

# The face of internet recruitment: Evaluating the labor markets of online crowdsourcing platforms in China

Xiaojun Li<sup>1</sup>, Weiyi Shi<sup>2</sup> and Boliang Zhu<sup>3</sup>

## Abstract

Zhubajie/Witmart and other online crowdsourcing platforms have proliferated in China, and researchers have increasingly used them for subject recruitment. One critical question remains, however: what is the generalizability of the findings based on these online samples? In this study, we benchmark the demography of an online sample from Zhubajie to nationally representative samples and replicate commonly asked attitudinal questions in national surveys. We find that online respondents differ from the general population in many respects. Yet, the differences become smaller when comparison is made with the internet users in benchmark surveys. Importantly, when predicting attitudes, our online sample with post-stratification weights is able to produce similar coefficients in most cases as these internet-active subsamples. Our study suggests that online crowdsourcing platforms can be a useful tool for subject recruitment, especially when researchers are interested in making inferences about Chinese netizens. We further analyze the political and social desirability issues of online subjects. Finally, we discuss caveats of using crowdsourcing samples in China.

## Keywords

Crowdsourcing, subject recruitment, online sample, external validity, response bias, China

## Introduction

Recruiting subjects for public opinion surveys typically imposes high costs and logistical burdens on researchers. These challenges are particularly pronounced in an authoritarian environment such as China. The advent of Chinese online crowdsourcing platforms provides researchers with a promising alternative for subject recruitment in China, where popular international platforms such as Amazon Mechanical Turk (MTurk) remain inaccessible.<sup>1</sup> Scholars in social science and other disciplines have been increasingly using Chinese crowdsourcing platforms such as Zhubajie for subject recruitment. For instance, a key word search of “Zhubajie” and “crowdsourcing” on Google Scholar returns 139 results since 2010.<sup>2</sup> Nevertheless, we have yet to systematically evaluate the external validity of crowdsourcing samples. How do subjects obtained from online recruitment differ from a nationally representative sample in China? Are we able to make generalizable inferences based on their responses?

In this study, we take a first step toward providing answers to these questions. We do so by comparing a host of demographic and attitudinal measures between a sample from Zhubajie, China’s largest crowdsourcing platform, and samples from four widely used population-based surveys. We find that the Zhubajie sample is more representative of Chinese netizens than of the general population. Importantly, in predicting public attitudes the Zhubajie sample with post-stratification weights is able to produce

<sup>1</sup>Department of Political Science, University of British Columbia, Canada

<sup>2</sup>School of Global Policy and Strategy, University of California, USA

<sup>3</sup>Department of Political Science, Pennsylvania State University, USA

## Corresponding author:

Boliang Zhu, Department of Political Science, Pennsylvania State University, 317 Pond Lab, University Park, PA 16802, USA.  
Email: bxz14@psu.edu



similar point estimates as the internet-user subsamples of the population-based surveys. Our finding suggests that online crowdsourcing platforms can be a useful tool for subject recruitment, especially when researchers are interested in making inferences about Chinese netizens. China's online population is itself an important subject for scholars interested in China. Netizens represent more than half of China's population, and the number continues to grow quickly. They are the group of people who have almost instant access to news about major sociopolitical events in China, who are more likely to express discontent with the government, and whose opinions tend to have more influence on policy decision-making (Shirk, 2007).

We further explore the political and social desirability issue and find that online subjects are more likely to hide their truthful preferences toward politically sensitive issues than socially sensitive ones. This suggests that the benefit extended by the presumed anonymity of self-administered online subject recruitment does not eliminate the fear of government monitoring in an authoritarian environment such as China's, where netizens are highly aware of extensive governmental surveillance of online activities. However, this aversion to political sensitivity could be even more pronounced in the context of in-person surveys. It is important to examine this question in future research to further assess the benefits of anonymous online surveys in authoritarian countries.

