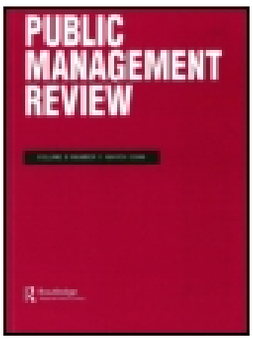

Public Management Review

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rpxm20

# Who can serve as the proxy for public employees in public administration experiments? a cross-sample comparison

Xiaoli Lu , Weijie Wang & Hao Xu

**To cite this article:** Xiaoli Lu , Weijie Wang & Hao Xu (2020): Who can serve as the proxy for public employees in public administration experiments? a cross-sample comparison, Public Management Review, DOI: 10.1080/14719037.2020.1864014

**To link to this article:** https://doi.org/10.1080/14719037.2020.1864014

PUBLIC MANAGEMENT REVIEW
https://doi.org/10.1080/14719037.2020.1864014

Routledge
Taylor & Francis Group



# Who can serve as the proxy for public employees in public administration experiments? a cross-sample comparison

Xiaoli Lu[a], Weijie Wang[b] and Hao Xu[a]

[a]Behavior and Data Science Lab, School of Public Policy and Management, Tsinghua University, Beijing, China; [b]Truman School of Public Affairs, University of Missouri, Columbia, MO, USA

**ABSTRACT**

This article examines whether convenience samples such as undergraduate students, MPA students, and online subjects can replicate the findings based on public officials in experimental research. We used a 2 × 2 factorial experimental design (High/Low Help-deservingness Clients × With/Without Situational Stress) with scenarios of discretionary decision-making by street-level bureaucrats in China. The four samples showed a consistent pattern in the impact of client help-deservingness on discretionary decision-making, but differed in the effects of situational stress on discretionary decision-making. We suggest that researchers be cautious in using convenience samples as surrogates for professional bureaucrats when the scenarios require professional expertise.

**KEYWORDS** Survey experiment; behavioural public administration; street-level bureaucrats; replication; discretion

## Introduction

Over the last decade, the public administration field has experienced a boom in the use of experimental methods in both research and practice (Li and Van Ryzin 2017). According to our recent review, 249 papers using experimental methods had been published across 21 public administration journals by the end of 2019.[1] As shown in Figure 1, the number of experimental publications has increased steadily since 2012, with an especially sharp increase over the last five years (from 21 papers in 2015 to 41 papers in 2019). In public administration practice, randomized controlled trials and ideas such as Behaviour Insights have been applied to public policy and management practices in various countries (European Union 2016).

In experimental publications, students and online participants have long been two major groups of experiment subjects. Of the 328 experiments reported in the 249 published articles,[2] 72 (21.95%) used students as experimental subjects, including 13 (3.96%) samples of MPA/MBA students (see Figure 2). A further 89 experiments (27.13%) used online panel samples, of which 33 (10.06%) recruited participants from the crowdsourcing platform MTurk and 56

**CONTACT** Weijie Wang wangweij@missouri.edu

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

Supplemental data for this article can be accessed at https://doi.org/10.1080/14719037.2020.1864014

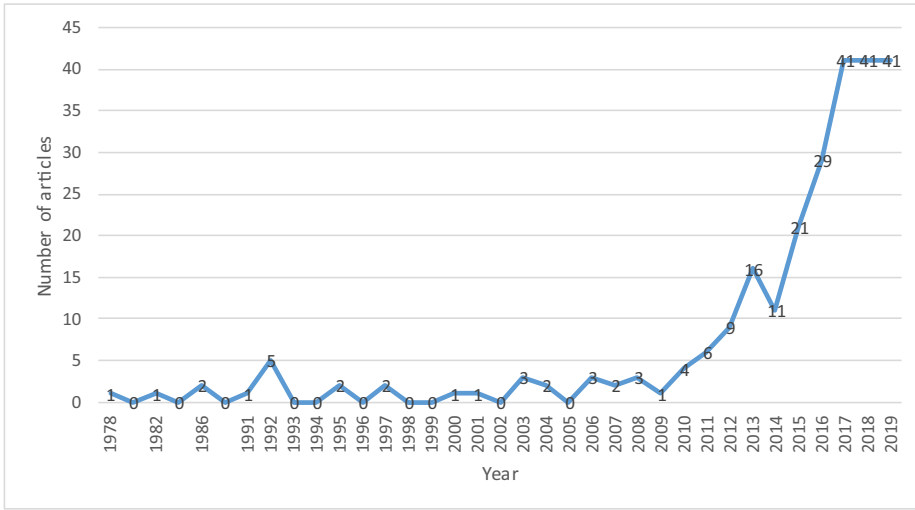


**Figure 1.** Trends of experimental publications in 21 PA journals by year (1978–2019).

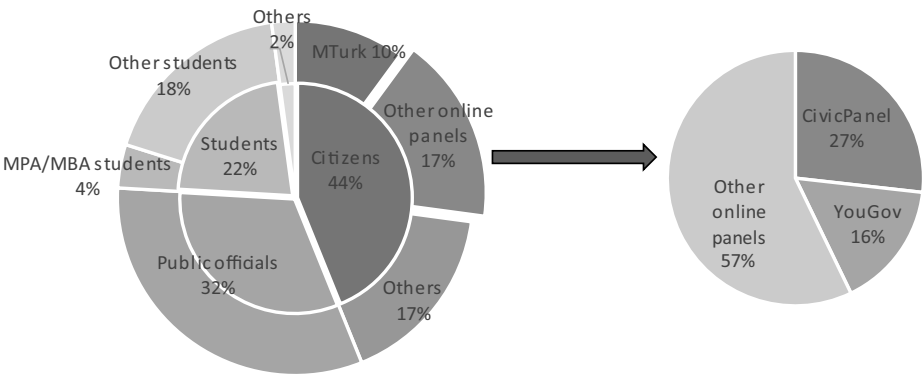


**Figure 2.** Subject distributions of experimental publications in 21 PA journals ($N$ = 328).

(17.07%) recruited from other platforms (including 15 from the CivicPanel project and 9 from YouGov). One hundred and five experiments (32.01%) used public officials as subjects.

Researchers in public administration commit to study the behaviour of public managers or policy-makers; however, gaining access to these professionals is notoriously difficult (Hermann and Ozkececi-Taner 2011). The use of online or student samples (including MPA/MBA students) to surrogate professionals in experimental studies has therefore become a popular option. In our dataset, 35 experiments (10.67%) used student samples to surrogate professionals in their experiment designs. Of these, ten used MPA/MBA students. Moreover, five experiments used online participants as surrogates (as shown in Figure 3).

The rapid increase in experimental research, however, has not been accompanied by replications (Walker, James, and Brewer 2017). Replication is critical to producing generalizable social science theories as it tests propositions in different contexts and with different populations. Human behaviours in experimental settings are affected by

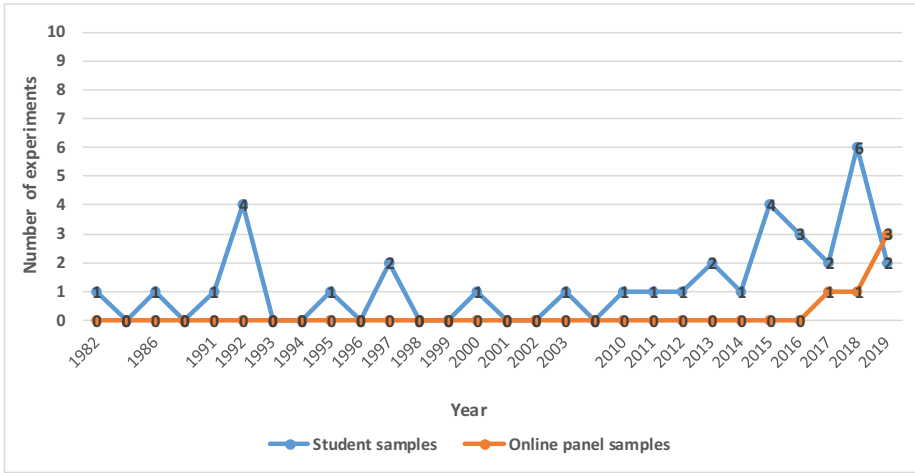


**Figure 3.** The distribution of experiments using student and online samples as surrogate professionals.

factors such as social desirability bias or the rules of the experiments, and subjects, especially those who self-select into experiments, may be systematically different from the targets that researchers hope to infer about (Levitt and List 2007). Replications are thus needed to test the external validity of findings. In public administration research, insufficient replication often leads to the criticism that experimental research lacks external validity (Mullinix et al. 2016; Krupnikov and Levine 2014). Replication efforts have been made to compare samples in given experimental scenarios in some research fields, such as political science, international relations, marketing, logistics, business management, and psychology. However, existing public administration research has provided little insight into whether online samples and student samples can replicate findings based on real public employees or policy-makers (James, Jilke, and Van Ryzin 2017; Stritch, Pedersen, and Taggart 2017; Walker, James, and Brewer 2017).

