Journal of Behavioral
Public Administration
Vol 8, pp. 1-19
DOI: 10.30636/jbpa.81.364

Research Article

More Random Than Not? A Review of the Logic of Inference in Experimental Public Administration

Spiro Maroulis*, Ulrich Thy Jensen*, Youngjae Won*, Jesper Asring Hansen†, Christian Bøtcher Jacobsen†, Ole Helby Petersen††

Abstract: The emphasis on causal inference in public administration research has spurred a proliferation of experimental studies, primarily due to the internal validity attributed to findings from random assignment of treatment conditions. However, experimental studies often highlight numerous findings that draw on varying logics of inference. Some inferences are based entirely on random assignment, others involve interactions to conduct subgroup or exploratory analyses, and some are rooted in non-experimental observations emanating from the study. We examine the logic of inference in experimental public administration. Drawing on a sample of experimental studies, we find that 57.5% of the findings rely solely on randomized inference, 18.8% involve interactions between randomly and non-randomly assigned factors, and 23.7% are based on logics unrelated to random assignment. Additionally, we investigate how these "upstream" findings are interpreted in the "downstream" studies that cite them. We find that 77.4% of downstream citations use the upstream findings to support causal claims. Of those, 41.6% are not rooted in a logic based on randomization, suggesting a misalignment between the logic of inference underlying results and their application by citing researchers. This misalignment has significant implications for the accumulation of scientific knowledge in public administration and may worsen if not addressed. We offer advice to upstream researchers on clearly stating the logic of inference underlying their key findings and for downstream researchers to carefully evaluate this logic to ensure accurate interpretation of research claims.

Keywords: Experiments, causal inference, randomization, public administration

Introduction

Experiments offer a vehicle for testing, developing, and refining theories central to public administration scholarship and have been pivotal for the surge of behavioral public administration research (Grimmelikhuijsen et al., 2017). The proliferation of experiments has placed randomized controlled trials as a mainstream element in public administration researchers' toolbox (James et al., 2017) and ascribed a premium to experimental results for their ability to elicit evidence about cause and effect. Scholars have, for example, used experimental designs to advance theories on how to recruit and retain public employees (Linos, 2018; Linos et al., 2022), how citizens and decision-makers respond to performance information (Baekgaard & Serritzlew, 2016; James & Van Ryzin, 2017; Nielsen & Moynihan, 2017), and how leadership affects employee motivation and organizational performance (Jacobsen et al., 2022; Jensen et al., 2019). Regardless of the topic, results from experimental studies play a constituent role in how we think about the interrelationships and implications of concepts core to public administration and management scholarship.

However, with the surge in experimentation has come substantial heterogeneity in how researchers conduct experiments. Numerous insightful reviews have emerged, aiming to guide best practices in experimental research (Baekgaard et al., 2015; Hansen & Tummers, 2020), assess the evidentiary value of experimental studies (Vogel & Xu, 2021), and map the topics of experimental studies in our field (Battaglio et

*Arizona State University, †Aarhus University, ††Roskilde University
Address correspondence to Spiro Maroulis at spiro.maroulis@asu.edu.

al., 2019). Unlike these reviews, we do not seek evidentiary value or research practices to correct. Instead, our review of experimental literature in public administration focuses on an issue that is even more fundamental to the use of experiments and the dissemination of their findings: the observation that, even within experiments following best practices, substantial heterogeneity exists concerning how researchers draw causal inferences. Some findings in experimental studies are entirely rooted in random manipulation, others involve interactions to conduct subgroup or exploratory analyses, and some build entirely on other observations emanating from the broader context of the experiment. This variation is often well-founded and provides insights of great importance for our field, such as when researchers investigate heterogeneous treatment effects across conditions that are not randomly assigned (e.g., gender, ethnicity, job type, education, employment sector). However, current practices can also lead to confusion about the logic of inference on which a particular finding rests. Stated differently, the internal validity benefits presumed of "experimental findings" do not necessarily apply equally to all findings within a study simply because the study featured an experiment. Researching current practices in experimental research can offer crucial insights to aid public administration and management scholarship in interpreting and applying findings from experimental studies.

Despite the fundamental nature of questions relating to causality and inference, the basis of inference among findings from experimental studies has received relatively little attention. This topic is crucial to evaluate now, as experimental work is propagating within our field and the standards for conducting and interpreting experimental studies are path dependent. Thus, our field will likely get locked into the established norms and their ramifications for years. Consequently, it is crucial to carefully examine the state of experimental work in public administration and thereby offer insights of importance to testing, developing, and refining theories central to public administration scholarship. To do so, we conduct an in-depth and systematic review of the underlying basis of inference for the primary empirical findings reported in all experimental studies published in the *Journal of Public Administration Research and Theory* (JPART) between 1991-2020. We identified all findings emphasized by authors in article abstracts and coded their logic of inference after carefully examining each article's empirical design and statistical analysis. We refer to these findings as the "upstream" findings. Importantly, we match this descriptive analysis with an in-depth citation analysis of the propagation of these findings in a randomly selected subset of citing articles across all journals in public administration. We refer to these citation instances as the "downstream" citations.