### Subject recruitment through Zhubajie

Subject recruitment on Zhubajie (zjb.com) works in a similar fashion to MTurk (see Berinsky et al., 2012) and is highly cost-efficient compared to in-person surveys. We describe the recruitment procedure in detail in Online Appendix A. For our study, we recruited a total of 1419 subjects aged 18 or above and living in Mainland China between July 21 and August 21, 2016. We modeled our advertisement after those posted by other social scientists recruiting survey respondents on the platform: a brief introduction to the purpose of the survey and participation rules, followed by an external link to Qualtrics.<sup>3</sup> This helped to ensure that our recruitment process would closely mirror current studies using the platform.

Repetitive participation is one of the most common violations of participation rules. We identify repetitive participation by tracking respondents' Internet Protocol (IP) addresses in Qualtrics, taking only the first response from a given IP address (and thereby eliminating 58 repeat respondents, or 4.1% of the total respondents). We checked the "prevent-ballot-box-stuffing" box in Qualtrics but did not restrict IP addresses in our recruitment process.<sup>4</sup> Even so, this rate of repeat participation is only marginally higher than what Berinsky et al. (2012) identify in their MTurk sample.<sup>5</sup> We also embedded instructional manipulation check questions (Oppenheimer et al., 2009). We

found 213 respondents (15.01%) failed to pass the attentiveness check, a rate that is somewhat lower than the typical rate on MTurk (see Berinsky et al., 2012). Our final sample used for the demographic comparisons below includes responses that are unique by IP address.<sup>6</sup> We focus on attentive respondents in coefficient equality tests and the analysis of political and social desirability bias. This leaves us with a total of 1089 valid responses. We address the inattentiveness issue in Online Appendix D. We do not find systematic evidence that inattentive respondents on zjb.com invalidate statistical inferences. Our results suggest that respondents' cognitive ability is the most likely factor driving inattentiveness.

### Demographic comparisons

We compare our Zhubajie sample with four national benchmarks: the 2013 Chinese General Social Survey (CGSS), the third wave (2011) of the Asian Barometer Survey (ABS), the sixth wave (2012) of the World Value Survey (WVS), and the 2008 China Survey (CS) hosted at Texas A&M University (see Online Appendix F.1 for the sampling methods and implementation of the four benchmark surveys). To our knowledge, these four are the best benchmark surveys that we can use at this stage. We are aware that these benchmark surveys were conducted three to four years before our Zhubajie survey and that the CS is already eight years earlier than ours. The difference in the timing may result in some discrepancies in both the demographic composition and public attitudes.<sup>7</sup>

We compare a total of six demographic variables of common interest between the Zhubajie sample and the four benchmark surveys: age, gender, education, household registration type (urban or rural), Chinese Communist Party (CCP) membership, and employment status. We chose these six demographic variables because they are commonly used in survey research and the questions have similar wording and answer choices for each of these variables in the five surveys, allowing us to use the same measures across the board.<sup>8</sup>

Table 1 presents the comparisons. For the Zhubajie sample, we calculate the summary statistics using both the raw data and two post-stratification weights based on two strata: age and gender. The first weight variable is computed from the 2010 population census.<sup>9</sup> The second weight variable is derived using data from the latest China Statistical Report on Internet Development (CSRID), issued by the China Internet Network Information Center (CNNIC) in January 2017.<sup>10</sup> Summary statistics for all benchmark surveys except the CGSS<sup>11</sup> are calculated with the built-in weights.

There are statistically significant differences between the unweighted Zhubajie sample and the benchmark full samples.<sup>12</sup> Recruitment on Zhubajie tends to yield more male than female respondents, and they tend to be younger and live in urban areas.<sup>13</sup> They are also highly educated and more likely to be CCP members. Post-stratification weighting

**Table 1.** Comparing Zhubajie (ZBJ) and benchmark samples on age, gender, education, household registration, Chinese Communist Party (CCP) membership, and employment status.