To address the question, this article replicates an experimental study conducted with street-level bureaucrats with commonly-used samples, including undergraduate students, MPA students and online participants. The objective is to examine if these convenience samples could lead to the same causal inference as the professional bureaucrat sample. Specifically, this article attempts to answer the following questions.

- Do undergraduate students, MPA students, and online respondents make discretionary decisions similar to street-level bureaucrats in a law enforcement scenario?
- Is the discretionary decision-making of undergraduate students, MPA students, online subjects, and real professionals affected by the theoretically-relevant factors in the same ways?

This paper makes several contributions to behavioural public administration research. First, it helps to address the deficit of replications. Although more scholars have recently replicate experimental studies conceptually or empirically in public management (Van Ryzin, Riccucci, and Li 2017; Filtenborg, Gaardboe, and Sigsgaard-Rasmussen 2017), replications in our field have still been limited. Walker, James, and Brewer (2017) concluded after a search of articles published in leading public administration journals

that 'the lack of replication of experimental public management research makes it urgent to address this deficit' (p.442). This study replicates an experiment with multiple new samples and provides valuable insights on research designs for future research. Second, it makes a methodological contribution by comparing experimental results between proxy samples and real street-level bureaucrats. To the best of our knowledge, this is the first study in public administration that compares how well convenience samples perform against a professional bureaucrat sample in causal analysis. Results show that the treatment effects based on convenience samples cannot replicate those based on the sample of professional street-level bureaucrats. The findings addressed some 'chronical concerns' in experimental public management research, one of which is whether MPA students can serve as proxies to public managers (Walker, James, and Brewer 2017). The findings thus carry important implications for selecting subjects in behavioural public administration research. Third, existing research that conducted cross-sample comparisons for evaluating external validity focused heavily on students and online participants from the U.S. or Europe, while there has been very limited attention to sample comparisons in other countries, including China, with a few exceptions such as Li, Shi, and Zhu (2018) and Li, Liang, Xu and Liu (2018). With data collected from students and online participants from China, this study also carries important implications for sample selection for future experimental research in China.

We begin with an examination of the existing findings from other social science fields using student and online samples as surrogates in experiments. Following this, we present the data source and data collection method for this study, and then report the findings. We conclude with lessons learned from the replications and suggestions for using convenience samples as proxies in experimental research to study public officials.

## Re-Examining the 'science of the sophomore'

Student samples have often been used in experimental studies in public management, political science, marketing, and other areas of social science because of their convenience and accessibility to researchers. Rosenthal and Rosnow (1969) even argued that, due to the wide use of students as experimental subjects, social experiments are a science of 'punctual college sophomore' volunteers. In public management research, students have often served as proxies for public employees or public managers. For example, MPA and undergraduate students have served as subjects in laboratory and survey experiments on a variety of topics, such as public employees' intrinsic motivations and public service motivations (Kroll and Porumbescu 2019), unethical behaviour (Belle and Cantarelli 2019), organizational citizenship behaviour (Jacobsen and Jensen 2017), individual performance (Anderson and Stritch 2015), prosocial behaviour (Esteve et al. 2016), and decision-making (Lee, Moon, and Kim 2017). The assumption is that these students, especially MPA students who mostly work in the public sector, have psychological processes similar to those of real public employees.

Although the external validity of student samples has been a major concern for researchers in different research areas, the evidence from existing research is mixed. A central topic is the extent to which student samples can produce experimental treatment effects similar to population-based samples (Mullinix et al. 2016). Among political science studies, Druckman and Kam (2011) showed that college students and the nonstudent general population were indistinguishable in some key covariates of interest, such as partisanship, ideology, views about homosexuality, and social trust.

Using Monte Carlo simulations, they further showed that student samples performed well if the treatment effect was homogenous, but using a student sample became problematic when the treatment effect was heterogeneous or the researchers failed to model a moderating effect. In experimental studies that tested how framing changes people's attitudes, Mullinix et al. (2016) found that student samples produced treatment effect estimates similar to those of population-based samples in three experiments. Similarly, Krupnikov and Levine (2014) conducted four parallel experiments with a student sample, an adult convenience sample from MTurk, and a nationally representative sample. Their findings suggest that the student sample performed better than the MTurk sample in replicating findings from the nationally representative sample, especially when the relevant moderators were taken into account.

Using student subjects is not without its problems, which have been well documented in the literature. Students are a relatively homogenous group (Lupton 2018), and they are not representative of the overall population in some social demographic dimensions, such as education, race, and age (Krupnikov and Levine 2014). Social psychologists have long suggested that college students differ from other people in systematic and marked ways (Sears 1986). For example, students tend to have unstable, changeable, weak, and inconsistent social and political attitudes, and usually have strong cognitive skills (Sears 1986). Student subjects are thus different from the populations to which they are generalized, such as public employees or public managers. The differences can occur in unmeasured ways, leading to biased estimations of treatment effects (Mullinix et al. 2016). In a survey experiment on the decision to approve the naturalization applications of immigrants in Switzerland, the student sample performed much worse than the nationally representative sample in recovering the qualitative pattern of the actual naturalization referendums (Hainmueller, Hangartner, and Yamamoto 2015). To investigate whether students can serve as proxies for professionals in decision-making in the counterterrorism area, Mintz, Redd, and Vedlitz (2006) compared the results of the same experiment conducted with a sample of military commanders and a sample of students. They found that the students made markedly different choices compared with the military commanders, and that the students used more information in their decision-making processes and were more likely to adopt a maximizing decision-making strategy. To the best of our knowledge, there has not been any studies that compared how well student samples perform in comparison with professional bureaucrats in experimental public management research. Walker, James, and Brewer (2017) considers this as a 'chronical concern' that should be addressed.

In business research, researchers have compared students (including working adults in part-time education programmes) with consumers (Jones and Sonner 2001), investors (Elliott et al. 2007; Liyanarachchi 2007), line managers (Remus 1986), loan officers (Abdel-Khalik 1974), and real managers (Hughes and Gibson 1991). The findings have again been mixed. Jones and Sonner's (2001) consumer studies revealed that a traditional student sample could not surrogate real customers, while part-time working adult students could. Remus's (1986) experimental research on production scheduling decisions confirmed that MBA students could surrogate line managers in their decisions. In contrast, neither Abdel-Khalik's (1974) loan evaluation and decision-making experiment nor Hughes and Gibson (1991) decision experiments on adopting a decision support system delivered positive results, with MBA students failing to replicate loan officers' or real managers' decisions. In two financial

accounting experiments conducted by Elliott et al. (2007), MBA students served as a good proxy for investors when the tasks had a low level of integrative complexity.

What conclusions can we draw regarding the use of student subjects in experimental studies based on the mixed findings in the current literature? In experimental studies in which student samples performed as well as population-based samples, the student subjects were typically not asked to play a role for which they were not well equipped. For example, Mullinix et al. (2016) and Krupnikov and Levine (2014) tested how differences in the framing of political issues altered citizens' attitudes. The target population was the general public, of which students are a part. In contrast, in studies that found that students samples did not perform well, the student subjects were asked to play the role of professionals who would be equipped with professional experience, skills, or knowledge that these students did not have (Mintz, Redd, and Vedlitz 2006). In these cases, student subjects did not match the population to which generalization was intended. Therefore, researchers need to be cautious when student subjects are used as proxies for professionals or elites.