Before we outline our methodological approach and present the results of our analyses, we want to emphasize that this review is not a search for mistakes or errors, nor an attempt to impose normative judgments on the value of findings derived from various bases of inference. Discoveries based on interactions, subgroup analyses, or post-hoc exploratory analysis are likely to offer valuable insights for the field of public administration. Nor are we asserting that randomization should be equated with "causality." Indeed, high internal validity for a treatment effect based on randomization does not always ensure an understanding of the underlying mechanisms or its replicability in other contexts. Randomization, however, increasingly plays a pivotal role in establishing the internal validity of a finding. Therefore, we advocate for careful consideration of what the heterogeneity in the logic of inference in experimental studies implies for theory building and testing in our field. These insights also hold implications for the cumulative body of knowledge we draw from causal evidence and for how we continue to develop and refine norms and standards for scholarly evidence.

Methods

Our approach to reviewing the literature consisted of four stages. The first stage involved selecting the experimental articles for review. In the second stage, we identified the primary findings of each study. In the third stage, we coded each upstream finding for the underlying approach and evidence used to draw the inference. In the fourth stage, we examined how others interpreted the selected study by identifying articles that cited the original studies downstream and coded how the citing authors utilized the original findings. All coders were authors of this article and collectively have extensive experience designing and reporting causal research in public administration.

Article Selection

To identify a corpus of experimental studies, we read the abstracts of every article published in JPART from its inception in 1991 through 2020. 2020 was chosen as end-year to allow time for the studies to be cited in subsequent years, as explained below. If the abstract contained any words cuing an experimental design (e.g., "experiment," "trial," "random," or "manipulation."), we cross-checked the article's "design" section to confirm that the empirical study fit our more specific definition of experiment as a design where a researcher either randomly assigned participants to treatment conditions or randomly manipulated a factor presented to participants for the purpose of drawing causal inference. This resulted in 64 articles that comprise the population of original, "upstream" experimental articles for our analyses. Articles that cite the 64 upstream experimental articles are termed "downstream" articles, and the coding of these studies is referred to as the "downstream citation analysis," which we describe in more detail below. Collectively, the sample of upstream and downstream studies included 301 studies from 73 journals.

Two considerations were paramount in focusing on JPART as the source journal for upstream experimental findings. First, given the depth of coding required for each article, practical considerations precluded the possibility of reviewing abstracts from multiple journals within the scope of a single study. Second, since our analysis included coding interpretations drawn within downstream citations (as opposed to only within original articles), a long history of commitment to publishing experimental work that enabled widespread citation was needed. Table 1 presents the total number of experimental studies published in selected public administration journals by time period, as well as the percentage of total studies published within that time period that were experimental (similarly determined by abstracts that contained words cuing an experimental design). As Table 1 demonstrates, JPART has a long history and high density of published experimental research compared to other journals in the field.

Table 1. Number (Percent) of Experimental Studies Published in Selected Journals by Time Period

Journal	Pre-2013	2013-2014	2015-2016	2017-2018	2019-2020	Total
JPART	15(3%)	5 (7%)	8 (9%)	18 (28%)	18 (25%)	64 (8%)
PAR	0	7 (4%)	15 (9%)	18 (12%)	18 (10%)	58 (3%)
PMR	0	0	5 (4%)	8 (5%)	4 (2%)	17 (2%)
PA	4 (0%)	2 (2%)	1 (1%)	8 (7%)	8 (7%)	23 (1%)
ARPA	0	0	0	3 (3%)	1 (1%)	4 (<1%)

To ensure a broadly representative sample of downstream citations, for each upstream article, we randomly selected five downstream articles that cited it (as indicated by the Web of Science citation data in August 2022). Several recently published upstream articles had fewer than five citations, and we included all their citations for these articles. A downstream study could cite an upstream study more than once. In those cases, we randomly selected one "citation instance" within the downstream article. Eight of the 64 experimental articles we identified originated before 2003 and did not figure in Web of Science, the database we used to collect citation data. Therefore, we excluded these eight articles from the downstream citation analysis. This process resulted in a sample of 237 downstream citation instances relating to the 56 upstream articles (an average of 4.2 citations for each upstream article). Table 2 provides a distribution of those downstream citation instances across citing journals.