	Age	Gender	Education	Urban	CCP	Employed
ZBJ (unweighted)	26.43	0.35	13.81	0.47	0.14	0.76
ZBJ (census weights)	32.54	0.48	13.48	0.62	0.17	0.83
ZBJ (CNNIC weights)	28.88	0.48	13.30	0.56	0.14	0.68
<i>Benchmark (full sample)</i>						
ABS (weighted)	44.61	0.51	5.82			0.75
CGSS	48.59	0.50	8.65	0.36	0.05	0.40
CS (weighted)	39.50	0.45	8.31	0.28	0.09	0.85
WVS (weighted)	43.02	0.49	9.19		0.07	0.74
<i>Benchmark (netizens)</i>						
ABS (weighted)	33.37	0.52	8.30			0.70
CGSS	32.64	0.45	12.94	0.56	0.15	0.71
CS (weighted)	30.15	0.39	11.75	0.61	0.16	0.91
WVS (weighted)	35.21	0.45	11.55		0.12	0.79

Notes: CNNIC, China Internet Network Information Center; ABS, Asian Barometer Survey; CGSS, Chinese General Social Survey; CS, China Survey; and WVS, World Value Survey.

helps ameliorate some of the demographic disparity, especially in terms of age and gender (which achieve parity), but at the same time widens the gap between the crowdsourcing and benchmark samples in measures such as urban household registration and employment status.

Service providers on Zhubajie are active internet users because it requires participants to track tasks constantly and complete them online in a timely manner. As such, samples recruited from Zhubajie presumably should be more representative of China's netizens, who account for 53.2% of the Chinese population according to the 39<sup>th</sup> biannual report issued by the CNNIC in January 2017. Among China's netizens, 52.4% are male and 30.3% are between the ages of 20 and 29.

This is indeed the case when we compare the same demographic variables for Zhubajie respondents and internet-active respondents in the benchmark surveys (bottom panel of Table 1). We consider respondents to be internet-active if they said that: they use the internet "almost daily or at least once a week" (ABS and WVS); the internet is their "primary source of information" (CGSS); and they "have an internet connection" (CS). We see that all six demographic measures in the weighted Zhubajie samples are statistically indistinguishable from at least one, if not all, of the benchmark surveys. These results suggest that the online crowdsourcing sample is much more representative of Chinese netizens than of the general population.

### Comparisons on common attitudinal measures: trust, national pride, and inequality

We investigate five attitudinal measures about which existing Chinese surveys commonly ask: interpersonal trust, trust in the central government, trust in the local government,

national pride, and attitude toward inequality. The interpersonal trust question asks respondents whether "most people can be trusted" or "one must be very careful in dealing with people." The trust in government question asks respondents how much they trust the central and local government in China. The national pride question asks respondents how proud they are of being Chinese. The inequality question asks respondents to what extent they endorse this statement: "For equality, the government should give added help to the underprivileged in society." Table 2 depicts the responses to questions on trust, national pride, and inequality in the Zhubajie sample and the benchmark samples. For ease of comparison, we calculate the ratio of respondents with an affirmative answer to each question, collapsing the Likert scales into binary ones.

While public attitudes in the benchmark surveys are generally consistent with each other, they diverge substantially from those of online subjects. Compared with the population-based samples, respondents on Zhubajie appear to be more trusting of other people, but less so when it comes to both the national and local governments. In particular, Zhubajie respondents have a much lower trust in local government than those in the benchmark samples. They also exhibit less national pride and are less likely to think that the government should give help to the poor in the society.