## Online panel data

Since their emergence in the late 1990s, various online platforms have become prevalent sources of participants for experimental research (Li, Kuo, and Rusell 1999). These platforms include Amazon's Mechanical Turk (MTurk), YouGov, InnoCentive, Threadless, Lánzanos, iStockPhoto, ModCloth, Fiat Mio, and StudyResponse. There are also a few popular online platforms in China, such as QQ Survey, SoJump, and Diaochapai. Using online recruitment, researchers ask participants to complete human intelligence tasks, with participants receiving compensation after their completed tasks are verified by the requesting researchers (Buhrmester, Kwang, and Gosling 2011).

Debate is ongoing regarding the use of online panel data based on marketplace crowdsourcing in western world (Mullinix et al. 2016; Walter et al. 2019). Some researchers have held that online platforms provide quick access to a large and diversified population at a low cost (Berinsky, Huber, and Lenz 2017; Buhrmester, Kwang, and Gosling 2011; Paolacci and Chandler 2014).[3] Online platforms can also provide unique opportunities to study 'hard to reach populations' such as lesbian, gay, bisexual, and transgender individuals (Stritch, Pedersen, and Taggart 2017; Smith et al. 2015). In political science studies, MTurk can attract more young Hispanic females and young Asians (Huff and Tingley 2015). Moreover, it is easier to study some sensitive topics, such as workplace violence, discrimination, and abusive supervision, on these online platforms than in face-to-face experiments (Porter et al. 2019).

Empirically, some studies have demonstrated that participants on online platforms were equivalent to, or even more representative than, traditional data sources, such as student and population-based samples, in dimensions like age and socioeconomic background (Berinsky, Huber, and Lenz 2017; Casler, Bickel, and Hackett 2013; Stritch, Pedersen, and Taggart 2017; Levay, Freese, and Druckman 2016; Buhrmester, Kwang, and Gosling 2011; Johnson and Borden 2012). Walter et al.'s (2019) meta-analysis of 90 independent samples and 32,121 participants in psychology and business research confirmed that online panel data had 'similar psychometric properties and produces criterion validities that generally fall within the credibility intervals of existing meta-analytic results from conventionally sourced data' (425). Casler, Bickel, and Hackett (2013) conducted a test of a simple psychological tool selection scenario and found no difference between

the results based on samples from MTurk, social media, and face-to-face groups. In their test of the 'big five' personality traits, Feitosa, Joseph, and Newman (2015) found that MTurk data were as effective as those collected from students or organizational employees, but this finding only applied to IP addresses from native English-speaking countries. Based on samples from the U.S., Clifford, Jewell, and Waggoner (2015) confirmed that samples from MTurk could mirror two benchmark national samples collected online and face-to-face in terms of measuring political ideology.

In contrast, other scholars have expressed concerns over the representativeness of samples recruited from crowdsourcing platforms. The major concern is that the composition of online populations and their motivations for participating in these online activities are unknown, and might make them different from populations without access to the Internet (Krupnikov and Levine 2014; Paolacci and Chandler 2014; Li, Shi, and Zhu 2018). Some researchers have obtained preliminary findings regarding the composition of MTurk samples. For instance, in their comparison with a non-Internet population, Paolacci and Chandler (2014) found that online samples tended to be younger, better educated, underemployed, less religious, and more liberal. Pew's (2016) report based on 3,370 MTurkers confirmed that they were younger and better educated than the general working adult population in the U.S. (Hitlin 2016). Berinsky, Huber, and Lenz (2017) recently pointed out that MTurk samples were younger than non-student samples and had a higher education level. In terms of geographical distribution, MTurk participants are mostly based in the U.S. and India (Marvit 2014; Huff and Tingley 2015; Hitlin 2016).[4]

Similar to student samples, another major challenge with online panel data is that the participants are naïve to experimental tasks. Some tasks are based on respondents' work experience or expertise; however, online subjects may be unfamiliar with the tasks or their contexts. Under these conditions, it is difficult for the treatment to trigger the necessary stimulus for the subjects. Porter et al. (2019) noted that it is best to ask general online participants to complete tasks that do not require particular knowledge, skills, or abilities.

An additional challenge is the repeated participation of some survey takers on these online platforms (known as habitual survey takers or professional survey takers). Pew's report indicated that around 63% of MTurkers performed a task every day (Hitlin 2016). Repeated participants have the potential of becoming savvier over time by learning from previous experiment experience, resulting in a reduction of effect size through experimental manipulation (Krupnikov and Levine 2014). Chandler et al.'s (2015) findings based on a replication of classical decision experiments on the MTurk platform confirmed that the use of non-naïve participants tended to reduce effect size. In their studies of political attitudes, based on a survey from the 2010 YouGov Cooperative Congressional Election Study, Hillygus, Jackson, and Young (2014) found that repeat participants tended to respond to survey questions in a less thoughtful manner than the population benchmark.

Last but not least, data collected from online crowdsourcing platforms could be contaminated by cheating behaviours. Psychologists began to worry about the data quality of MTurk in August 2018, when cheating behaviour was first exposed with some evidence of nonsense answers to open-ended questions and respondents with duplicate GPS locations (Bai 2018; Dreyfuss 2018; Stokel-Walker 2018). Ryan's (2018) analysis of his own data showed that at least 9.38% of the responses were fraudulent. Kennedy et al.'s (2018) analysis demonstrated that HIT approval rate was not a reliable

indicator for quality control. They found that even when researchers set a standard 95% HIT approval rate in their procedures, the percentage of fraudulent respondents who used a VPN or non-US IP address could sometimes reach 20%.

## Data and Method

Originally, we conducted a vignette-based survey experiment among *Chengguan* officers who are urban law enforcement professionals in two Chinese cities (See Appendix 4 of the supplementary material for experimental design and vignettes description). The experimental vignettes were discretionary decision-making scenarios that *Chengguan* officers often face in their law-enforcement encounters with clients (specifically, regulating street vendors). Using a 2 × 2 factorial and between-subjects experimental design, we aimed to examine the impact of client help deservingness, crowd situational stress, and their interaction effects on officers' discretionary decision making. As shown in Appendix 3 of the supplementary material, we used age difference (with an old man standing for a client with high help deservingness and a young man standing for a client with low help deservingness) to operationalize the client help-deservingness cue and the presence or absence of bystanders gathering at the scene to operationalize the crowd situational stress cue. Discretion was measured by the size of the fine imposed on the vendors.

We were given permission to collect data from two cities' Urban Management and Law Enforcement Bureaus. We invited *Chengguan* officers to join the survey via their official social media groups (including WeChat and QQ). The respondents were randomized into one of the four treatment conditions via Lediaocha, a Chinese e-survey platform. We received 467 responses in total and 442 effective responses were retained after data cleaning (as shown in Appendix 3 of the supplementary material). The original experiment showed that clients' help deservingness affects their discretionary decision making, and *Chengguan* officers tend to impose a smaller fine to clients that were perceived as deserving of help. On the other hand, situational stress in the form of the presence of bystanders alone did not influence *Chengguan* officers' discretionary decision making, but it weakens the effect of clients' help deservingness (Lu, Xu, and Wang 2019).

We then replicated the same survey experiment with three samples that have often been used in experimental research in public management: undergraduate students, MPA students, and online participants. If the convenience samples could serve as valid proxies, we should be able to arrive at the same causal inference in that the factors that influenced discretionary decision-making should stay the same, and the differences in the amounts of fine imposed should not be statistically significant.

To restrict the compliance pressure generated in classroom settings (Sears 1986), we posted our survey link in their official WeChat groups of MPA students and undergraduate students majoring in public administration, business administration, and economics at a public university in eastern China. To improve response rates, we asked our contacts in these WeChat groups to send several reminders during the study period. Randomization was again conducted in Lediaocha. Similar to previous cross-sample analyses, such as Krupnikov and Levine (2014), we attempted to match the sample sizes with the benchmark sample to achieve similar statistical power. We received 443 valid responses from the undergraduates and 297 from the MPA students.

The online subjects were recruited from a Chinese crowdsourcing platform, SoJump.[5] SoJump is similar to Amazon Mechanical Turk as a platform for recruiting

online questionnaire respondents, but does not provide other recruitment services. When we recruited online participants, we paid the SoJump platform 3 Chinese yuan (about US $0.43) per effective response. According to SoJump, the platform has 2.6 million registered online users, of whom 1 million are daily active users.[6] SoJump's released data show that 52% of its registered subjects are male, 29.34% belong to the 26–30 age group, 39.2% are working professionals, and 26.3% are students (as shown in Appendix 2 of the supplementary material). Most respondents are based in economically developed provinces, such as Guangdong (14.81%), Beijing (10.73%), Shanghai (7.73%), Zhejiang (6.85%), Jiangsu (6.32%), Shandong (5.24%) (as shown in Figure 2 in Appendix 2 of the supplementary material). SoJump respondents are even younger than MTurkers.