Table 2. Number of Downstream Citation Instances by Journal

Downstream Journal	Number of Citation Instances
Journal of Public Administration Research and Theory	34
Public Administration Review	31
Public Management Review	28
International Public Management Journal	22
Public Administration	14

Public Performance and Management Review	13
American Review of Public Administration	5
Governance	5
Journal of Public and Nonprofit Affairs	4
Local Government Studies	4
Other (3 or fewer citation instances each)	77
Total	237

Identification of Findings

Studies often report several results throughout an article. To identify the primary findings of each study, we limited ourselves to the findings presented explicitly in the article's abstract. For example, in the abstract from Teodoro and An (2018, p. 321), shown in Figure 1, the final sentence states: "We find consistent evidence that agencies' brands positively affect support for federal management, but also that partisanship conditions agencies' brand favorability." This sentence contains two findings, one corresponding to each clause. The first clause contains one finding indicating that brands of agencies affect public support. The second clause contains a second finding showing that the partisanship of survey respondents moderates the effect.

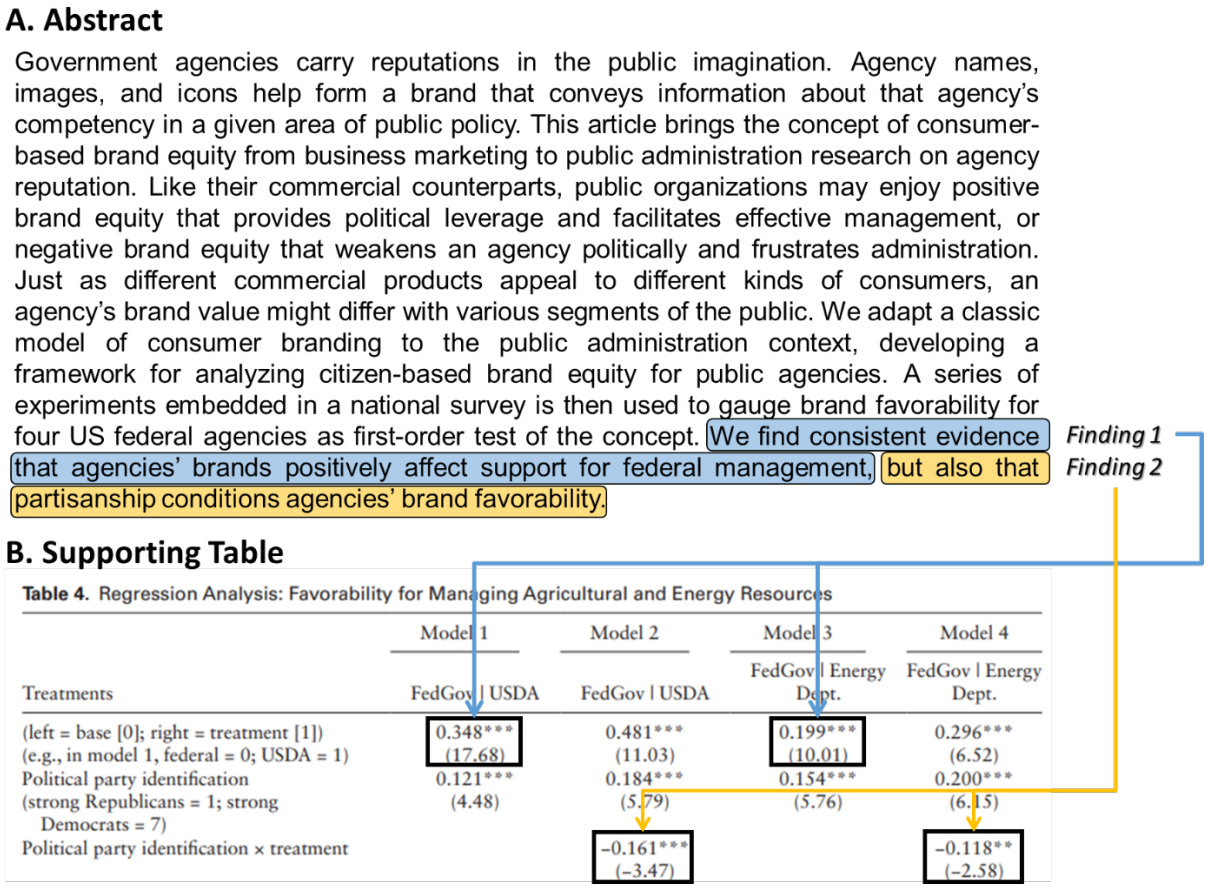

Figure 1. Representative abstract and findings

The process was not always as straightforward as the example above. Some abstracts contained findings related to multiple outcome variables in one sentence. For example, the abstract in Andersen & Hjortskov (2016, p. 647) contained the following sentence: "However, in a number of experiments we show that

perceptions of performance and satisfaction are formed in ways that are not so consistent and better explained by an intuitive mode of thinking." This sentence refers to two outcomes: performance and satisfaction. In such cases, we treated the single sentence in the abstract as providing a separate finding for each outcome.