The pattern of these differences mirrors how the internet-using subsamples differ from their non-internet-using subsamples in all of the national benchmark surveys. In other words, internet users in the national surveys also tend to exhibit higher personal trust, lower governmental trust, less national pride, and a more moderate view toward equality compared to non-internet-using respondents. Post-stratification weighting helps narrow the gap between the Zhubajie sample and the internet-active subsamples of

**Table 2.** Comparisons on common attitudinal measures between Zhubajie (ZBJ) and benchmark samples.

	Interpersonal trust	Trust in central govt.	Trust in local govt.	National pride	Equality
ZBJ (unweighted)	0.75	0.74	0.36	0.76	0.57
ZBJ (census weights)	0.77	0.77	0.41	0.74	0.58
ZBJ (CNNIC weights)	0.77	0.76	0.37	0.77	0.60
<i>Benchmark (full sample)</i>					
ABS (weighted)	0.54	0.97	0.79	0.90	
CGSS	0.28				
CS (weighted)	0.54			0.90	0.89
WVS (weighted)	0.63	0.92		0.90	
<i>Benchmark (netizens)</i>					
ABS (weighted)	0.56	0.93		0.88	
CGSS	0.34				
CS (weighted)	0.61			0.90	0.87
WVS (weighted)	0.63	0.89	0.73	0.86	

Notes: CNNIC, China Internet Network Information Center; ABS, Asian Barometer Survey; CGSS, Chinese General Social Survey; CS, China Survey; and WVS, World Value Survey.

the benchmark surveys to some extent, but the differences are still statistically significant.<sup>14</sup>

One critical question is whether the differences between Zhubajie respondents and other survey respondents can be attributed to observable characteristics or are due to unobservable attributes that make them more likely to participate in a crowdsourcing platform. To explore this question more rigorously, we utilize the five questions as dependent variables and regress them on a set of demographic variables for the Zhubajie and the benchmark samples. Our expectation is that if the differences between Zhubajie and other survey respondents for these five common public opinion indicators are driven by observable demographics, then the estimated coefficients for demographics should be similar across regressions.

We run ordinary least squares regressions of the five attitudinal measures on a set of demographic variables and conduct coefficient equality tests (regression results are shown in Online Appendix B). We use the original values on the Likert scales in the regressions. The five dependent variables are not highly correlated except for the two variables regarding trust in national/local government ( $r = 0.53$ ) (see Table G in Online Appendix F.3). This gives us confidence that these variables are not measuring a common latent characteristic.

Table 3 presents the coefficient equality test results.<sup>15</sup> We see that in five out of 11 models, the coefficient estimates from the Zhubajie sample are not significantly different from those produced by the internet-active subsamples of the benchmark surveys. Post-stratification weighting helps improve the estimates.<sup>16</sup> With post-stratification weighting, only two out of 11 models produce significantly different coefficients between the Zhubajie sample and the internet-active subsamples of the benchmark surveys. Since the Zhubajie sample does not have respondents older than

aged 55, we repeat the analysis by excluding those respondents in the benchmark samples. The results with post-stratification weights remain the same.

Taken together, our results show that the Zhubajie sample with post-stratification weights can produce similar coefficients in most cases as the internet-user subsamples of the national surveys. This finding suggests that results based on crowdsourcing samples can be used to make inferences about the Chinese netizens.

### Political and social desirability biases in online surveys

One critical challenge of survey research is that respondents may conceal their beliefs and behavior, especially when it comes to politically and socially sensitive issues. In other words, respondents may choose to conform to social norms or political correctness by hiding their true actions or opinions, therefore invalidating statistical inferences (Blair and Imai, 2012; Bullock et al., 2011; Presser and Stinson, 1998). One crucial question is, then, how would online subjects respond to politically and socially sensitive issues in China where internet activity is highly monitored by the government?

