) to a high of 58% (SM16), and the differences are significant in two out of three instances. Roughly, 68% of respondents in both the GSS and MTurk16b believe it is acceptable for police to strike citizens under some circumstances. But the MTurk17 sample is significantly more supportive of police use of force (77%). The online respondents (MTurk17) are also significantly less likely than GSS respondents to want law enforcement spending increased (35% vs. 54%). Finally, there is a statistically significant, 13-percentage-point difference in fear of crime between the GSS and SM16 samples (31% vs. 44%).

Next, our attention turns to the multivariate findings. Figure 2 presents the regression coefficients and confidence intervals from the models estimating support for law enforcement spending in the GSS and MTurk17 samples.¹² In the GSS sample, there are five statistically significant predictors of spending preferences: *White*, *Male*, *Age*, *Moderate* and *Conservative*. Four out of these five relationships are in the same direction in the online sample, but only three are statistically significant. Of course, differences in significance levels are to be expected because of sampling error and differences in sample size. There are three significant predictors in the online sample—*Male*, *Age*, and *Conservative*—and all three match the direction of their corresponding GSS coefficient. In total, five of the

¹¹ Typically, to compare logistic regression coefficients across samples, we would need to use heterogeneous choice models to control for the confounding effects of group differences in residual variation (Williams 2009). But as one reviewer pointed out, the GSS and online samples are assumed to represent the same population, and thus we should not expect differences in residual variation across the samples absent selection bias.

¹² In addition to the figures, tables comparing the weighted GSS and unweighted online estimates can be found in the online supplementary materials.

Table 3 Descriptive statistics

	GSS		Online nonprobability samples											
			MTurk15		MTurk16a		MTurk16b		SM16		MTurk17			
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD		
LE spending	2.43	.68	–	–	–	–	–	–	–	–	2.18*	.69		
Less (%)	10.87	–	–	–	–	–	–	–	–	–	16.42*	–		
Same (%)	35.27	–	–	–	–	–	–	–	–	–	49.05*	–		
More (%)	53.86	–	–	–	–	–	–	–	–	–	34.53*	–		
Fear of crime (%)	31.22	–	–	–	–	–	–	–	43.93*	–	–	–		
Use of force (%)	68.41	–	–	–	–	–	67.54	–	–	–	76.57*	–		
Death penalty (%)	60.92	–	56.16*	–	46.76*	–	–	–	58.19	–	–	–		
White (%)	73.55	–	83.80*	–	83.80*	–	80.49*	–	84.11*	–	83.03*	–		
Latino (%)	14.45	–	5.59*	–	6.36*	–	7.63*	–	7.25*	–	5.31*	–		
Male (%)	45.23	–	46.41	–	57.68*	–	50.51*	–	46.29	–	49.80*	–		
Age	47.56	17.59	37.01*	12.08	34.49*	10.85	36.06*	11.87	–	–	39.37*	12.95		
Age (ordinal)	2.61	1.07	1.99*	.88	1.83*	.82	1.90*	.87	2.57	1.08	2.13*	.91		
18–29 (%)	19.43	–	31.89*	–	38.56*	–	36.70*	–	20.74	–	26.20*	–		
30–44 (%)	26.26	–	44.37*	–	44.89*	–	42.61*	–	26.68	–	43.67*	–		
45–59 (%)	27.85	–	16.60*	–	11.67*	–	14.58*	–	27.09	–	20.68*	–		
60+ (%)	26.46	–	7.14*	–	4.89*	–	6.12*	–	25.49*	–	9.44*	–		
Bachelor's degree (%)	28.86	–	52.76*	–	52.67*	–	52.44*	–	57.24*	–	54.30*	–		
Income (ordinal)	2.49	.62	2.39*	.59	–	–	2.45	.59	–	–	–	–		
\$0–\$9,999 (%)	6.78	–	5.59	–	–	–	5.07	–	–	–	–	–		
\$10,000–\$49,999 (%)	37.47	–	50.05*	–	–	–	45.19*	–	–	–	–	–		
\$50,000+ (%)	55.75	–	44.36*	–	–	–	49.75*	–	–	–	–	–		
Income (binary) (%)	55.75	–	44.36*	–	–	–	49.75*	–	56.14	–	–	–		
Ideology	2.05	.79	1.72*	.79	1.67*	.78	1.76*	.83	1.89*	.78	1.81*	.83		
Liberal (%)	28.55	–	49.05*	–	52.61*	–	48.79*	–	36.67*	–	45.44*	–		

Table 3 (continued)

	GSS		Online nonprobability samples									
	Mean	SD	MTurk15		MTurk16a		MTurk16b		SM16		MTurk17	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Moderate (%)	37.16	–	29.59*	–	28.03*	–	26.01*	–	37.62	–	28.08*	–
Conservative (%)	34.29	–	21.36*	–	19.35*	–	25.20*	–	25.71*	–	26.48*	–
Racial resentment	3.69	1.34	3.13*	1.38	–	–	2.92*	1.39	–	–	–	–

