JPS
Journal of Policy Studies

Meier, K., An, S., Davis, J., & Park, J. (2023). Secondary Benefits of Manipulation Checks: Three Illustrations From Behavioral Public Administration. *Journal of Policy Studies*, *38*(4), 1–8. https://doi.org/10.52372/jps38401

Articles

Secondary Benefits of Manipulation Checks: Three Illustrations From Behavioral Public Administration

Kenneth Meier[1], Seung-ho An[2], Jourdan Davis[3], Joohyung Park[4]

[1] Public Administration and Policy, American University, [2] School of Government and Public Policy, University of Arizona, [3] University of North Carolina Charlotte, [4] Public Administration and Public Policy, American University

Keywords: manipulation checks, experiments, behavioral public administration, public sector bias, internet recruitment of subjects, incentives

https://doi.org/10.52372/jps38401

Vol. 38, Issue 4, 2023

Manipulation checks in behavioral public administration are commonly used and reported to determine if the experimental and control group have received different treatments. This paper uses three experiments to argue that manipulation checks for experimental treatments can have secondary benefits that can be used to improve the quality of behavioral work in the field. The three cases address the importance of using more clear terms in experimental manipulations (government v. public), using different on-line platforms to recruit experimental subjects (Mechanical Turk, Prolific, and Data.Spring), and whether larger payments more produce more attentive subjects.

Although the behavioral approach to public administration and public policy has long historical roots (see Roethlisberger & Dickson, 1939; Simon, 1947; Martin and Sanderson 2009), in recent years experimental research has significantly increased with both vignette experiments with the general public (Campbell, 2023) and using public employees as subjects (Orey & Craemer, 2023). Although there are useful guides to best practices in experiment research (James et al., 2020), any experimental process will generate some new insights on a regular basis that can be used to increase the validity of future research. This study examines how a common tool used to assess the internal validity of experiments, manipulation checks, can provide insight into three issues: (1) how the terms used in experiments can enhance the validity of results, (2) the reliability of various crowdsourcing platforms that generate samples of convenience, and (3) whether incentives matter in recruiting experimental subjects. These secondary benefits can then contribute to improving experimental design by better choice of phrasing, selection of the recruitment platform, or using greater incentives in recruitment of subjects.

The Value of Manipulation Checks

A common concern in experimental research is determining whether the intended experimental treatment actually was applied to the experimental subjects and not the control group (Ejelöv & Luke, 2020; Mutz & Pemantle, 2015). This concern exists whether the experiment is in medicine and patients do not take the medicine or follow the treatment specifications, in the public policy behavioral nudge literature (John, 2018), in lab or survey experiments where the treatment is verbal or visual (Mutz, 2011). In a wide variety of areas within behavioral public administration including sector bias (Hvidman & Andersen, 2016), performance information (Petersen, 2020), audit studies (Lahey & Beasley, 2009), subjects are given cues, often subtle cues in mere mention studies, that may or may not be picked up by the experimental subjects. Experimental scholars have long heeded Leon Festinger's (1953, p. 145) admonition that "It is rarely safe to assume beforehand that the operations used to manipulate variables will be successful and will tie in directly with the concept the experimenter has in mind." Using a post-treatment manipulation check to determine if the experimental subjects perceived the treatment and the control group subjects did not is advocated as a best practice in experimental research whenever possible in political science (Mutz & Pemantle, 2015), psychology (Flake et al., 2017), organizational research (Highhouse, 2009), operations research (Bachrach & Bendoly, 2011) and other behavioral sciences.[1]

The logic for treatment effects is simple and direct. Subjects are randomly (R) assigned to the experimental (t for treatment) and control (c) groups.[2] The experimental subjects are then exposed to an experimental treatment (X)

1 There is a literature raising a question about pretreatment manipulation tests and whether or not that generates a framing effect that might bias the experiment (Fayant et al., 2017; Hauser et al., 2018). Our discussion only involves post-treatment manipulation tests and thus any potential framing problems should not be relevant. Our discussion does not directly deal with attention tests that seek to determine if respondents are responding randomly or simply being inattentive but do not apply directly to the treatment.