To answer this question, we turn to the item count technique or list experiment that has become increasingly popular in social science research to elicit truthful answers to sensitive questions (e.g., Corstange, 2009; Gilens et al., 1998; Kuklinski et al., 1997). In list experiments, respondents are randomly assigned to a control or a treatment group. The control group contains a list of non-sensitive items. The treatment group is asked identical questions except for an additional sensitive item. Respondents in both groups are then asked to provide the total number of items on the list to which they answer affirmatively, not exactly

**Table 3.** Equality of estimated coefficients between Zhubajie and benchmark samples.

Survey	Dependent Variable	Zhubajie versus internet users				Zhubajie versus internet users (Age ≤ 55)			
		Unweighted		Weighted		Unweighted		Weighted	
		F-Stat	P	F-Stat	P	F-Stat	P	F-Stat	P
ABS	Trust national government	1.02	(0.41)	0.29	(0.94)	1.17	(0.32)	0.41	(0.87)
	Trust local government	1.54	(0.16)	0.85	(0.53)	1.93*	(0.07)	1.11	(0.35)
	Interpersonal trust	2.17**	(0.04)	1.72	(0.11)	1.76*	(0.10)	1.36	(0.23)
	National pride	3.30***	(0.00)	2.99***	(0.01)	3.38***	(0.00)	3.22***	(0.00)
WVS	Trust national government	2.06*	(0.06)	0.53	(0.78)	1.93*	(0.07)	0.62	(0.72)
	Interpersonal trust	1.07	(0.38)	0.24	(0.96)	0.87	(0.52)	0.21	(0.98)
	National pride	4.70***	(0.00)	2.34**	(0.03)	4.82***	(0.00)	2.43**	(0.02)
CS	Interpersonal trust	1.69*	(0.09)	1.37	(0.20)	1.61	(0.11)	1.26	(0.26)
	National pride	1.17	(0.31)	1.51	(0.14)	1.13	(0.34)	1.44	(0.17)
	Equality	0.48	(0.89)	1.36	(0.20)	0.58	(0.82)	1.38	(0.19)
CGSS	Interpersonal trust	3.58***	(0.00)	1.33	(0.22)	3.28***	(0.00)	1.16	(0.32)

Notes: \* significant at 10%; \*\* significant at 5%; \*\*\* significant at 1%; ABS, Asian Barometer Survey; WVS, World Value Survey; CS, China Survey; CGSS, Chinese General Social Survey.

which ones, such that their privacy is protected. Due to randomization, the difference in means between the control and treatment groups reflects the proportion of responses to the sensitive item, under certain assumptions (Blair and Imai, 2012).

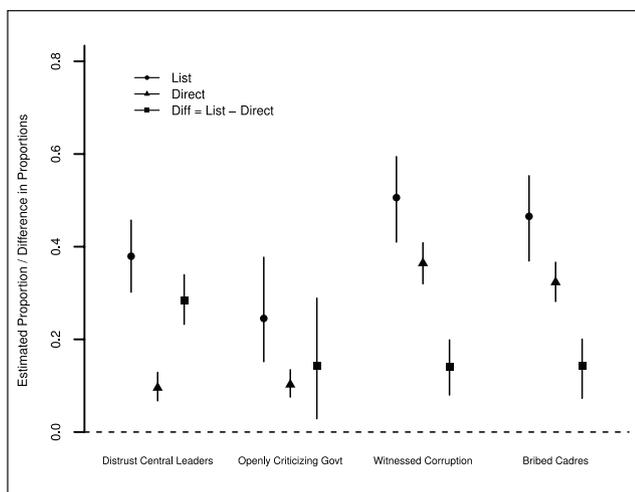
We embedded four list experiments to investigate response bias regarding politically and socially sensitive questions in online surveys. In the list experiments, we utilized the same sensitive items as in Tang (2016) and replaced some non-sensitive items to address potential ceiling effects.<sup>17</sup> The four sensitive items are distrusting central government leaders, openly criticizing central government leaders, witnessing government officials' corruption, and bribing government officials. The first two questions were to test political desirability biases, while the last two were to test social desirability biases.<sup>18</sup> For comparison, we also asked respondents direct questions regarding the sensitive items in the control groups after the list experiments. Following Blair and Imai (2012), we test design effects in the four list experiments and cannot reject the null hypothesis of no design effect.