SD Standard deviation

* $p < .05$ (two-tailed) for difference from GSS

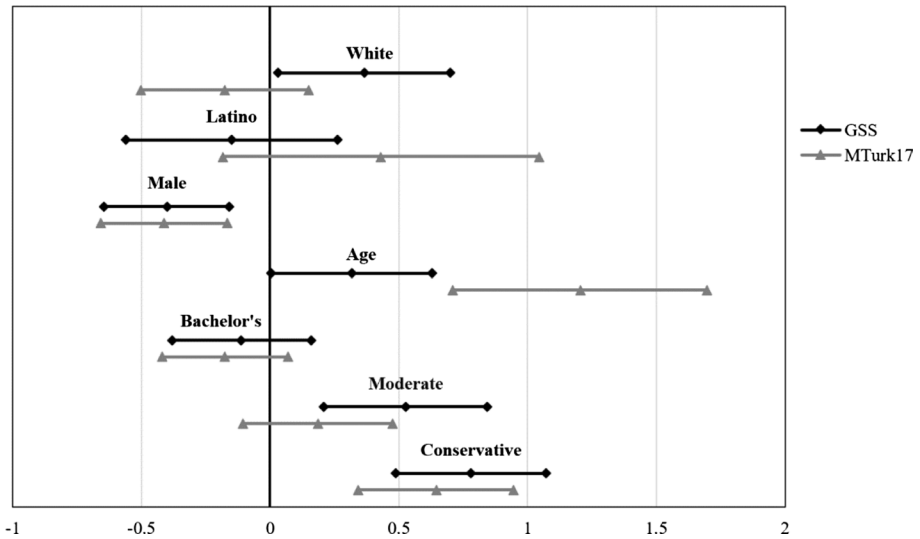


Fig. 2 Coefficients and confidence intervals from ordered logistic regression models predicting law enforcement spending preferences. *Note* Figure shows unstandardized regression coefficients and 95% confidence intervals

seven coefficients, whether significant or nonsignificant, are in the same direction in both samples.

Although in the same direction, many of the coefficients appear to differ considerably in magnitude across the two samples. As noted previously, one straightforward approach to comparing the two sets of results is to examine whether the coefficient for a given variable in the online nonprobability sample falls within the confidence interval for the same variable in probability sample (see, e.g., Open Science Collaboration 2015). Out of the seven variables, four (57%) fail this test: *White*, *Latino*, *Age* and *Moderate*. Using product terms with combined samples (GSS and MTurk17), two of these variables, *White* and *Age*, have coefficients that differ significantly between the two samples.

Figure 3 presents the results for fear of crime. In the GSS sample, there are three statistically significant predictors: *Male*, *Age*, and *Income*. These three relationships are in the same direction in the online sample, and two of the three (*Male* and *Age*) are statistically significant, and these two coefficients are the only significant predictors in the online sample. Across all eight variables included in the two models, four (50%) have coefficients that are in the same direction in both samples. The coefficients for most of the variables, however, differ considerably in magnitude across the two models. Indeed, three of the eight variables (38%)—*Latino*, *Age*, and *Income*—have coefficients in the online sample that fall outside of the respective 95% confidence interval in the GSS sample. When tested formally using product terms with a combined sample (GSS and SM16), only the coefficient for *Income* is significantly different across the two samples.

Figure 4 presents the results for support for police use of force. Here there are four models, because two online samples have this outcome variable and can be compared against the GSS. Different independent variables are available in each online sample. Across the models, there is a total of sixteen variables, and thus we can compare sixteen pairs of coefficients. Thirteen of the sixteen (81%) variables have coefficients in the same direction in

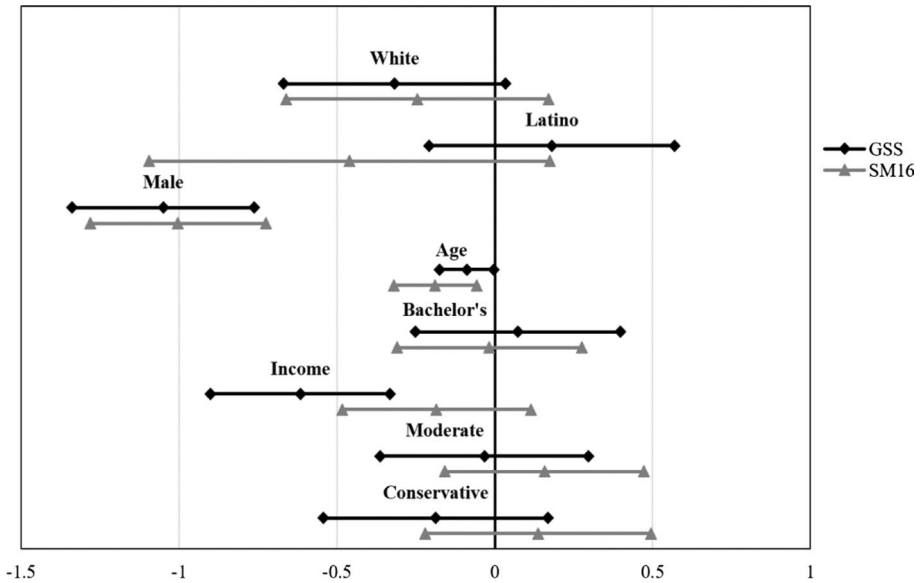


Fig. 3 Coefficients and confidence intervals from binary logistic regression models predicting fear of crime. *Note* Figure shows unstandardized regression coefficients and 95% confidence intervals

the GSS and online samples. Eight of the eleven (73%) coefficients that are statistically significant in the GSS are in the same direction in the online samples, and seven of the eight (88%) coefficients that are significant in the online samples are in the same direction in the GSS. Of the sixteen coefficients estimated for the online samples, 10 (63%) fall outside of the GSS confidence intervals. Six of these ten (60%) coefficients differ significantly from those in the GSS sample.

Finally, Fig. 5 shows the multivariate regression coefficients and confidence intervals from regression models predicting support for the death penalty. This outcome variable was available in three online samples, and each had slightly different independent variables. Altogether, then, we can compare 25 sets of regression coefficients from identical models estimated with the GSS sample and the three online samples. Most of the coefficients in the online samples—17/25 or 68%—are in the same direction as in the GSS sample. In the GSS sample, there are 17 statistically significant coefficients, and 13 (76%) are in the same direction in the online samples, although many are non-significant. Likewise, there are nine significant predictors in the online samples, all of which match the direction of the GSS coefficients. Despite these directional similarities, however, the magnitude of the coefficients again appears to differ across the GSS and online samples. Inspection of the Figure reveals that 13/25 (or 52%) of the coefficients in the online samples fall outside of the 95% GSS confidence intervals. And 6 of these 13 (46%) coefficients differ significantly from those in the GSS sample.