We used a manipulation check in the questionnaire design to improve response validity. 73 respondents(13.85%) failed to pass the check and were dropped from our final analysis.[7] In total, we received 454 responses from SoJump and excluded 6 of them in the data cleaning process (the procedures for which are shown in Appendix 3 of the supplementary material). Finally, we arrived at 448 valid responses. Dropping responses that failed a manipulation check following treatment assignment may cause bias (Aronow, Baron, and Pinson 2019). To guard against that, we conducted robustness tests with a sample that included the dropped responses, and the results are highly consistent with main results presented in Table 3. Details of the robustness tests are presented in Appendix 8 of the supplementary material.

We compared the demographic differences between our samples and data from two national-level censuses: the Statistical Report on Internet Development in China produced by the China Internet Network Information Centre (CNNIC) and the Sixth National Population Census (as shown in Appendix 1 of the supplementary material). First, the *Chengguan* officer sample was older than the other samples, while the undergraduate student sample was much younger than the other samples. Second, the majority of the *Chengguan* officers were male (88.6%), and the proportion of males was much higher than in the other samples, while the undergraduate student sample and the MPA student sample had more females (66.59% and 57.91% respectively) than both benchmarks. Third, the *Chengguan* officer sample was better educated than the online sample and the national benchmark, and the education level of the SoJump online sample was quite similar to that of MTurk in the U.S., with the majority of the workers having received higher education. Moreover, we expected the online participant sample to have similar demographics to the netizens; however, the results indicate that the online participant sample was better educated and had a higher employment rate than the general Internet population.

To ensure the homogeneity of the different treatment groups, we checked the differences in all of the available background variables between the four sampled groups. As shown in Appendix 5 of the supplementary material, the group difference tests on gender, age, and familiarity with *Chengguan* were all insignificant (at the 0.10 level). Therefore, the randomization of the different treatment groups was effective.

## Results and Analysis

In this section, we present a series of comparisons between the proxy samples and the benchmark sample. First, we compared the fine amounts imposed by each group under

each treatment condition. Second, we compared whether the factors that had statistically significant effects on the discretionary decision-making of the four samples were different.

## Comparison of the four samples

Figure 4 presents the average fine amounts imposed by each sample under the four treatment conditions. The first three columns in Table 1 present the *t*-test results comparing each of the three proxy samples with the benchmark group (*Chengguan* officers). The results reveal whether the average fine amount imposed by each specific group was significantly different from that imposed by the *Chengguan* officers. The last column presents the results of a cross-sample analysis of variance (ANOVA) test to determine whether the average fine amounts across the four samples were significantly different.

Under Treatment Condition 1 (low help-deservingness and no situational stress), there was a statistically significant difference between the four groups as determined by

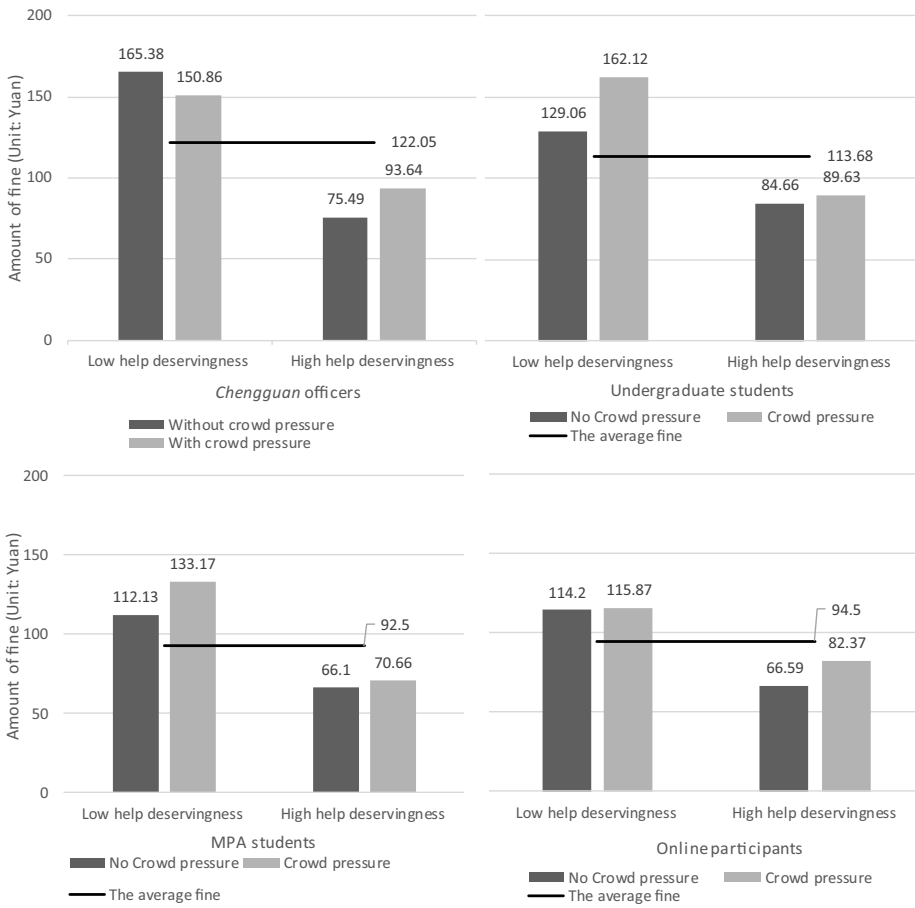


**Figure 4.** The average fines imposed by the four samples under the four treatment conditions.

**Table 1.** Summary of results of *t*-tests and ANOVA.

| Sample Treatment condition | Undergraduate | MPA | Online | ANOVA |
|---|---|---|---|---|
| 1 Low help deservingness/no situational stress | $t = 3.346$***, $p < .01$ | $t = 4.084$***, $p < .01$ | $t = 4.726$ ***, $p < .01$ | $F = 11.153$, $p < .01$ |
| 2 High help deservingness/no situational stress | $t = -0.983$. $p = .327$ | $t = 0.993$, $p = .322$ | $t = 1.135$, $p = .258$ | $F = 2.037$, $p = .108$ |
| 3 Low help deservingness/ situational stress | $t = -0.943$, $p = .347$ | $t = 1.379$, $p = .170$ | $t = 3.115$ ***, $p < .01$ | $F = 6.365$, $p < .01$ |
| 4 High help deservingness/ situational stress | $t = 0.602$, $p = .548$ | $t = 2.391$**, $p < .05$ | $t = 1.167$, $p = .244$ | $F = 1.926$, $p = .125$ |

*t*-tests are between each proxy sample and the *Chengguan* sample. *p < .1; **p < .05; ***p < .01.

ANOVA ($F = 11.153$, $p < .01$). Our *t*-test results showed that all of the proxy samples imposed statistically significantly lower fine amounts than that of the benchmark sample. This suggest that the undergraduate students, MPA students, and online participants were all more lenient in their decisions under this condition. Specifically, the MPA students were the most lenient group in their discretionary decision-making, followed by the online respondents and then the undergraduate students.

Under Treatment Condition 2 (high help deservingness and no situational stress), there were no statistically significant differences between groups in mean fine amounts as determined by ANOVA ($F = 2.037$, $p = .108$). The differences in the average fine amounts between the proxy samples and *Chengguan* officers were also not statistically significant.

When the treatment condition was low help deservingness with the presence of bystanders (Treatment Condition 3), the ANOVA result ($F = 6.365$, $p < .01$) suggests that there was a statistically significant difference between the four groups. Further *t*-test results showed that the online respondents imposed significantly lower fines than that of the benchmark group ($t = 3.115$, $p < .01$).

Under Treatment Condition 4 (high help deservingness with the presence of bystanders), there were no statistically significant differences between groups in mean fine amount as determined by ANOVA ($F = 1.926$, $p = .125$). The *t*-test results indicated that the MPA respondents imposed significantly smaller fines on vendors than the benchmark group.