Both political and social desirability biases are the tendency of survey respondents to hide their beliefs and behavior and provide answers that will be viewed more favorably by others. The issue of political desirability bias could be more pronounced in authoritarian countries because citizens might be fearful of getting into trouble by revealing truthful answers to politically sensitive questions (Tsai, 2010; Zhu, 1996). Anonymous self-administrated online surveys are supposed to reduce these response biases (Heerwegh, 2009; Kreuter et al., 2008). Yet, given the extensive government surveillance of online activities, the role of anonymous online surveys in alleviating political desirability bias might be limited in China.

Figure 1 presents an overview of political and social desirability biases across the four list experiments. For the two politically sensitive questions, we find that 38% and 25% of the respondents distrust and would openly criticize central government leaders in indirect questioning, compared to 10% in direct questioning for both cases. Our results show a relatively high level of political trust even among netizens, who tend to be more critical of the government (King et al., 2013). These results are consistent with many previous findings that the Chinese government enjoys high levels of trust among the public (e.g., Lewis-Beck et al., 2013; Shi, 2001; Tang, 2016; Wang, 2005). In addition, our results indicate a large political desirability bias for the distrusting central government leaders question and a moderate bias for the openly criticizing central leaders question. The differences between indirect and direct questioning are 28 and 14 percentage points, respectively, both of which are statistically significant.

For the two socially sensitive questions, in the two list experiments, about 51% of respondents reported that they witnessed corruption and 47% admitted they bribed government officials, in comparison to 36% and 32% in direct questions. The difference between indirect and direct questioning in our sample is approximately 14 percentage points in both cases, and both are statistically significant. Within our sample, the results show that respondents are more likely to hide truthful preferences in response to politically sensitive questions than socially sensitive ones.

In comparison, Tang (2016), based on a nationally representative sample, finds little political desirability bias: 4 percentage points for the question regarding distrusting central government leaders and 1 percentage points for openly criticizing central government leaders. The magnitudes of social desirability bias found in Tang (2016) are



**Figure 1.** Political and social desirability bias.

Notes: Plot of estimated proportions of respondents answering the sensitive item in the list experiments and the direct questions affirmatively as well as the differences between direct and indirect questioning. Solid circles, triangles, and squares indicate estimated proportions in the list experiments, proportions in the direct questions, and differences between indirect and direct questioning, respectively. Vertical lines are 95% confidence intervals.

quite similar to ours: 13 and 10 percentage points for the questions regarding witnessing government officials' corruption and bribing government officials, respectively. That our results show a larger political desirability bias than Tang's is surprising. Although the Chinese internet is subject to heavy governmental surveillance, we would still expect anonymous self-administrated online surveys to mitigate respondents' concern over political sensitivity compared to face-to-face surveys (Heerwegh, 2009; Kreuter et al., 2008). We should, however, emphasize that Tang's results may have underestimated the level of political desirability bias because of the ceiling effect in the list experiments, as acknowledged by the author himself (Tang, 2016: 147). As such, our study cannot yet provide a conclusive answer to whether anonymous self-administrated online surveys help alleviate response bias to politically sensitive questions in the Chinese context. It is important to further assess the advantage of online surveys relative to in-person surveys in mitigating response biases in future research.

## Concluding remarks

Our analyses yield three important findings regarding the external validity of online samples from Zhubajie. First, Zhubajie samples are much more representative of Chinese netizens than of the general population on a host of demographic and attitudinal measures. Second, Zhubajie samples with post-stratification weights can produce similar coefficients as the internet-active subsamples of national population-based surveys. In this sense, Zhubajie samples are useful and valid for making inferences about China's netizens.

Third, we show that respondents from Zhubajie are more likely to answer politically sensitive questions untruthfully than socially sensitive ones. While our sample comparisons are conducted in less than ideal situations, we believe that our findings demonstrate the usefulness of online samples, especially if researchers are interested in the public opinions of netizens, an important constituency for regime stability and leadership survival in China (Shirk, 2007).