When taken together, what do the results say about the generalizability of the findings from crowdsourced and opt-in samples? The results are summarized in Table 4. Altogether, we saw in Figs. 2, 3, 4 and 5 results from 14 multivariate regression models, seven with the GSS and seven with online non-probability samples, predicting four different outcome variables. In total, the seven models estimated with online samples yielded 56 regression coefficients. These coefficients were mostly in the same direction as those in the GSS (39/56,

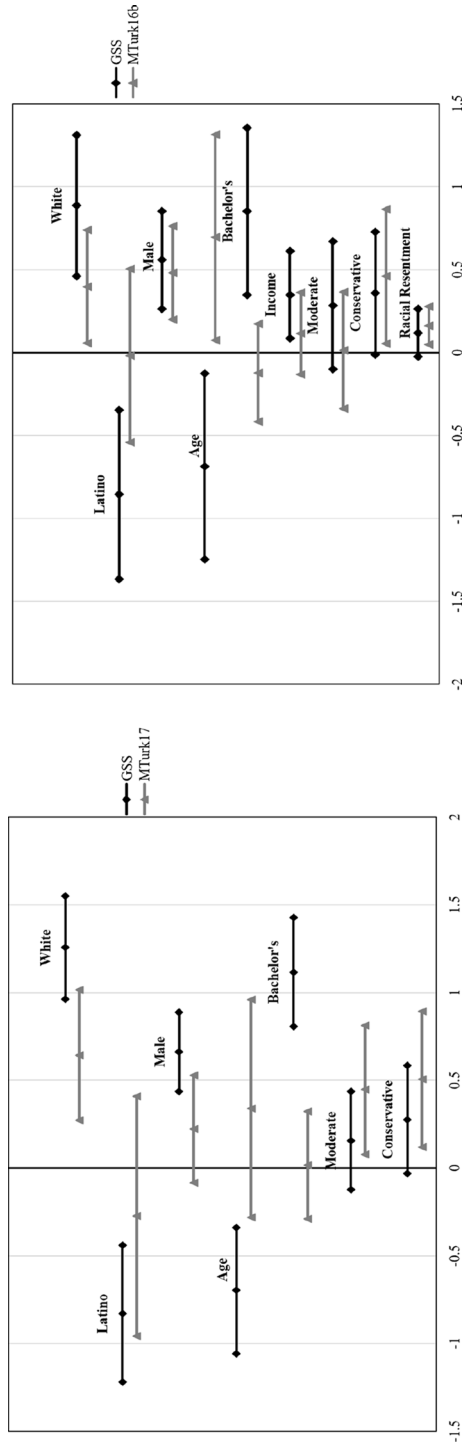


Fig. 4 Coefficients and confidence intervals from binary logistic regression models predicting global support for police use of force. Note Figure shows unstandardized regression coefficients and 95% confidence intervals