## *The effect of client help deservingness and crowd situational stress on discretionary decision-making*

The results of our original survey experiment on *Chengguan* officers serve as the benchmark in this study. To briefly summarize the benchmark results, *Chengguan* officers gave old vendors who were perceived as deserving of help smaller fines than they gave to young vendors, which is consistent with previous research. The difference was statistically significant at the 1% level. Countering conventional thinking, the presence of bystanders alone did not have a statistically significant effect on the amount of the fine imposed. The impact of the interaction term between the above two factors on discretionary decision was statistically significant, suggesting that the effect of help deservingness was contingent on the presence of bystanders. Specifically, the

Table 2. Summary of replication findings across samples.

| | Original experiment | Experimental replication | | |
|---|---|---|---|---|
| | *Chengguan* officers ($n$ = 422) | Students ($n$ = 443) | MPA students ($n$ = 297) | Online participants ($n$ = 448) |
| H1 Help deservingness→discretion [a] | - *** | - *** | - *** | - *** |
| H2 Crowd pressure→discretion | n.s. | + *** | + * | n.s. |
| H3 Interaction effect | + ** | - * | n.s. | n.s. |

*p < .1; **p < .05; ***p < .01. n.s.: Non-significant. '+': positive effect; '-': negative effect. [a] the fine amount (Chinese yuan). Based on regression results without controls.

*Chengguan* officers imposed a larger fine on old vendors if there were bystanders watching, but were likely to fine young vendors less in the presence of bystanders.

Using the same research design and treatment approach, three samples that have been often used as proxies for street-level bureaucrats in public management research produced mixed findings when compared with the original sample of street-level bureaucrats (as shown in Table 2). The only consistent finding was the impact of clients' help deservingness on respondents' discretionary decision making. None of the three proxy samples produced the same pattern of statistical significance for other causal effects.

### Client help deservingness cue

Table 3 presents the detailed results of the regression analyses. We used effect coding to analyse the factorial experimental data to ensure that the regression coefficients would be equivalent to the classically defined main effects and interactions. We coded street vendors with high help deservingness as 0.5 and street vendors with low help deservingness as −0.5. We coded the situational stress cue in a similar way, with 0.5 for situational stress and −0.5 for without situational stress. As Table 3 shows, the help deservingness cue had a negative and statistically significant effect on the fine amounts across all of the samples, which means that all of the respondents tended to be more lenient when they faced a help-deserving street vendor. However, the main effects of help deservingness were different, and all of the proxy samples returned smaller effect sizes than the benchmark sample, consistent with the results of the previous ANOVA test.

### Crowd pressure cue

The four samples responded very differently to the crowd pressure cue. For the original *Chengguan* officers, the presence of bystanders alone did not have a statistically significant effect on their discretionary decisions. The online sample had the same response as the *Chengguan* officers, in that bystanders did not produce statistically significant effects. However, both undergraduate and MPA respondents tended to impose larger fines and to be less lenient if there were bystanders watching.

### Interaction effect

The interaction effect between the crowd pressure cue and help deservingness cue was statistically significant for the *Chengguan* officers. As Figure 4 shows, they tended to

**Table 3.** Regression results for the main and interaction effects of the treatments on the discretion of the four sampled groups.

| | Discretion (*Chengguan* officers) | | | Discretion (undergraduate students) | | | Discretion (MPA students) | | | Discretion (online panel sample) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Model 1–1 | Model 1–2 | Model 1–3 | Model 2–1 | Model 2–2 | Model 2–3 | Model 3–1 | Model 3–2 | Model 3–3 | Model 4–1 | Model 4–2 | Model 4–3 |
| F1 Help deservingness | −74.403*** | −73.554*** | −73.729*** | −57.932*** | −58.453*** | −58.119*** | −53.722*** | −54.270*** | −54.971*** | −40.587*** | −40.553*** | −39.692*** |
| | (8.074) | (8.055) | (8.000) | (7.315) | (7.297) | (7.343) | (7.329) | (7.342) | (7.441) | (6.086) | (6.084) | (5.970) |
| F2 Situational stress | 1.864 | 1.821 | 0.982 | 17.986** | 19.019*** | 19.063*** | 12.001 | 12.799* | 12.253* | 8.820 | 8.725 | 8.262 |
| | (8.085) | (8.055) | (8.033) | (7.300) | (7.297) | (7.314) | (7.310) | (7.342) | (7.382) | (6.085) | (6.084) | (5.966) |
| Interaction (F1 × F2) | | 32.677** | 32.012** | | −28.092* | −26.907* | | −16.468 | −15.344 | | 14.100 | 9.799 |
| | | (16.109) | (15.984) | | (14.594) | (14.652) | | (14.684) | (14.842) | | (12.167) | (11.961) |
| Control variables | | | | | | | | | | | | |
| Gender | | | −4.999 | | | −6.247 | | | −7.712 | | | −10.672* |
| (Male = 1) | | | (12.601) | | | (7.768) | | | (7.617) | | | (6.112) |
| Age | | | 0.326 | | | −1.651 | | | −1.479 | | | −1.630*** |
| | | | (1.200) | | | (2.130) | | | (1.720) | | | (0.387) |
| Tenure | | | 0.166 | | | – – | | | – – | | | – – |
| | | | (1.230) | | | | | | | | | |
| City (City X = 1; City Y = 0) | | | −32.680*** | | | – – | | | – – | | | – – |
| | | | (11.449) | | | | | | | | | |
| Familiarity with *Chengguan* | | | – – | | | 2.276 | | | 3.729 | | | 5.025 |
| | | | | | | (3.862) | | | (3.149) | | | (3.494) |
| Tenure in public sector | | | – – | | | – – | | | 1.036 | | | – – |
| | | | | | | | | | (1.637) | | | |
| Education level | | | – – | | | – – | | | – – | | | −0.581 |
| | | | | | | | | | | | | (4.644) |
| Constant | 121.922*** | 121.342*** | 140.834*** | 115.970*** | 116.368*** | 146.429*** | 95.287*** | 95.515*** | 125.476*** | 94.789*** | 94.757*** | 136.344*** |
| | (4.032) | (4.027) | (36.195) | (3.654) | (3.649) | (44.091) | (3.667) | (3.671) | (45.156) | (3.043) | (3.042) | (22.921) |
| R-squared | 0.169 | 0.177 | 0.199 | 0.132 | 0.139 | 0.142 | 0.159 | 0.162 | 0.171 | 0.094 | 0.097 | 0.145 |
| *N* | 422 | 422 | 422 | 443 | 443 | 443 | 297 | 297 | 297 | 448 | 448 | 448 |

*p < .1; **p < .05; ***p < .01. Unstandardized coefficients are reported. Standard errors in parentheses

impose a larger fine on old vendors if there were bystanders watching, while they were likely to fine young vendors less if there were bystanders. The presence of bystanders thus changed their discretionary decision making in enforcing laws. The interaction effect was also statistically significant for the undergraduate students, but the direction of the effect was the opposite of that for the *Chengguan* sample. The interaction term suggests that undergraduate respondents tended to impose a larger fine on low help-deserving vendors if they were being watched by bystanders. For the MPA respondents and online respondents, the interaction effects were not statistically significant, meaning that the effect of help deservingness did not depend on the situational factor, or vice versa. Figure 5 visualizes the interaction effect for each group.