2 The control group might not be an actual control group but a designated comparison group. For example, gender bias studies might compare women to men, motivated reasoning explanations might compare those with strong pre-existing attitudes to those without, or Bayesian decision experiments might compare those with priors to those without.

and the control group is not. The outcome dependent variable is then observed for both the experimental group (O_t) and the control group (O_c). The experimental and control group are both then asked if they observed the treatment (M_t and M_c, respectively). For a 2 x 2 between subjects experiment, it takes the following logical form:

$$\begin{array}{cccc} R & X & O_t & M_t \\ R & & O_c & M_c \end{array}$$

The results from the manipulation check are then compared to the actual treatment as illustrated by following table to determine if the experimental group differed from the control group in exposure to the manipulation test:

	Treatment	
Manipulation Check	X	Not X
M_t	a	b
M_c	c	d

The test might be done by comparing the percentage of the experimental group that correctly perceived the experimental condition to the percentage of the control group who falsely perceived the experimental condition (comparing a to c) with a f-test or as Mutz and Pemantle (2015) suggest using all the data in the table to calculate a chi-square test. Significant results of either test are an indication that the control and experiment groups differ in terms of the perceived treatment. The advantage of the chi-square test over the percentage test is that it not only takes advantage of all the data and allows for misperceptions that might be common to both groups, but it is easier to apply in situations with multiple control groups or if one includes an "unsure" category in the responses for the manipulation check. Since the current set of illustrations will at times be using experiments with different sample sizes and the chi-square calculations are affected by the number of cases, we will rely the percentage of experiment subjects whose response to the manipulation check match the actual experimental condition (a/(a+c)).

In addition to this key role in assessing the internal validity of the experiment, we suggest that there can be a variety of second-order benefits to post-treatment testing for manipulation effects. One common use that will not be discussed here is using the manipulation check as an instrumental variable to estimate local average treatment effects rather than the impact of the "intent to treat" (Angrist & Imbens, 1995; Mourifié & Wan, 2017). Our concern will be using the manipulation checks for either substantive or methodological information for either hypothesis testing or to improve research designs. Petersen (2020), for example, used information from manipulation check results to determine if motivated reasoning varied by whether information was positive or negative. He found that the manipulation checks revealed that negative information resulted in less attention to the accuracy of information and thus less need for motivated reasoning. Such a use is rare as Ejelöv and Luke (2020) conclude in their extensive survey of manipulation checks in social psychology, "In our sample, manipulation checks (of any type) were rarely used for analytic purposes other than data exclusion."

Experiment 1: Question Wording - Public or Government?

Mere mention experiments simply make a brief mention of some experimental condition thought to influence results (Gaines et al., 2007). Such survey or field experiments are used in audit studies to probe discrimination where fictitious job applications or requests for information are sent to individuals or organizations (Lahey & Beasley, 2009), in survey experiments that might assess sector bias (Hvidman & Andersen, 2016; Marvel, 2016), studies of blame avoidance via contracting or other forms of delegation (Johnson et al., 2019; Piatak et al., 2017), or examinations of questions symbolic representation (Riccucci et al., 2014) and similar studies of gender or racial bias (Funk, 2019) among others.

The assumption behind "mere mention" experiments is that a brief mention will convey a specific meaning to the respondent. The experiment used to illustrate the utility of manipulation checks for question wording was an experiment on sector bias in the delivery of services. This literature asks if public sector organizations are systematically perceived as less effective (or some other evaluative criteria) than private sector organizations when performance outcomes are equal, or alternatively if private organizations get more credit for positive performance results than public ones do (see Hvidman & Andersen, 2016). Hvidman and Andersen (2016, p. 113) specifically suggest just the word "public" might trigger biases: "Given that there exist negative stereotypes of public sector organizations, we would expect the word 'public' to prime respondents for beliefs about low performance and, therefore, make them evaluate the performance of an organization labeled 'public' worse than otherwise identical organizations." The normative concern is that such misperceptions of performance have implications for trust in government and diffuse support for the political system which are key elements in the relationship between democracy and administration. The literature is somewhat mixed on the sector bias question and has been applied to only a few types of services (mail services, hospitals, nursing homes, see Hvidman & Andersen, 2016; Marvel, 2016; Meier et al., 2022) so the question of where and under what conditions sector bias exists remains important in public administration.