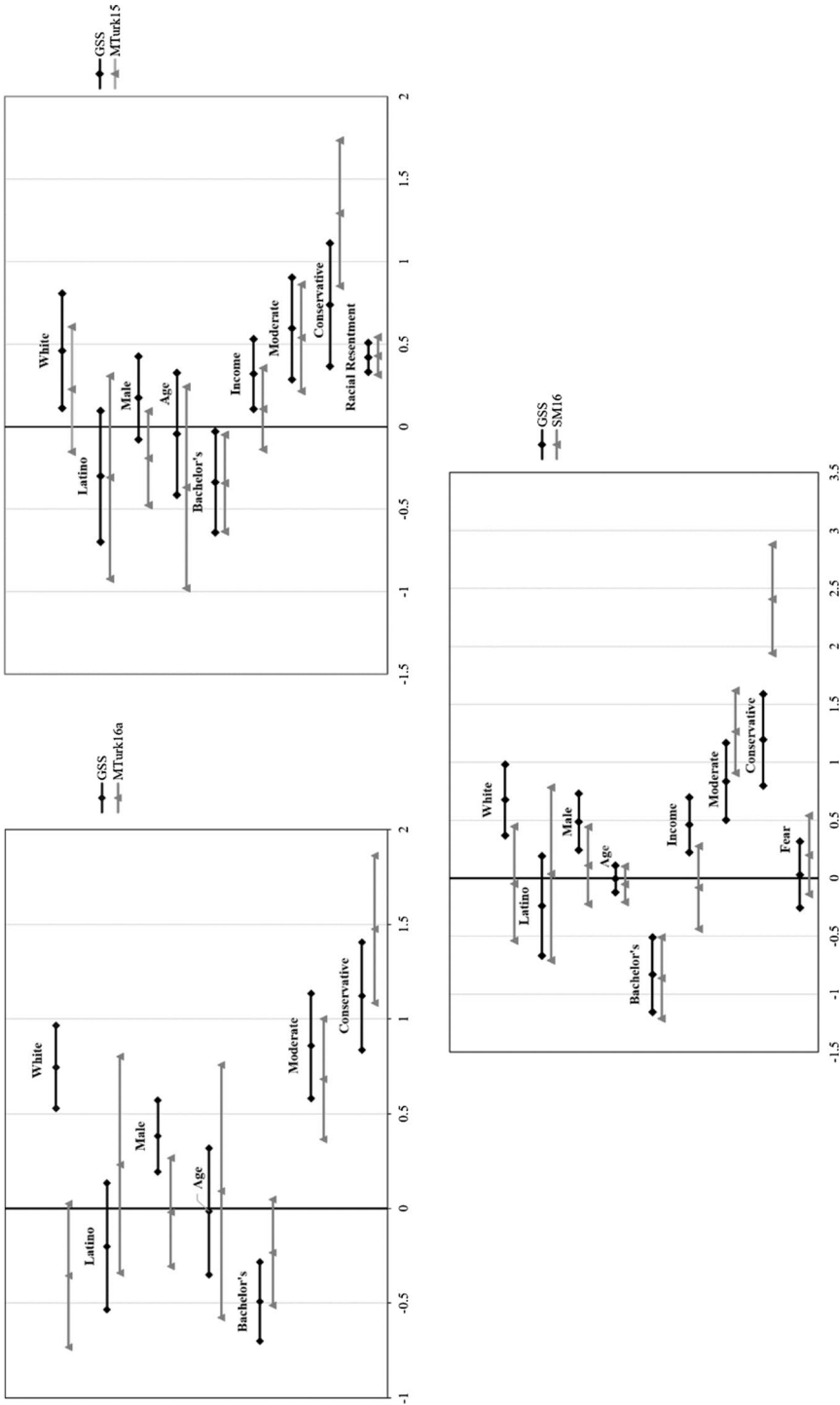


Fig. 5 Coefficients and confidence intervals from binary logistic regression models predicting death penalty support. Note Figure shows unstandardized regression coefficients and 95% confidence intervals

Table 4 Summary of findings when the online non-probability samples are unweighted versus weighted

All estimates	Unweighted	Weighted
Online and GSS coefficients are in the same direction	39/56 (70%)	35/56 (63%)
Online coefficient is outside the GSS confidence interval	30/56 (54%)	32/56 (57%)
Online confidence interval excludes the GSS coefficient	22/56 (39%)	18/56 (32%)
Confidence interval discrepancy AND coefficients differ significantly	15/56 (27%)	14/56 (25%)
Statistically significant estimates		
GSS coefficient is significant AND the online coefficient is in the same direction	28/36 (78%)	26/36 (72%)
Online coefficient is significant AND the GSS coefficient is in the same direction	21/22 (95%)	14/14 (100%)

or 70%), especially if the coefficient reached statistical significance in either sample. Of the 36 coefficients that were statistically significant in GSS, 28 (or 78%) were in the same direction in the online samples, although many were non-significant. Similarly, all but one of the 22 coefficients that were statistically significant in the online samples were in the same direction in the GSS.

Unfortunately, although the coefficients were normally in the same direction in the GSS and online samples, they tended to differ considerably in magnitude. More than half of the coefficients in the online samples (30/56, or 54%) fell outside of the 95% confidence interval in the GSS.¹³ A large portion of these coefficients (15/30, or 50%) also differed significantly from those in the GSS, when tested formally using product terms in combined samples. So, while using an online sample will generally allow for correct inferences about the *direction* of relationships, especially if those relationships are statistically significant, it often will lead to inaccurate estimates of the *size* of those relationships in the general population.

Supplementary Analyses

In supplementary models, we examined the effect of weighting the online samples on the findings. Using an iterative proportional fitting algorithm (raking), we weighted the online samples to match the weighted GSS sample on race, sex, education, income, and age. The results are provided in Tables B1–B7 of the online supplement, and are summarized in Table 4. The substantive conclusions are the same as those reported above: over half (57%) of the coefficients in the online samples fall outside the GSS confidence intervals. The only real difference is that the confidence intervals in the online samples are wider, so they exclude fewer (32% vs. 39%) of the GSS coefficients. This, of course, is to be expected, because estimating a weighted regression model with weights based on the same variables

¹³ If we reverse the comparison, and focus instead on the confidence intervals in the online samples, we find that 22 out of 56 (or 39%) excluded the GSS regression coefficient.

already included in the model will not remove any additional bias, but will reduce the precision of estimates (Solon et al. 2015; Winship and Radbill 1994).¹⁴

Conclusion

Criminologists often justify their use of online non-probability samples by citing research from other fields on the comparability of relational inferences across sample types. For example, Pickett et al. (2013: 737–738) argued, “compared with correlational results from probability samples interviewed either by telephone or in person, findings from non-probability Internet samples yield remarkably similar relational inferences.” The problem with this interpretation, however, is that all of the cited research focused on outcomes other than criminal justice attitudes. Different outcomes will be impacted by selection to different degrees (Simmons and Bobo 2015), because selection bias occurs when there are differences between the sample and population on variables related to the specific outcomes of interest (Elwert and Winship 2014). Our findings suggest that relational inferences for criminal justice attitudes are not as resistant to sample quality as previously thought. This is important given that the use of online non-probability samples for criminological research is increasing (e.g., Enns and Ramirez 2018; Jones and Olderbak 2014; Lageson et al. 2018; Lehmann and Pickett 2017; Vaughan et al. 2019), and is likely to continue to increase, given the growing costs and difficulties of fielding probability samples (Callegaro et al. 2015; Groves et al. 2009).

Do our findings mean that criminologists should stop using online non-probability samples altogether? This would be an overreaction, we think. A more measured response would be to exercise caution when making inferences, whether univariate or relational, about criminal justice attitudes from data collected via online crowdsourcing or opt-in platforms. Specifically, we recommend that criminologists, as well as researchers in other disciplines, take two steps to assess, as best as possible, the extent to which selection bias and effect heterogeneity undermine the internal and external validity of their findings. The first is to conduct a provisional generalizability check. Bhutta (2012) made a similar recommendation. She explained:

[I]ncluding several items in my survey pulled directly from two probability-based surveys enabled me to assess the extent to which Facebook usage correlated with relationships of interest. These benchmarks also enabled me to measure the extent of the bias in the data as well as to evaluate the ability of surveys weights to improve its representativeness (p. 77).