We do offer some caveats about using Zhubajie samples. First, researchers should be careful when asking online subjects directly about politically sensitive issues, as direct approaches tend to yield more bias. Second, as of the present, Zhubajie subjects are not suitable for making inferences about the general population. That said, the validity of Zhubajie samples may improve as the pool of service providers increases and China's internet population continues to grow. Finally, Zhubajie subjects are more likely to be male, young, and highly educated; hence, similar to MTurkers, they may not be suitable for some research topics (Berinsky et al., 2012).

## Acknowledgements

Earlier versions of the paper were presented at the 2016 American Political Science Association mini-conference on Chinese Politics and University of California, San Diego, Global Policy and Strategy Junior Brownbag Seminar. We are grateful to Jesse Driscoll, Jean Oi, Eric Plutzer, Molly Roberts, Yiqing Xu, and conference and seminar participants for helpful comments and suggestions.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Supplementary material

The supplementary material is available at: <http://journals.sagepub.com/doi/suppl/10.1177/2053168018759127>.

## Notes

1. The external validity of MTurk samples has been studied extensively. See, for example, Berinsky et al. (2012), Clifford and Jerit (2014), Clifford et al. (2015), and Huff and Tingley (2015).
2. The search was performed on October 6, 2017.
3. We obtained institutional review board approval for this approach to subject recruitment on Witmart and Zhubajie.
4. With the "prevent-ballot-box-stuffing" option, Qualtrics places a cookie on the browser when respondents submit their answers to help prevent them from taking a survey multiple times.
5. In their sample Berinsky et al. (2012) found that a total of seven Internet Protocol addresses produced two responses

each to their demographic survey (i.e., 14 of 551 responses or 2.4% of the total responses). This is a slightly lower rate than our sample. Note, however, that this pattern does not provide conclusive evidence for repeat survey taking. As Berinsky et al. (2012) note, it could be the case that multiple people took the survey from the same large company, university, home, or coffee shop. We cannot rule out this possibility in our sample.

6. Inattentive respondents are included in the demographic comparisons. We took only the first bid from the same account, which eliminated 76 responses.
7. For example, the current administration's nationalistic rhetoric could either spawn greater nationalism among netizens or elicit a backlash. In addition, the demographics of the population may change over the eight years. Yet our coefficient equality tests suggest that the correlations between demographic variables and public attitudes remain relatively stable over time.
8. The Asian Barometer Survey did not ask questions on household registration and Chinese Communist Party membership. The World Value Survey had a question on household registration in the auxiliary version, not the publicly available one.
9. Population census is conducted every ten years. We exclude populations older than 55, since our Zhubajie sample has no respondents in this age category.
10. These weights are raked and based on the entire internet population because the China Statistical Report on Internet Development does not report fine-grained data by age and gender.
11. The publicly available Chinese General Social Survey (CGSS) dataset does not include weights. We use unweighted data for CGSS in all analyses, which may explain the larger discrepancy between the CGSS and the other three national surveys.
12. For the employment status variable, only the difference between the Zhubajie sample and the Chinese General Social Survey full sample is statistically significant.
13. Note that the exact demographic profile tends to vary by samples drawn from Zhubajie. For example, male respondents accounted for 54% of the sample in Huang (2015), 54–56% in Huang (2015), and 52% in Li and Zeng (2017). This suggests the importance of using post-stratification weights when analyzing Zhubajie samples.
14. One factor that may contribute to the differences in respondents' attitudes between the Zhubajie and the benchmark samples is timing. Ideally, we would like to implement the survey on Zhubajie in parallel with a national population-based survey to do benchmark comparisons. We leave this for future research.
15. We obtain similar results when excluding constants in the coefficient equality test (see Online Appendix B). We also conduct coefficient equality tests comparing the Zhubajie sample with the full benchmark samples. Results show that in most cases they produce significantly different point estimates (see Online Appendix C).
16. *P*-values increase in most cases. We experimented with different ways to construct post-stratification weights. It turns out that post-stratification weights based on age and gender yield the closest results to the benchmark samples.
17. Online Appendix F.2 presents the design of the four list experiments. *T*-tests suggest that the treatment and control groups in the four list experiments are largely balanced. See Online Appendix F.4 for sample balance checks.