It could be argued that random students and online participants may not be familiar with the context in which street-level bureaucrats work. Respondents who are familiar with the context may have psychological processes similar to *Chengguan* officers. Are respondents who are familiar with the context therefore more legitimate surrogates for street-level bureaucrats? To test this, we restricted our online sample to respondents

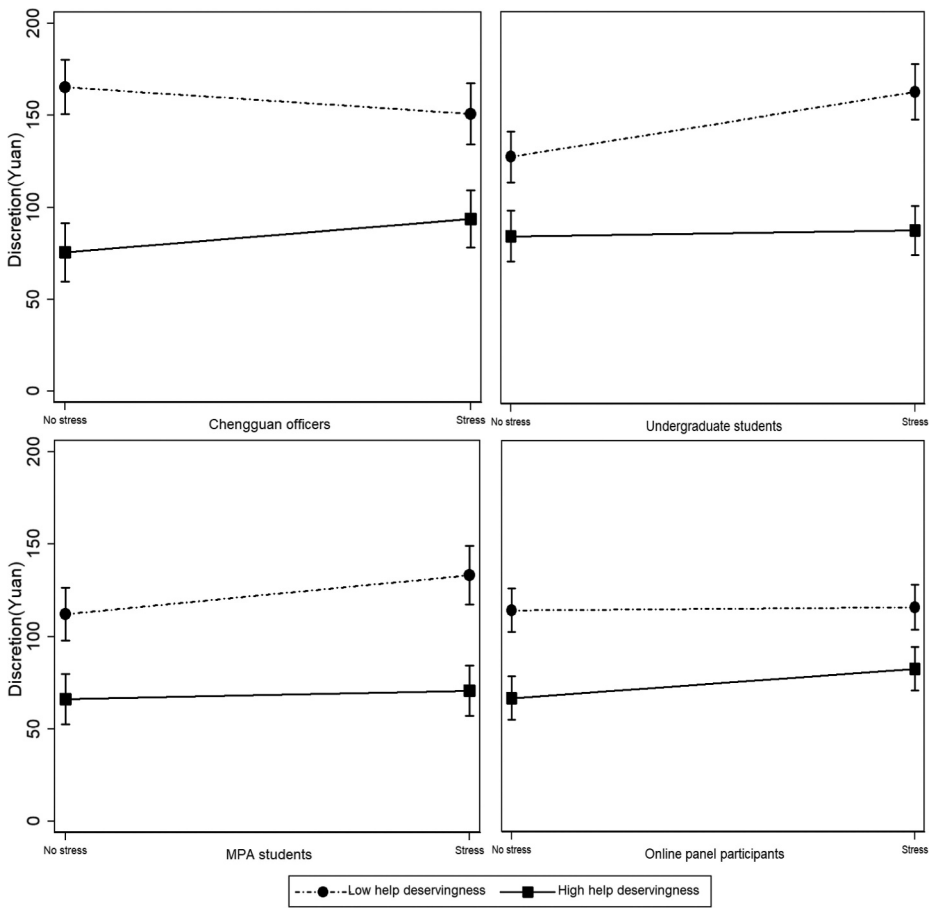


**Figure 5.** Two-way interaction between situational stress and help deservingness of the four sampled groups (analysis without controls; with 95% confidence intervals).

**Table 4.** Regression results for the restricted online samples[1].

| | Discretion (Online panel sample) | | |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| F1 Help deservingness | −33.889*** | −34.128*** | −32.868*** |
| | (8.432) | (4.213) | (8.328) |
| F2 Situational stress | 6.486 | 6.634 | 5.429 |
| | (8.433) | (8.426) | (8.328) |
| Interaction (F1 × F2) | | −19.852 | −23.049 |
| | | (16.852) | (16.729) |
| Control variables | | | |
| Gender (Male = 1) | | | −6.965 |
| | | | (8.558) |
| Age | | | −1.263** |
| | | | (0.543) |
| Constant | 93.527*** | 93.961*** | 138.605*** |
| | (4.201) | (4.213) | (17.734) |
| R-squared | 0.084 | 0.091 | 0.126 |
| *N* | 182 | 182 | 182 |

Note: [1]Analysis restricted to online respondents who claimed that they were very familiar or familiar with *Chengguan* working practices. *p < .1; **p < .05; ***p < .01. Unstandardized coefficients are reported. Standard errors in parentheses.

who claimed they were familiar with the work practices of *Chengguan* officers and ran the regressions again. The results from this restricted online sample, which are presented in Table 4, showed no major differences in the patterns of statistical significance or the fine amounts, thus indicating that self-claimed familiarity with the context did not make these respondents behave more like real *Chengguan* officers. We were not able to perform a similar analysis for the undergraduate sample because only about 60 students replied that they were familiar with the work of *Chengguan* officers, and regression analyses with such a small sample would not have been robust.

## *When can we use students, MPA students, or online samples as surrogates for professionals in public administration research?*

This section provides a few lessons based on the comparison of the four subject groups, which might be useful for future experimental designs, especially for selecting subjects as surrogates for public professionals.

Undergraduate students, MPA students, and online subjects are not perfect surrogates for *Chengguan* officers in the tests of decision making. The primary assumption behind using convenience samples as surrogates is that these samples show similar psychological processes to real-world professionals, such as the *Chengguan* officers in this study. Among the proxy samples we analysed, MPA students are mostly adults with work experience in the public sector, which is a major reason why they are often used as proxies for public officials. Our research results bring both good and bad news for experimental researchers who often use proxy samples.

The three proxy groups and the original sample converged in one respect: vendors with high help-deservingness tended to receive a smaller fine. This is consistent with findings from other contexts that street-level bureaucrats tend to prioritize clients with certain characteristics, especially those who are perceived as deserving of help (Jilke and Tummers 2018). This suggests that there may be a few strong psychological

mechanisms that are shared not just by the proxy samples and the original *Chengguan* officers but also street-level bureaucrats in other contexts.

The more discouraging finding is that the interaction effect identified in the benchmark group, which challenged conventional thinking, was not reported in any of the three proxy groups. One possible explanation for this contradiction is that naiveté matters. Situational stress in law enforcement is a kind of psychological feeling that laymen cannot easily relate to. This finding brought us the first lesson in using proxies for public professionals:

*Lesson 1: Even though public professionals might share some psychological mechanisms with students and with online recruited labour, we should be very cautious in using convenience samples as surrogates for public professionals, especially when there are psychological feelings involved that laymen cannot easily relate to.*

Our study also examined the subjects' familiarity with the working practices of *Chengguan* officers. The experimental results based on the subsample of online participants who claimed that they were either familiar or very familiar with *Chengguan* law enforcement were no different to the findings based on the entire sample. Moreover, the interaction effect found among the original *Chengguan* officers could not be replicated in the subsample. This suggests that claimed familiarity is not a reliable indicator to help select subjects, which leads to the second lesson:

*Lesson 2: Claimed familiarity with professionals' working practices cannot be simply used as a criterion in selecting subjects to serve as surrogates for public professionals.*

This study highlighted a few problems with a Chinese online crowdsourcing platform (SoJump) that might need to be considered in the external generalization of research findings. Online subjects in China share similar demographic characteristics with their counterparts in other parts of the world. Specifically, online subjects tend to be better educated than the general body of Internet users and the general population. When compared with the surrogated population, the online sample was more gender balanced and younger, and had more diversified age groups. Our analysis of the online sample also showed that gender and age had a significant impact on the amount of fine imposed on vendors. Furthermore, we conducted regressions of different groups of online participants (males, females, and different age groups), and the results did not indicate differences between these groups and the entire online sample (as shown in Appendix 5 of the supplementary material).

Our manipulation check in the online data collection process indicated that nearly 14% of the respondents could not pass the single manipulation check question, which is much higher than the reported 9% of failures based on multiple questions in the U.S. (Ryan 2018). This provides a warning that cheating behaviours do exist in the Chinese platform, which might contaminate the data if there is no manipulation check. It is thus necessary to use manipulation checks in experimental research. However, researchers should be careful when dealing with responses that failed manipulation checks. Aronow, Baron, and Pinson (2019) warned that dropping responses that failed manipulation checks following treatment assignment might lead to biased estimates. The mechanism of how dropping subjects induces biased estimates is still underexplored. At least, this informs us that we cannot automatically drop subjects who failed manipulation checks in future research, and we need some robustness tests to compare results with and without dropped responses to guard against bias.

*Lesson 3: Online platforms can provide more diversified samples than most data sources*

*Lesson 4: Data contamination is a serious problem that might impact the effectiveness of online samples. The design of manipulation check is necessary, but researchers should be careful in dealing with responses that failed manipulation checks.*

## Conclusion

Different forms of replication are important for social science research because they serve multiple purposes: they help not only check internal validity of previous findings but also establish external validity by examining whether previous findings hold across populations and contexts (Tsang and Kwan 1999). In this study, we use replication to address an important problem for behavioural public administration: would we arrive at the same causal influence if we conduct survey experiments with commonly-used convenience samples, such as students and online participants, to surrogate professional bureaucrats? The answer to this question, based on our findings, is 'No'. Given the popularity of using students or online participants as surrogates for professional bureaucrats, our findings raise some timely alerts about subject recruitment. When the underlying theoretical questions concern the behaviours of professional bureaucrats, convenience samples such as students or online participants are not necessarily good surrogates. Moreover, the study shows the value of replications in behavioural public administration research. One direction to further the development of behavioural public administration research is to pay more attention to generalizability and produce more generalizable theories through replications.