To conduct a provisional generalizability check, researchers using online non-probability samples should, in addition to the questions measuring their main variables of interest, include a few relevant substantive questions that are identical to those used in a recent high-quality probability-based survey, such as the GSS or the American National Election Studies. These “test” questions should measure variables that are closely related to the main variables of interest. For example, when using an online probability sample to examine punitiveness or sentencing preferences, researchers might include the GSS questions

¹⁴ We also estimated supplementary models with both the GSS and online samples unweighted (see Tables C1–C7 in the online supplement). The substantive conclusions were unchanged.

about the death penalty (cappun) and court harshness (courts). In online surveys examining attitudes about gun control, researchers might include the GSS question on gun permits (gunlaw). Online surveys examining attitudes toward police might include the questions from the GSS on police use of force (polhitok) or law enforcement spending (natcrimy), and those that examine cognitive or emotional reactions to crime (e.g., perceived risk, concern, or fear) could include the GSS question on victimization fear (fear).

Researchers can then use the included test questions to obtain (and provide readers with) preliminary evidence about the generalizability of their findings. Specifically, in addition to their main models for the variables of interest, they can also estimate a set of supplementary models using both their online non-probability sample and the probability sample from which the test question was pulled. These supplementary models would be identical in the two samples, taking responses to the test question as the outcome, and whatever other variables are available in both samples (demographic or attitudinal) as the predictors. In our view, the inclusion of these models in an appendix should be a requirement for all studies using online non-probability samples to assess criminal justice attitudes. When relationships for the test variable in the online non-probability sample closely resemble (in direction and magnitude) the benchmarks in the probability sample, it becomes more likely that the main results for the related outcome will generalize. The reasoning is that when differences between the sample and population bias the results for the main variable of interest (e.g., punitiveness), they should also usually bias the results for the related test variable (e.g., death penalty support).

The second step that researchers should take to assess the generalizability of their findings is to conduct a model-specification test by comparing unweighted and weighted results. Weighting and statistical control (through regression) are both forms of model-based adjustment (Morgan and Winship 2015), and both will produce conditional exchangeability in the same circumstance: when the adjustment variables are the variables affecting the probability of selection into the sample (Gelman 2007; Gelman and Carlin 2002). If the variables in the model are the wrong adjustment variables, weighting on them would not help. If they are the right variables, unweighted regression would control for the sampling design and yield more precise estimates (Solon et al. 2015; Winship and Radbill 1994).

Even in this situation, however, differences between weighted and unweighted findings can be informative. They may signal that the unweighted model is improperly specified, which would be the case if the included variables had heterogeneous effects and the researcher failed to model that heterogeneity (omitted necessary interaction terms) (Winship and Radbill 1994; Solon et al. 2015). Weighting does not fix the effect-heterogeneity problem, but it can send up a red flag, alerting researchers that they need to include an interaction term (Solon et al. 2015). Additionally, when weights are a function of other variables not included in the model (more on this shortly), weighted regression may remove selection bias not accounted for by the statistical controls. It is for this reason that statisticians have long suggested testing for significant differences between weighted and unweighted results as a way to check for model misspecification (DuMouchel and Duncan 1983; Pfeffermann 1993). And there are now many methods available for doing this (Bollen et al. 2016). Researchers using online non-probability samples should leverage these methods (e.g., the Stata command “`wgttest`”) to investigate whether they may need to respecify their models to account for effect heterogeneity.

The bad news, and the main obstacle to obtaining valid inferences from online nonprobability samples, is that any form of model-based adjustment, statistical control or weighting, will fail unless researchers use the correct adjustment variables (Mercer et al. 2017).

Our study indicates that controlling for (or weighting on) demographics and political ideology is insufficient to ensure that crowdsourced and opt-in sampling designs are exogenous for criminal justice outcomes, like fear of crime and support for the death penalty. This is consistent with an earlier study that found the level of death penalty support was significantly lower in an MTurk sample than in a probability sample, even after adjusting for demographics and political beliefs (Levay et al. 2016). More generally, research on quasi-experimental design shows that adjusting for demographics is rarely sufficient to account for selection (Shadish et al. 2008). So, the challenge for criminologists using online non-probability samples is to identify the relevant design variables—the variables associated with both SONS and criminal justice outcomes.

This again is why conducting a provisional generalizability check is important. Researchers using online non-probability samples often estimate weighted regression models with weights that are a function of a large number of variables not included in the model, hoping that doing so results in conditional exchangeability. Indeed, opt-in panel providers, like YouGov, often provide these weights to researchers. If, as we have recommended, researchers have included test questions to conduct a provisional generalizability check, they can use the broader design weights to estimate weighted regression models for the test variables in the online non-probability sample. They can then examine whether the adjustments increase the similarity of the results to those in the probability sample. Improvements provide evidence that the weighting adjustments account for the relevant differences between the sample and population, increasing confidence in the generalizability of the findings (See Bhutta 2012; Mercer et al. 2017).

The good news is that with the correct model of the data generation process and appropriate adjustments, researchers can obtain accurate estimates even with very unrepresentative non-probability samples. Wang et al. (2015), for example, used daily polling conducted on an online gaming platform, Xbox, to accurately model the 2012 election. Impressively, the authors were also able to render relatively accurate predictions of sample subgroups, even for groups rare to the Xbox platform such as women aged 65 and older. The ability to demonstrate success by remedying sampling recruitment flaws through an empirical model, a model-based path to generalizability, is critical for non-probability samples (Baker et al. 2010). A focus on model-based inference also challenges researchers using online non-probability samples to identify a priori potentially confounding covariates that should be actively accounted for during data collection and analysis (Mercer et al. 2017). All this is to say that greater theoretical and empirical attention should be devoted to understanding the relationship between SONS and criminal justice outcomes.