18. Tang (2016: 144) noted that lying about witnessing corruption and bribing government officials are more of an issue of social desirability.

### Carnegie Corporation of New York Grant

This publication was made possible (in part) by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the authors.

### References

- Berinsky AJ, Huber GA and Lenz GS (2012) Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* 20(3): 351–368.
- Blair G and Imai K (2012) Statistical analysis of list experiments. *Political Analysis* 20(1): 47–77.
- Bullock W, Imai K and Shapiro JN (2011) Statistical analysis of endorsement experiments: Measuring support for militant groups in Pakistan. *Political Analysis* 19(4): 363–384.
- Clifford S and Jerit J (2014) Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science* 1(2): 120–131.
- Clifford S, Jewell RM and Waggoner PD (2015) Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics* 2(4): 1–9.
- Corstange D (2009) Sensitive questions, truthful answers? Modeling the list experiment with LISTIT. *Political Analysis* 17(1): 45–63.
- Gilens M, Sniderman PM and Kuklinski JH (1998) Affirmative action and the politics of realignment. *British Journal of Political Science* 28(1): 159–183.
- Heerwegh D (2009) Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research* 21(1): 111–121.
- Huang H (2015) International knowledge and domestic evaluations in a changing society: The case of China. *American Political Science Review* 109(3): 908–922.
- Huang H (2017) A war of (mis)information: The political effects of rumors and rumor rebuttals in an authoritarian country. *British Journal of Political Science* 47(2): 283–312.
- Huff C and Tingley D (2015) “Who are these people?” evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics* 2(3): 1–12.
- King G, Pan J and Roberts ME (2013) How censorship in China allows government criticism but silences collective expression. *American Political Science Review* 107(2): 326–343.
- Kreuter F, Presser S and Tourangeau R (2008) Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly* 72(5): 847–865.
- Kuklinski JH, Cobb MD and Gilens M (1997) Racial attitudes and the “New South.” *The Journal of Politics* 59(2): 323–349.
- Lewis-Beck MS, Tang W and Martini NF (2013) A Chinese popularity function. *Political Research Quarterly* 67(1): 16–25.

- Li X and Zeng K (2017) Individual preferences for FDI in developing countries: Experimental evidence from China. *Journal of Experimental Political Science* 4(3): 195–205.
- Oppenheimer DM, Meyvis T and Davidenko N (2009) Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology* 45(4): 867–872.
- Presser S and Stinson L (1998) Data collection and social desirability bias in self-reported religious attendance. *American Sociological Review* 63(1): 137–145.
- Shi T (2001) Cultural values and political trust: A comparison of the People's Republic of China and Taiwan. *Comparative Politics* 33(4): 401–419.
- Shirk SL (2007) *China: Fragile Superpower*. New York, NY: Oxford University Press.
- Tang W (2016) *Populist Authoritarianism: Chinese Political Culture and Regime Sustainability*. New York, NY: Oxford University Press.
- Tsai LL (2010) Quantitative research and issues of political sensitivity in rural China. In: Carlson A, Gallagher ME, Lieberthal K, et al. (eds) *Contemporary Chinese Politics: New Sources, Methods, and Field Strategies*. New York, NY: Cambridge University Press, pp.246–265.
- Wang Z (2005) Before the emergence of critical citizens: Economic development and political trust in China. *International Review of Sociology* 15(1): 155–171.
- Zhu J-H (1996) “I don't know” in public opinion surveys in China: Individual and contextual causes of item non-response. *Journal of Contemporary China* 5(12): 223–244.