The example is drawn from a study of sector bias in the US nursing home industry (Meier et al., 2022) that seeks to evaluate information credibility as well as sector bias. The pretest reported here was conducted for two reasons. First, there is a great deal of misinformation in the US among who owns and operates nursing homes including among individuals who actually have placed family members in such homes (Ben-Ner et al., 2019). In such cases, mere mention cues might be ineffective. Second, while studies have traditionally framed the experiments as "public" and "private" organizations, less attention has been paid to what subjects might think of as a public organization. Based on the the-

oretical discussions in this literature (Rainey et al., 1976 and subsequent work), researchers often simplify the distinction to conceive of public organizations as those owned and operated by government and private organizations as those operated by private individuals (although a few studies distinguish between private for profit and private nonprofit organizations, see Meier & An, 2020). It is possible that a mere mention of a "public" organization might not trigger the perception that the organization is government owned and operated. After all, in the US a public corporation is a private organization owned by stockholders; a private club is privately owned and not open to the public whereas a public club would be privately owned but open to the public for doing business.

To address this concern with whether "public" was the appropriate term to use, during the pretest of the experiment respondents were randomly assigned different vignettes that described a nursing home as either a "public" nursing home or a "government owned" nursing home (the experiment also included private for profit nursing homes and private nonprofit homes). Other information on performance and evaluators were also randomly assigned. After the subjects were asked to evaluate the performance of the nursing home on a variety of dimensions, they were asked on a separate page to respond to manipulation checks and some demographic questions. One manipulation check asked the subject to identify if the nursing home was "Public or government owned," "private for profit," or "private nonprofit." Subjects were also allowed to check a "don't know" category. The relevant responses are in the table below:

Table 1. Government is Superior to Public as a Mere Mention Cue

	Cue	
	Public	Government
Subject's Response		
Public or Government	26.8%	46.5%
Other Response	73.2%	54.5%
N	381	396

Although 46.5% would not be a manipulation check that stands out in the literature, it is a clear improvement over 26.8% and compares favorably to those who received the for-profit cue and misidentified the home as government/public (7.8%), and those who received the non-profit cue and misidentified the home as government/public (13.2%). The results using "government" show a manipulation result strong enough to conclude that the government treatment was distinct from the other sector treatments and clearly superior to using the term "public." This simple wording distinction is relevant for substantial research that is experimental or even surveys of the general public (Gupta et al., 2023) given the ambiguity about how some services are delivered (Fitriningrum et al., 2023) or the complexity of organizations that do not precisely fit in existing categories (Oh et al., 2023).

Experiment 2. Evaluating Platforms for Recruiting Subjects in Internet Experiments

Convenience samples are frequently used in behavioral public administration, and the rise of internet recruitment platforms has dramatically lowered the cost of doing so. While several papers have demonstrated that internet samples from Mechanical Turk (MTurk) compared favorably to other convenience samples and at times even to more expensive representative sampling processes (Berinsky et al., 2012; Casler et al., 2013; Hauser & Schwarz, 2016), in some countries scholars have two or more choices for internet recruitment platforms. In the US for example, MTurk, Prolific, Lucid, YouGov, SurveyJunkie and others can be used for online experiments. Even a relatively small country such as Korea has several options (Data.Spring, Do It Survey, Embrain-Macromill group). A scholar interested in conducting an experiment ideally would like to know the quality of the subjects in addition to the cost (see below). The latter is not systematically available and currently relies on informal communication among scholars.

In a recent survey experiment involving public responses to government actions in regard to the covid-19 pandemic across 8 countries, we were forced to consider alternatives to the MTurk default either because MTurk had few workers in the country or did not operate at all (Amirkhanyan et al., 2023). Although not set up to systematically test the quality of the survey respondents, this provided an opportunity to get a rough indicator of the quality of responses on three different survey platforms: MTurk, Prolific, and Data.Spring. Respondents in each country were asked to evaluate the response of a hypothetical government to covid-19 on a variety of performance dimensions. Three treatment variables were included: the generic policy action of the government (democratic or autocratic), the evaluation of the policy action by an independent international organization (positive or negative), and inequality of the impact (whether low income individuals were more detrimentally affected or not).