Some sampling approaches used with opt-in panels, such as the YouGov's sample-matching method, are designed around model-based inference, and may yield more representative online non-probability samples than those we analyzed (Rivers 2007). Indeed, the Cooperative Campaign Analysis Project uses online non-probability samples drawn from an opt-in panel via sample matching (Ansolabehere and Rivers 2013). And several recent studies in our field use matched samples (e.g., Enns and Ramirez 2018; Lehmann and Pickett 2017). However, matching is just another form of adjustment. As with any adjustment method, the effectiveness of matching on a given set of variables, as YouGov does, will vary depending on the specific outcome variable in the analysis (Mercer et al. 2018; Simmons and Bobo 2015). There is thus a need for future research that explores whether matched online non-probability samples yield similar relational inferences as probability samples for criminal justice outcomes. The use of matching aside, we would still suggest

that researchers using these samples report the results of a provisional generalizability check and a model-specification test.

Future studies should also explore whether experiments on criminal justice attitudes that use online non-probability samples yield externally valid findings. Our study focused only on observational inferences about relationships between variables, which should be more sensitive to sample quality because selection threatens both their internal and external validity. In experiments, however, with a large enough sample (and on expectation), selection threatens only the external validity of findings (Shadish et al. 2002). After reviewing the relevant experimental research, Callegaro et al. (2014) concluded, “[T]he limited evidence so far does not suggest there are substantial differences in either replication or size of effects across probability and non-probability samples” (p. 43). Mullinix et al. (2015) came to the same conclusion. Again, however, the accuracy of this conclusion will depend on the outcome. When effect heterogeneity exists—which will be often for criminal justice outcomes, because race and political ideology tend to moderate other effects (Peffley and Hurwitz 2007; Roche et al. 2016; Simmons 2017)—experiments using online non-probability samples will not generate accurate estimates of population average effect sizes, unless they account for the heterogenous treatment effects (Weinberg et al. 2014).

Before closing, several limitations of our analysis bear mention, which provide opportunities for future research seeking to build on our study. First, in our study the survey mode differed across the samples. The GSS was administered via face-to-face interviews, whereas the five online surveys were self-administered. This is a common limitation in studies comparing findings from online non-probability samples and probability samples (e.g., Bhutta 2012; Pasek 2016; Pasek and Krosnick 2010; Simmons and Bobo 2015). The consequence is that we cannot ascertain the extent to which differences between the samples herein are due to errors of observation (e.g., satisficing), errors of non-observation, or some combination of both. Future research comparing online non-probability samples with probability samples should attempt to make the survey mode consistent across samples.

Second, our study examined only four measures of criminal justice attitudes. Of course, it possible that relational inferences for other types of criminal justice attitudes—for example, views about gun control, sanction perceptions, or perceptions of police procedural justice—from online non-probability samples would be less sensitive to sample quality and thus more similar to those obtained with a probability sample. Future studies should explore this possibility. More broadly, greater empirical attention should be devoted to understanding how the selection process in online sampling impacts different types of inferences (univariate, relational) for criminal justice attitudes.

To conclude, our study suggests that if researchers use an online nonprobability sample to examine criminal justice attitudes, they should expect the following: (1) the unweighted univariate prevalence estimates will likely be inaccurate; (2) the relationships between variables will likely be in the correct direction, but either over- or under-estimated in magnitude, often to a large extent; and (3) adjusting for demographics will not result in conditional exchangeability. We advise that researchers using these samples always conduct and report the results of a provisional generalizability check using test questions that measure constructs related to the main variables of interest and that have been pulled directly from a recent probability-based survey. We also suggest researchers test for model misspecification by examining whether weighted and unweighted findings differ significantly.

Acknowledgements The authors thank Jasmine Silver, Sean Roche, Luzi Shi, Megan Denver, and Shawn Bushway for their help collecting data.