While we believe that our study contributes to behavioural public administration in general and sample selection in particular, the study is limited by the small number of comparisons we were able to make. We were only able to compare four samples by replicating one experiment. Future research could compare more experiments and samples and provide more insights on sample selection and generalizability.

## Notes

1. We selected 48 journals in the Public Administration category from Web of Science database. Then, we excluded journals that primarily focused on public policy and non-profit management and journals in non-English languages, resulting in 21 journals. These journals include *Administration & Society, the American Review of Public Administration*, the *Australian Journal of Public Administration, Canadian Public Administration*, the *International Public Management Journal, International Review of Administrative Sciences*, the *Journal of Homeland Security and Emergency Management*, the *Journal of Policy Analysis and Management*, the *Journal of Public Administration Research and Theory, Local Government Studies, Public Administration, Public Administration and Development, Public Administration Review, Public Management Review, Public Money & Management, Public Performance & Management Review, Public Personnel Management, Regulation & Governance, Review of Public Personnel Administration*, and *Transylvanian Review of Administrative Sciences*.
2. Some articles reported more than two experiments.
3. According to Paolacci and Chandler (2014), the popular crowd-sourcing platform MTurk already had 500,000 registered workers from 190 countries by 2014. Data from the Web tracking company Alexa.com show that MTurk had around 750,000 visitors in December 2015 (cited in Hitlin 2016).
4. There were around 57% American participants and 37% Indian in 2010, according to Marvit (2014).
5. SoJump is a professional survey platform in China, and its website is https://www.wjx.cn/.

6. https://www.wjx.cn/sample/service.aspx (last accessed on 1 May 2019).
7. The manipulation check question employed is 'In the above vignette, which type of citizen-client is fined by the *Chengguan* officer? (A store merchant or a street vendor)'. For details of manipulation check tool, see Oppenheimer, Meyvis, and Davidenko (2009).

## Acknowledgement

An earlier version of this paper was presented at the 77th Midwest Political Science Association Conference. We would like to thank Asmus Leth Olsen, Ling Zhu and other scholars for helpful comments, Ma Ben, Zhang Haibo, Jiang Zhenyu, Zhu Xian for helping with data collection, and anonymous reviewers for their helpful comments and suggestions.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the National Science Foundation of China Project [71774098; 71790611]; Ministry of Education, Youth Project of Humanities and Social Sciences [17YJC630301].

## Notes on contributors

*Xiaoli Lu* is an Associate Professor at Tsinghua University's School of Public Policy and Management. He received his PhD in public administration at the Utrecht School of Governance, Utrecht University. His current research examines the effects of situational factors on street-level bureaucrats' discretion and sensemaking. He also is interested in crisis and disaster management and multimodal organizational analysis.

*Weijie Wang* is an Assistant Professor at the Truman School of Public Affairs, University of Missouri. He received his PhD in Public Policy and Management from USC Price of School of Public Policy. His current research examines how personnel management and reforms affects organizational performance.

*Hao Xu* is a Ph.D. candidate at Tsinghua University's School of Public Policy and Management. His research mainly focuses on street-level bureaucracy and crisis management.

## ORCID

*Xiaoli Lu* http://orcid.org/0000-0001-7633-7222

## References

Abdel-Khalik, A. R. 1974. "On the Efficiency of Subject Surrogation in Accounting Research." *The Accounting Review* 49: 743–750.

Anderson, D. M., and J. M. Stritch. 2015. "Goal Clarity, Task Significance, and Performance: Evidence from a Laboratory Experiment." *Journal of Public Administration Research and Theory* 26 (2): 211–225. doi:10.1093/jopart/muv019.

Aronow, P. M., J. Baron, and L. Pinson. 2019. "A Note on Dropping Experimental Subjects Who Fail A Manipulation Check." *Political Analysis* 27 (4): 572–589. doi:10.1017/pan.2019.5.

Bai, H. (2018). Evidence that a Large Amount of Low Quality Responses on MTurk Can Be Detected with Repeated GPS Coordinates. https://www.maxhuibai.com/blog/evidence-that-responses-from-repeating-gps-are-random

Belle, N., and P. Cantarelli. 2019. "Do Ethical Leadership, Visibility, External Regulation, and Prosocial Impact Affect Unethical Behavior? Evidence from a Laboratory and a Field Experiment." *Review of Public Personnel Administration* 39 (3): 349–371. doi:10.1177/0734371X17721301.

Berinsky, A. J., G. A. Huber, and G. S. Lenz. 2017. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20: 351–368.

Buhrmester, M., T. Kwang, and S. D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-quality, Data?" *Perspectives on Psychological Science* 6 (1): 3–5. doi:10.1177/1745691610393980.

Casler, K., L. Bickel, and E. Hackett. 2013. "Separate but Equal? A Comparison of Participants and Data Gathered via Amazon's MTurk, Social Media, and Face-to-face Behavioral Testing." *Computers in Human Behavior* 29 (6): 2156–2160. doi:10.1016/j.chb.2013.05.009.

Chandler, J., G. Paolacci, E. Peer, P. Mueller, and K. A. Ratliff. 2015. "Using Nonnaive Participants Can Reduce Effect Sizes." *Psychological Science* 26 (7): 1131–1139. doi:10.1177/0956797615585115.

Clifford, S., R. M. Jewell, and P. D. Waggoner. 2015. "Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology?" *Research & Politics* 2 (4): 1–9. doi:10.1177/2053168015622072.

Dreyfuss, E. (2018). A Bot Panic Hits Amazon's Mechanical Turk. https://www.wired.com/story/amazon-mechanical-turk-bot-panic/

Druckman, J. N., and C. D. Kam. 2011. "Students as Experimental Participants: A Defense of the 'Narrow Data Base'." In *Handbook of Experimental Political Science*, edited by J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia, 41–57. New York: Cambridge University Press.

Elliott, W. B., F. D. Hodge, J. J. Kennedy, and M. Pronk. 2007. "Are M.B.A. Students a Good Proxy for Nonprofessional Investors?" *The Accounting Review* 82 (1): 139–168. doi:10.2308/accr.2007.82.1.139.

Esteve, M., D. Urbig, A. Van Witteloostuijn, and G. Boyne. 2016. "Prosocial Behavior and Public Service Motivation." *Public Administration Review* 76: 177–187.

European Union. (2016). Behavioural Insights Applied to Policy: Overview across 32 European Countries. http://publications.jrc.ec.europa.eu/repository/bitstream/JRC100146/kjna27726enn_new.pdf

Feitosa, J., D. L. Joseph, and D. A. Newman. 2015. "Crowdsourcing and Personality Measurement Equivalence: A Warning about Countries Whose Primary Language Is Not English." *Personality and Individual Differences* 75: 47–52. doi:10.1016/j.paid.2014.11.017.

Filtenborg, A. F., F. Gaardboe, and J. Sigsgaard-Rasmussen. 2017. "Experimental Replication: An Experimental Test of the Expectancy Disconfirmation Theory of Citizen Satisfaction." *Public Management Review* 19 (9): 1235–1250. doi:10.1080/14719037.2017.1295099.

Hainmueller, J., D. Hangartner, and T. Yamamoto (2015). Validating Vignette and Conjoint Survey Experiments against Real-world Behavior. *Proceedings of the National Academy of Sciences*, *112*, 2395–2400.

Hermann, M. G., and B. Ozkececi-Taner. 2011. "The Experiment and Foreign Policy Decision Making." In *Cambridge Handbook of Experimental Political Science*, edited by J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia, 430–442. New York, NY: Cambridge University Press.

Hillygus, D. S., N. Jackson, and M. Young. 2014. "Professional Respondents in Nonprobability Online Panels." In *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, P. Lavrakas, J. Krosnick, J. Bethlehem, and A. Gritz, 219–237. West Sussex, UK: Wiley.

Hitlin, P. (2016). Research in the Crowdsourcing Age, a Case Study. https://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/

Huff, C., and D. Tingley. 2015. ""Who are These People?" Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research & Politics* 2 (3): 1–12. doi:10.1177/2053168015604648.