Although one might define the quality of subjects in a variety of ways, one minimum standard might be that subjects pay attention to the experiment. Variation in correct responses to manipulation checks might be a reasonable indicator of the quality of subjects. Table 2 presents the average percentage of subjects who correctly identified each of the three treatments in the post evaluation manipulation check. In general the manipulation checks show a strong treatment effect with values generally ranging be-

Table 2. Comparison of Survey Platforms: Percent Correct on Three Manipulation Checks

Country	Platform	% Correct	Subjects
United States	MTurk	81.9	986
Germany	Prolific	93.2	987
Italy	Prolific	91.6	996
Spain	Prolific	86.3	987
Canada	Prolific	93.0	1000
U.K.	Prolific	92.1	999
South Korea	Data.Spring	73.9	1007
Denmark[1]	Prolific	88.3	117

[1] We were unable to attract sufficient responses in Denmark via Prolific but included those results since they appear in the original study where they are highly consistent with the results from the other seven countries.

tween 85 and 95%. Although we cannot separate out country effects from survey platform effects (that would require within country comparisons),[3] the results appear to indicate that Prolific generates the highest quality respondent pool (90.8%) compared to MTurk (81.9) and Data.Spring (73.9).

Any definitive conclusions are premature, however, given that the alternative hypothesis of country differences in subject pools cannot be ruled out. Similar assessments within a country that use pools from different providers are needed to make that assessment. A meta analysis of existing studies, however, might provide some corroborating evidence.

Experiment 3: Do You Get What You Pay For?

Quality of subjects is only one consideration for a researcher; the cost of subjects also places a limit on the research one can conduct. Unlike Prolific and Data.Spring which set the cost of the respondents, MTurk provides some flexibility in how much subjects are paid (the variation in wage rates including minimum wage rates makes determining pricing difficult). Although many experiments provide only token compensation, a logical question to pose is whether higher levels of compensation might result in a higher quality sample of subjects.[4]

We investigated the relationship between payment amount and subject quality by fielding two blame avoidance experiments 30 days apart on MTurk (An & Meier, 2021). The survey experiments involved the Federal Aviation Administration and who might be blamed for airplane crashes based on a fact pattern for the Boeing 737 Max. In the first experiment, subjects were paid $0.80 for a five minute survey; in the second experiment they were offered half that amount ($0.40). Individuals were not permitted to participate in both surveys to avoid learning effects. Two manipulation checks were asked: First a question about who appointed the head of the FAA and the second about whether or not the FAA contracted out the regulatory work for a failed safety system. The results in Table 3 show that while higher compensated subjects were slightly more likely to pass the more difficult manipulation check (it was embedded in the vignette rather than in the first sentence), they were slightly less likely to pass the easier presidential appointment check. Neither difference, however, is anywhere near statistically significant; the relative compensation appears to be unrelated to subject quality.

Table 3. Increasing Token Payments Does Not Improve Subject Quality (% Correct)

Manipulation Check	Low Pay	High Pay
Who Appoints the FAA Head	96.1	95.3
Was the Safety Work Contracted	45.8	47.3
N	448	463

Why might incentives have not worked as predicted in this case? One possible explanation is that low quality MTurkers might be the largest share of potential subjects and they rapidly fill up the demand regardless of the price. After all any worker willing to work for the lower wage should also be willing to work for the higher wage given the equal nature of the work. A second possibility is that the difference in wages which are small to start off with are not large enough to create any incentive effects. It is quite possible that much larger differences could generate differences in the quality of the respondents. The third possibil-

3 We get some fragmentary evidence that separates out country effects given that we initially tried to use MTurk in Italy, Spain, and Canada, but in all three cases were unable to get sufficient subjects and abandoned those subjects and recruited a full panel via Prolific. That evidence is very mixed as shown by the respective Prolific and MTurk results for Italy (91.6 v. 90.4), Spain (86.3 v. 86.6), and Canada (93.0 v. 88.4).

4 Payment rates do appear to affect participation, that is, how quickly the number of needed subjects participate in the experiment (Buhrmester et al., 2011) but no studies have examined how payment rates affect quality.

ity is that there were no screens to distinguish quality before allowing individuals to take the survey (other than the screening for non-US IP addresses and the screens to eliminate bots) and thus there were simply no limits on the ability of any quality respondent to apply.

Conclusions

Ejelöv and Luke (2020, p. 7) stress the importance of manipulation checks, "Given that successfully manipulating independent variables is the sine qua non of experimental methodology, it is highly important that researchers take seriously the task of vetting their manipulations." Although manipulation checks play this crucial role in establishing the internal validity of experiments and can also be used to estimate local average treatment effects, this research argued that they can have additional second order value in both methodological and substantive terms. The three illustrations were presented – determining appropriate word choice, assessing the quality of recruitment platforms, and determining appropriate incentives – do not exhaust the possibilities. The word choice illustration has multiple permutations in terms of how treatment effects might be framed in terms of style of presentation, order of presentation, and degree of emphasis. Many such decisions are made in the design of experiments, often via pretests or focus groups that would be valuable if shared with other scholars. Although much work has been done on the various ways to recruit experimental subjects (see Berinsky et al., 2012), it is clear that additional work could be done by constructing better comparisons (within country or within subject type) for internet samples or other types of convenience samples. And direct payment of subjects is only one type of incentive that can be used to recruit subjects; normative appeals (Bellé, 2013) or lottery entry appeals Samuels & Zucco, 2013) can also be used.