References

- Ansolabehere S, Rivers D (2013) Cooperative survey research. *Annu Rev Polit Sci* 16:307–329
- Baker R, Blumberg SJ, Brick JM, Couper MP, Courtright M, Dennis JM, Dillman D, Frankel MR, Garland P, Groves RM, Kennedy C, Krosnick JA, Lavrakas PJ, Lee S, Link M, Piekarski L, Rao K, Thomas RK, Zahs D (2010) Research synthesis: aAPOR report on online panels. *Public Opin Q* 74:711–781
- Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, Dever JA, Gile KJ, Tourangeau R (2013) Summary report of the AAPOR task force on non-probability sampling. *J Sur Stat Methodol* 1:90–143
- Berk RA (1983) An introduction to sample selection bias in sociological data. *Am Sociol Rev* 48:386–398
- Berk RA, Ray SC (1982) Selection biases in sociological data. *Soc Sci Res* 11:352–398
- Berryessa CM (2018) The effects of psychiatric and “biological” labels on lay sentencing and punishment decisions. *J Exp Criminol* 14:241–256
- Bhutta C (2012) Not by the book: facebook as a sampling frame. *Sociol Methods Res* 41:57–88
- Blair J, Czaja RF, Blair EA (2013) *Designing surveys: a guide to decisions and procedures*. Sage, Thousand Oaks
- Bollen KA, Biemer PP, Karr AF, Tueller S, Berzofsky ME (2016) Are survey weights needed? A review of diagnostic tests in regression analysis. *Annu Rev Stat Appl* 3:375–392
- Brandon DM, Long JH, Loraas TM, Mueller-Phillips J, Vansant B (2013) Online instrument delivery and participant recruitment services: emerging opportunities for behavioral accounting research. *Behav Res Account* 26:1–23
- Brown EK, Socia KM (2017) Twenty-first century punitiveness: social sources of punitive American views reconsidered. *J Quant Criminol* 33:935–959
- Bullock JG, Green DP, Ha SE (2010) Yes, but what’s the mechanism? (don’t expect an easy answer). *J Pers Soc Psychol* 98:550–558
- Callegaro M, Villar A, Krosnick J, Yeager D (2014) A critical review of studies investigating the quality of data obtained with online panels. In: Callegaro M, Baker R, Bethlehem J, Goritz A, Krosnick J, Lavrakas P (eds) *Online panel research: a data quality perspective*. Wiley, New York, pp 23–53
- Callegaro M, Manfreda KL, Vehovar V (2015) *Web survey methodology*. Sage, Thousand Oaks
- Casey LS, Chandler J, Levine AS, Proctor A, Strolovitch DZ (2017) Intertemporal differences among MTurk workers: time-based sample variations and implications for online data collection. *SAGE Open* 7:1–15
- Chandler J, Shapiro D (2016) Conducting clinical research using crowdsourced convenience samples. *Annu Rev Clin Psychol* 12:53–81
- Chandler J, Mueller P, Paolacci G (2014) Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behav Res Methods* 46:112–130
- Chang L, Krosnick JA (2009) National surveys via RDD telephone interviewing versus the internet: comparing sample representativeness and response quality. *Public Opin Q* 73:641–678
- Couper MP (2011) The future of modes of data collection. *Public Opin Q* 75:889–908
- Denver M, Pickett JT, Bushway SD (2017) Criminal records and employment: a survey of experiences and attitudes in the United States. *Justice Q* 35:584–613
- Dum CP, Socia KM, Rydberg J (2017) Public support for emergency shelter housing interventions concerning stigmatized populations. *Criminol Public Policy* 16:835–877
- DuMouchel WH, Duncan GJ (1983) Using sample survey weights in multiple regression analyses of stratified samples. *J Am Stat Assoc* 75:535–543
- Elliott MR, Valliant R (2017) Inference for nonprobability samples. *Stat Sci* 32:249–264
- Elwert F, Winship C (2014) Endogenous selection bias: the problem of conditioning on a collider variable. *Annu Rev Sociol* 40:31–53
- Enns PK, Ramirez M (2018) Privatizing punishment: testing theories of public support for private prison and immigration detention facilities. *Criminology* 56:546–573
- ESOMAR 28: Surveymonkey Audience (2013) European Society for Opinion and Marketing Research, Amsterdam. <https://www.esomar.org/>
- Gelman A (2007) Struggles with survey weighting and regression modeling. *Stat Sci* 22:153–164
- Gelman A, Carlin JB (2002) Poststratification and weighting adjustments. In: Groves RM, Dillman DA, Eltinge JL, Little RJA (eds) *Survey nonresponse*. Wiley, New York, pp 289–302
- Gottlieb A (2017) The effect of message frames on public attitudes toward criminal justice reform for nonviolent offenses. *Crime Delinq* 63:636–656
- Greenland S (2003) Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiol* 14:300–306
- Groves RM, Fowler FJ, Couper MP, Lepkowski J, Singer E, Tourangeau R (2009) *Survey methodology*, 2nd edn. Wiley, Hoboken

- Holbert RL, Shah DV, Kwak N (2004) Fear, authority, and justice: crime-related TV viewing and endorsements of capital punishment and gun ownership. *Journal Mass Commun Q* 81:343–363
- Horton JJ, Rand DG, Zeckhauser RJ (2011) The online laboratory: conducting experiments in a real labor market. *Exp Econ* 14:399–425
- Hox JJ, De Leeuw ED, Zijlman EA (2015) Measurement equivalence in mixed mode surveys. *Front Psychol* 6:1–10
- Johnson D (2009) Anger about crime and support for punitive criminal justice policies. *Punishm Soc* 11:51–66
- Johnson D, Kuhns JB (2009) Striking out: race and support for police use of force. *Justice Q* 26:592–623
- Jones DN, Olderbak SG (2014) The associations among dark personalities and sexual tactics across different scenarios. *J Interp Viol* 29:1050–1070
- Keeter S, McGeeney K, Mercer A, Hatley N, Pattern E, Perrin A (2015) Coverage error in internet surveys. Pew Research Center, Washington. Retrieved from <https://www.pewresearch.org/methods/2015/09/22/coverage-error-in-internet-surveys/>
- King RD, Wheelock D (2007) Group threat and social control: race, perceptions of minorities and the desire to punish. *Soc Forces* 85:1255–1280
- Lageson SE, McElrath S, Palmer KE (2018) Gendered public support for criminalizing “Revenge Porn”. *Feminist Criminol* 14:560–583
- Lehmann PS, Pickett JT (2017) Experience versus expectation: economic insecurity, the Great Recession, and support for the death penalty. *Justice Q* 34:873–902
- Levy KE, Freese J, Druckman JN (2016) The demographic and political composition of Mechanical Turk samples. *Sage Open* 6:1–17
- Little A, Rubin DB (2002) Statistical analysis with missing data. Wiley, New York
- Mercer AW, Kreuter F, Keeter S, Stuart EA (2017) Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. *Public Opin Q* 81:250–271
- Mercer A, Lau A, Kennedy C (2018) For weighting online opt-in samples, what matters most?. Pew Research Center, Washington
- Morgan SL, Winship C (2015) Counterfactuals and causal inference. Cambridge University Press, Oxford
- Mullinix KJ, Leeper TJ, Druckman JN, Freese J (2015) The generalizability of survey experiments. *J Exp Pol Sci* 2:109–138
- Nicolaas G, Calderwood L, Lynn P, Roberts C (2014) Web surveys for the general population: How, why and when?. National Centre for Research Methods, Southampton. Retrieved from <http://eprints.ncrm.ac.uk/3309/3/GenPopWeb.pdf>
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716
- Page BI, Shapiro RY (1992) The rational public: fifty years of trends in Americans’ policy preferences. Chicago University Press, Chicago
- Pasek J (2016) When will nonprobability surveys mirror probability surveys? Considering types of inference and weighting strategies as criteria for correspondence. *Int J Public Opin Res* 28:269–291
- Pasek J, Krosnick JA (2010) Measuring intent to participate and participation in the 2010 census and their correlates and trends: comparisons of RDD telephone and non-probability sample internet survey data. Statistical Research Division of the US Census Bureau, Washington. Retrieved from https://www.mod.gov.se/digitalAssets/1456/1456661_pasek-krosnick-mode-census.pdf
- Peer E, Vosgerau J, Acquisti A (2014) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav Res Methods* 46:1023–1031
- Peffley M, Hurwitz J (2007) Persuasion and resistance: race and the death penalty in America. *Am J Pol Sci* 51:996–1012
- Peytchev A (2009) Survey breakoff. *Public Opin Q* 73:74–97
- Peytchev A (2011) Breakoff and unit nonresponse across web surveys. *J Off Stat* 27:33–47
- Pfeffermann D (1993) The role of sampling weights when modeling survey data. *Int Stat Rev* 61:317–337
- Pickett JT (2016) On the social foundations for crimmigration: latino threat and support for expanded police powers. *J Quant Criminol* 32:103–132
- Pickett JT, Mancini C, Mears DP (2013) Vulnerable victims, monstrous offenders, and unmanageable risk: explaining public opinion on the social control of sex crime. *Criminology* 51:729–759
- Pickett JT, Cullen F, Bushway SD, Chiricos T, Alpert G (2018) The response rate test: nonresponse bias and the future of survey research in criminology and criminal justice. *Criminologist* 43:7–11
- Rivers D (2007) Sampling for web surveys. Joint Statistical Meetings, Salt Lake
- Roche SP, Pickett JT, Gertz M (2016) The scary world of online news? Internet news exposure and public attitudes toward crime and justice. *J Quant Criminol* 32:215–236