Hughes, C. T., and M. L. Gibson. 1991. "Students as Surrogates for Managers in a Decision-making Environment: An Experimental Study." *Journal of Management Information Systems* 8 (2): 153–166. doi:10.1080/07421222.1991.11517925.

Jacobsen, C. B., and L. E. Jensen. 2017. "Why Not "Just for the Money"? an Experimental Vignette Study of the Cognitive Price Effects and Crowding Effects of Performance-related Pay." *Public Performance & Management Review* 40 (3): 551–580. doi:10.1080/15309576.2017.1289850.

James, O., S. R. Jilke, and G. G. Van Ryzin. 2017. "Behavioural and Experimental Public Administration: Emerging Contributions and New Directions." *Public Administration* 75 (4): 865–873. doi: 10.1111/padm.12363.

Jilke, S., and L. Tummers. 2018. "Which Clients are Deserving of Help? A Theoretical Model and Experimental Test." *Journal of Public Administration Research and Theory* 28 (2):226–238. doi:10.1093/jopart/muy002

Johnson, D. R., and L. A. Borden. 2012. "Participants at Your Fingertips: UsingAmazon's Mechanical Turk to Increase Student–faculty Collaborative Research." *Teaching of Psychology* 39 (4): 245–251. doi:10.1177/0098628312456615.

Jones, W. L., and B. S. Sonner. 2001. "Just Say No to Traditional Student Samples." *Journal of Advertising Research* 41 (5): 63–71. doi:10.2501/JAR-41-5-63-71.

Kennedy, R., S. Clifford, T. Burleigh, P. Waggoner, and R. Jewell (2018). The Shape of and Solutions to the MTurk Quality Crisis. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3272468

Kroll, A., and G. A. Porumbescu. 2019. "When Extrinsic Rewards Become "Sour Grapes": Anexperimental Study of Adjustments in Intrinsic and Prosocial Motivation." *Review of Public Personnel Administration* 39: 467–486. doi:10.1177/0734371X15608419.

Krupnikov, Y., and A. S. Levine. 2014. "Cross-sample Comparisons and External Validity." *Journal of Experimental Political Science* 1 (1): 59–80. doi:10.1017/xps.2014.7.

Lee, M. J., M. J. Moon, and J. Kim. 2017. "Insights from Experiments with Duopoly Games: Rational Incremental Decision-making." *Public Management Review* 19 (9): 1328–1351. doi:10.1080/14719037.2017.1282002.

Levay, K. E., J. Freese, and J. N. Druckman. 2016. "The Demographic and Political Composition of Mechanical Turk Samples." *SAGE Open* 6 (1): 1–17. doi:10.1177/2158244016636433.

Levitt, S. D., and J. A. List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21 (2): 153–174. doi:10.1257/jep.21.2.153.

Li, H., and G. G. Van Ryzin. 2017. "A Systematic Review of Experimental Studies in Public Management Journals." In *Experiments in Public Management Research: Challenges & Contributions*, edited by O. James, S. Jilke, and G. G. Van Ryzin, 20–36. New York, NY: Cambridge University Press.

Li, H., C. Kuo, and M. G. Rusell. 1999. "The Impact of Perceived Channel Utilities, Shopping Orientations, and Demographics on the Consumer's Online Buying Behavior." *Journal of Computer-Mediated Communication* 5: 1–20.

Li, H., J. Liang, H. Xu, and Y. Liu. 2018. "Does Windfall Money Encourage Charitable Giving? An Experimental Study." *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations* 30 (4): 841–848. doi:10.1007/s11266-018-9985-y.

Li, X., W. Shi, and B. Zhu. 2018. "The Face of Internet Recruitment: Evaluating the Labor Markets of Online Crowdsourcing Platforms in China." *Research & Politics* 5 (1): 1–8. doi:10.1177/2053168018759127.

Liyanarachchi, G. A. 2007. "Feasibility of Using Student Subjects in Accounting Experiments: A Review." *Pacific Accounting Review* 19 (1): 47–67. doi:10.1108/01140580710754647.

Lu, X., H. Xu, and W. Wang. 2019. "Clients' Help Deservingness, Crowd Situational Stress and Discretionary Decision-making: An Experimental Study of Regulatory Street-level Bureaucrats in China." *International Public Management Journal* 1–26. doi:10.1080/10967494.2019.1661892.

Lupton, D. L. 2018. "The External Validity of College Student Subject Pools in Experimental Research: A Cross-sample Comparison of Treatment Effect Heterogeneity." *Political Analysis* 27 (1): 90–97. doi:10.1017/pan.2018.42.

Marvit, M. Z. (2014). How Crowdworkers Became the Ghosts in the Digital Machine. The Nation. http://www.thenation.com/article/how-crowdworkers-became-ghosts-digital-machine/

Mintz, A., S. B. Redd, and A. Vedlitz. 2006. "Can We Generalize from Student Experiments to the Real World in Political Science, Military Affairs, and International Relations?" *Journal of Conflict Resolution* 50 (5): 757–776. doi:10.1177/0022002706291052.

Mullinix, K. J., T. J. Leeper, J. N. Druckman, and J. Freese. 2016. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2 (2): 109–138. doi:10.1017/XPS.2015.19.

Oppenheimer, D. M., T. Meyvis, and N. Davidenko. 2009. "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power." *Journal of Experimental Social Psychology* 45 (4): 867–872. doi:10.1016/j.jesp.2009.03.009.

Paolacci, G., and J. Chandler. 2014. "Inside the Turk: UnderstandingMechanical Turk as a Participant Pool." *Current Directions in Psychological Science* 23 (3): 184–188. doi:10.1177/0963721414531598.

Porter, C. O. L. H., R. Outlaw, J. P. Gale, and T. S. Cho. 2019. "The Use of Online Panel Data in Management Research: A Review and Recommendations." *Journal of Management* 45 (1): 319–344. doi:10.1177/0149206318811569.

Remus, W. 1986. "Graduate Students as Surrogates for Managers in Experiments on Business Decision Making." *Journal of Business Research* 14 (1): 19–25. doi:10.1016/0148-2963(86)90053-6.

Rosenthal, R. W., and R. L. Rosnow. 1969. *Artifact in Behavioral Research*. New York: Academic Press.
Ryan, T. J. (2018). Data Contamination on MTurk. http://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/
Sears, D. O. 1986. "College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51 (3): 515–530. doi:10.1037/0022-3514.51.3.515.
Smith, N. A., I. E. Sabat, L. R. Martinez, K. Weaver, and S. Xu. 2015. "A Convenient Solution: Using MTurk to Sample from Hard-to-reach Populations." *Industrial and Organizational Psychology* 8 (2): 220–228. doi:10.1017/iop.2015.29.
Stokel-Walker, C. (2018). Bots on Amazon's Mechanical Turk are Ruining Psychology Studies. https://www.newscientist.com/article/2176436-bots-on-amazons-mechanical-turk-are-ruining-psychology-studies/
Stritch, J. M., M. J. Pedersen, and G. Taggart. 2017. "The Opportunities and Limitations of Using Mechanical Turk (Mturk) in Public Administration and Management Scholarship." *International Public Management Journal* 20 (3): 489–511. doi:10.1080/10967494.2016.1276493.
Tsang, E. W. K., and K.-M. Kwan. 1999. "Replication and Theory Development in Organizational Science: A Critical Realist Perspective." *Academy of Management Review* 24 (4): 759–780. doi:10.5465/amr.1999.2553252.
Van Ryzin, G. G., N. M. Riccucci, and H. Li. 2017. "Representative Bureaucracy and Its Symbolic Effect on Citizens: A Conceptual Replication." *Public Management Review* 19 (9): 1365–1379. doi:10.1080/14719037.2016.1195009.
Walker, R. M., O. James, and G. A. Brewer. 2017. "Replication, Experiments and Knowledge in Public Management Research." *Public Management Review* 19 (9): 1221–1234. doi:10.1080/14719037.2017.1282003.
Walter, S. L., S. E. Seibert, D. Goering, and E. H. O'Boyle. 2019. "A Tale of Two Sample Sources: Do Results from Online Panel Data and Conventional Data Converge?" *Journal of Business and Psychology* 34 (4): 425–452. doi:10.1007/s10869-018-9552-y