A method of systematic reporting of such second-order examinations of manipulation checks or other similar assessments in behavioral public administration would be valuable to scholars in the field. It would create greater efficiencies in the design of research and contribute to the internal validity of experimental work. Publishing such work as formal articles likely sets a high barrier and might be perceived as imposing high relative costs on the researcher. A convenient and accessible reporting system via some type of searchable repository or blog might be an alternative way to communicate what could be a valuable information to the scholarly community.

Submitted: December 06, 2023 KST, Accepted: December 10, 2023 KST

References

Amirkhanyan, A. A., Meier, K. J., Song, M., Roberts, F. W., Park, J., Vogel, D., Bellé, N., Molina, A. L., Jr., & Guul, T. S. (2023). Liberté, Égalité, Crédibilité: An experimental study of citizens' perceptions of government responses to COVID-19 in eight countries. *Public Administration Review*, *83*(2), 401–418. https://doi.org/10.1111/puar.13588

An, S., & Meier, K. J. (2021). *When the Plane Crashes: Contracting and Blame Avoidance for Failures* [Unpublished paper].

Angrist, J. D., & Imbens, G. W. (1995). *Identification and estimation of local average treatment effects*. National Bureau of Economic Research. https://doi.org/10.3386/t0118

Bachrach, D. G., & Bendoly, E. (2011). Rigor in behavioral experiments: A basic primer for supply chain management researchers. *Journal of Supply Chain Management*, *47*(3), 5–8.

Bellé, N. (2013). Experimental evidence on the relationship between public service motivation and job performance. *Public Administration Review*, *73*(1), 143–153.

Ben-Ner, A., Hamann, D. J., & Ren, T. (2019). Does Ownership Matter in the Selection of Service Providers? Evidence from Nursing Home Consumer Surveys. *Nonprofit and Voluntary Sector Quarterly*, *47*(6), 1271–1295. https://doi.org/10.1177/0899764018790698

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. https://doi.org/10.1093/pan/mpr057

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Campbell, J. W. (2023). Public Participation and Trust in Government: Results From a Vignette Experiment. *Journal of Policy Studies*, *38*(2), 23–31. https://doi.org/10.52372/jps38203

Casler, K., Bickel, L., & Hackett, E. (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. https://doi.org/10.1016/j.chb.2013.05.009

Ejelöv, E., & Luke, T. J. (2020). "Rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology*, *87*, 103937. https://doi.org/10.1016/j.jesp.2019.103937

Fayant, M.-P., Sigall, H., Lemonnier, A., Retsin, E., & Alexopoulos, T. (2017). On the limitations of manipulation checks: An obstacle toward cumulative science. *International Review of Social Psychology*, *30*(1), 125–130. https://doi.org/10.5334/irsp.102

Festinger, L. (1953). Laboratory experiments. In L. Festinger & D. Katz (Eds.), *Research Methods in the Behavioral Sciences* (pp. 136–172). Dryden Press.

Fitriningrum, A., Pulungan, A. H., & Darusidhi, P. (2023). Government, state-owned enterprise, and liberalization of universal postal obligation services. *Journal of Policy Studies*, *38*(3), 25–40. https://doi.org/10.52372/jps38303

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378.

Funk, K. (2019). If the shoe fits: Gender role congruity and evaluations of public managers. *Journal of Behavioral Public Administration*, *2*(1). https://doi.org/10.30636/jbpa.21.48

Gaines, B. J., Kuklinski, J. H., & Quirk, P. J. (2007). The logic of the survey experiment reexamined. *Political Analysis*, *15*(1), 1–20. https://doi.org/10.1093/pan/mpl008

Gupta, A. K., Bhurtel, A., & Bhattarai, P. C. (2023). Service Users' Confidence in Accessing Public Services in Nepal: What Makes Differences? *Journal of Policy Studies*, *38*(1), 29–43. https://doi.org/10.52372/jps38103

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology*, *9*, 998. https://doi.org/10.3389/fpsyg.2018.00998/full

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods*, *12*(3), 554–566.