- Ross J, Irani L, Silberman M, Zaldivar A, Tomlinson B (2010) Who are the crowdworkers? Shifting demographics in Mechanical Turk. In: Edwards K, Rodden T, Proceedings of the ACM conference on human factors in computing systems. ACM, New York
- Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 66:688
- Sackett PR, Yang H (2000) Correction for range restriction: an expanded typology. *J Appl Psychol* 85:112–118
- Shadish W, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin, Boston
- Shadish WR, Clark MH, Steiner PM (2008) Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *J Am Stat Assoc* 103:1334–1344
- Sheehan KB, Pittman M (2016) Amazon's Mechanical Turk for academics: The HIT handbook for social science research. Melvin and Leigh, Irvine
- Silver JR, Pickett JT (2015) Toward a better understanding of politicized policing attitudes: conflicted conservatism and support for police use of force. *Criminology* 53:650–676
- Silver JR, Silver E (2017) Why are conservatives more punitive than liberals? A moral foundations approach. *Law Human Behav* 41:258–272
- Simmons AD (2017) Cultivating support for punitive criminal justice policies: news sectors and the moderating effects of audience characteristics. *Soc Forces* 96:299–328
- Simmons AD, Bobo LD (2015) Can non-full-probability internet surveys yield useful data? A comparison with full-probability face-to-face surveys in the domain of race and social inequality attitudes. *Sociol Methodol* 45:357–387
- Solon G, Haider SJ, Wooldridge JM (2015) What are we weighting for? *J Hum Resour* 50:301–316
- Stewart N, Chandler J, Paolacci G (2017) Crowdsourcing samples in cognitive science. *Trends Cogn Sci* 21:736–748
- Tourangeau R, Yan T (2007) Sensitive questions in surveys. *Psychol Bull* 133(5):859
- Tourangeau R, Frederick G, Conrad FG, Couper MP (2013) The science of web surveys. Oxford University Press, Oxford
- Unnever JD, Cullen FT (2010) The social sources of Americans' punitiveness: a test of three competing models. *Criminology* 48:99–129
- Unnever JD, Cullen FT, Jonson CL (2008) Race, racism, and support for capital punishment. *Crime Justice* 37:45–96
- Valliant R, Dever JA (2011) Estimating propensity adjustments for volunteer web surveys. *Sociol Methods Res* 40:105–137
- Vaughan TJ, Holleran LB, Silver J (2019) Applying moral foundations theory to the explanation of capital jurors' sentencing decisions. *Justice Q*. <https://doi.org/10.1080/07418825.2018.1537400>
- Wang W, Rothschild D, Goel S, Gelman A (2015) Forecasting elections with non-representative polls. *Int J Forecast* 31:980–991
- Weinberg JD, Freese J, McElhattan D (2014) Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsourcing-recruited sample. *Sociol Sci* 1:292–310
- Williams R (2009) Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociol Methods Res* 37:531–559
- Winship C, Radbill L (1994) Sampling weights and regression analysis. *Soc Methods Res* 23:230–257
- Yeager DS, Krosnick JA, Chang L, Javitz HS, Levendusky MS, Simpser A, Wang R (2011) Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opin Q* 75:709–747
- Zhou H, Fishbach A (2016) The pitfall of experimenting on the web: how unattended selective attrition leads to surprising (yet false) research conclusions. *J Pers Soc Psychol* 111:493–504

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Andrew J. Thompson¹  · Justin T. Pickett¹

¹ School of Criminal Justice, University at Albany, SUNY, 135 Western Avenue, Albany, NY 12222, USA