Hvidman, U., & Andersen, S. C. (2016). Perceptions of public and private performance: Evidence from a survey experiment. *Public Administration Review*, *76*(1), 111–120. https://doi.org/10.1111/puar.12441

James, O., Olsen, A. L., Moynihan, D. P., & Van Ryzin, G. G. (2020). *Behavioral public performance: How people make sense of government metrics*. Cambridge University Press. https://doi.org/10.1017/9781108761338

John, P. (2018). *How far to nudge?: assessing behavioural public policy*. Edward Elgar Publishing. https://doi.org/10.4337/9781786430557

Johnson, A. P., Geva, N., & Meier, K. J. (2019). Can hierarchy dodge bullets? Examining blame attribution in military contracting. *Journal of Conflict Resolution*, *63*(8), 1965–1985. https://doi.org/10.1177/0022002718824984

Lahey, J. N., & Beasley, R. A. (2009). Computerizing audit studies. *Journal of Economic Behavior & Organization*, *70*(3), 508–514. https://doi.org/10.1016/j.jebo.2008.02.009

Marvel, J. D. (2016). Unconscious bias in citizens' evaluations of public sector performance. *Journal of Public Administration Research and Theory*, *26*(1), 143–158.

Meier, K. J., & An, S. (2020). Sector bias in public programs: US nonprofit hospitals. *Journal of Behavioral Public Administration*, *3*(1), 1–8. https://doi.org/10.30636/jbpa.31.107

Meier, K. J., Song, M., Davis, J. A., & Amirkhanyan, A. A. (2022). Sector bias and the credibility of performance information: An experimental study of elder care provision. *Public Administration Review*, *82*(1), 69–82.

Mourifié, I., & Wan, Y. (2017). Testing local average treatment effect assumptions. *Review of Economics and Statistics*, *99*(2), 305–313. https://doi.org/10.1162/rest_a_00622

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton University Press.

Mutz, D. C., & Pemantle, R. (2015). Standards for experimental research: Encouraging a better understanding of experimental methods. *Journal of Experimental Political Science*, *2*(2), 192–215.

Oh, M., Lee, S.-Y., & Jung, Y. (2023). Loyal to the Public: Examining the Relationship Between Chief Executives and the Pursuit of Public Values in Quangos. *Journal of Policy Studies*, *38*(1), 1–14. https://doi.org/10.52372/jps38101

Orey, B. D., & Craemer, T. (2023). Black and Blue: Black Police Officers' Implicit and Explicit Biases in Split-Second Decisions to Shoot or Not to Shoot Unarmed Black Civilians. *Journal of Policy Studies*, *38*(3), 1–13. https://doi.org/10.52372/jps38301

Petersen, N. B. G. (2020). Whoever Has Will be Given More: The Effect of Performance Information on Frontline Employees' Support for Managerial Policy Initiatives. *Journal of Public Administration Research and Theory*, *30*(4), 533–547. https://doi.org/10.1093/jopart/muaa008

Piatak, J., Mohr, Z., & Leland, S. (2017). Bureaucratic accountability in third-party governance: Experimental evidence of blame attribution during times of budgetary crisis. *Public Administration*, *95*(4), 976–989. https://doi.org/10.1111/padm.12341

Rainey, H. G., Backoff, R. W., & Levine, C. H. (1976). Comparing public and private organizations. *Public Administration Review*, *36*(2), 233. https://doi.org/10.2307/975145

Riccucci, N. M., Van Ryzin, G. G., & Lavena, C. F. (2014). Representative bureaucracy in policing: Does it increase perceived legitimacy? *Journal of Public Administration Research and Theory*, *24*(3), 537–551. https://doi.org/10.1093/jopart/muu006

Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the Worker*. Harvard University Press.

Samuels, D. J., & Zucco, C. (2013). *Using Facebook as a subject recruitment tool for survey-experimental research. Available at SSRN 2101458*.

Simon, H. A. (1947). *Administrative behavior*. The Free Press.

Supplementary Materials

Download: https://jps.scholasticahq.com/article/88435-secondary-benefits-of-manipulation-checks-three-illustrations-from-behavioral-public-administration/attachment/190227.html