Two articles in the early 1990s – Schwartz and Shea (1991) and Thurmaier (1992)– had no abstract. We alternatively used conclusion paragraphs summarizing the main findings for those articles. If an article contained findings from several studies, one or more of which were not experimental, we included only the findings from the experimental study (e.g., Pedersen et al. (2018)).

The next step was to map the primary finding to the results presented in the article's main body. This mapping enabled us to characterize the logic and justification underlying the finding. More specifically, we mapped each finding to a supporting result in a table or figure in the study, typically, but not always, a regression coefficient. For instance, in the Teodoro and An (2018) example in Figure 1, the coefficient on treatment in Model 1 of Table 4 (0.348) supports the first finding. A coefficient on the interaction term (political party identification × treatment) in Model 2 of Table 4 (-0.161) supports the second finding. It was possible for a single finding in the abstract to map to more than one coefficient. For example, the coefficient on treatment in Model 3 (0.199) in Figure 1 also supports the first finding. When that occurred, we would use the coefficient with the largest absolute t-value for the basis of the coding described below. We also note that in this example, the authors make it clear in the body of the article that one of their hypotheses was being tested by an interaction term that utilized a non-randomized variable collected outside of the experimental study. That was not the case in all the upstream articles we reviewed.

Coding of Findings from Upstream Experimental Studies

We coded each finding from an abstract with respect to its underlying logic of inference. The logic of inference refers to the extent to which random manipulation of a treatment or focal variable underlies a finding. It can take one of four values: random, treatment interaction, non-random, and summary.

A finding was coded as *random* when the finding's coefficient or calculation of the effect size came directly from a variable the researchers randomly assigned. We did not distinguish whether the assignment was done at the individual or group level (see Jilke et al. (2019) for group-randomized trials).

We coded a finding as *treatment interaction* when the finding relied on the product between the randomized treatment variable and another study variable that was not randomly assigned. In some instances, the non-randomized variable was a characteristic of the study participant such as gender or race, as researchers sought to identify heterogeneous treatment effects across subgroups. Cases where an interaction model was being used to draw inference about an effect for a particular subgroup (not the difference between subgroups), and the treatment variable had been randomized, were coded as random and not included in the treatment interaction category. We also coded as random cases when units were randomly assigned to cells that interact levels of two variables, as in a 2x2 factorial design.

We coded findings based on variables measured (not manipulated) before or after the experiment as *non-random.* In general, these findings include variables that are not easily manipulatable. For example, variables such as students' majors, organization type (public or private), and decision-makers' professional backgrounds or political attitudes fall into this category. A finding from the interaction of two variables, neither of which was the treatment variable, was also included in this category as opposed to the treatment interaction category.

Finally*, summary* refers to findings based on a summary judgment of multiple coefficients or findings. These could be findings based on an observed pattern of coefficients within a single model, from separate models, or the same model. For example, Jankowski, Prokop, & Tepe's (2020) abstract states, "In all three samples, hiring decisions are primarily based on meritocratic attributes." This finding is supported by comparing coefficients of several meritocratic attributes (such as working experience and school education) to those of several non-meritocratic attributes (such as gender and age) in different models estimated on each of the three samples.

Coding of Downstream Citations

To code the downstream citation instances, we extracted 100 words before and after each instance, providing contextual information about the citation instances used in the downstream study. We then used the extracted text to determine how the downstream study utilized the upstream study. The first step was determining whether a citation instance referred to a finding. A citation instance was coded as a *finding* when we interpreted the purpose of the citation as providing some kind of support or evidence for a statement made within the text block we analyzed. In contrast, we did not code a citation instance as a finding if it sought to offer more context about a topic and its importance or to acknowledge other studies on the same topic.

For citations coded as a finding, we took several additional steps. First, we evaluated whether the citation instance presented the finding as causal. We categorized a citation as *causal* if, in describing the finding, the downstream author used language that implied a cause-and-effect relationship (e.g., "caused," "impacted," "produced"). In contrast, we categorized a finding as not causal if the downstream article explicitly described it as correlational (e.g., "associated with," "corresponds to") or referred to an empirical fact, observation, or insight that did not imply a cause-and-effect relationship. Second, we attempted to map the finding in the citation instance to the appropriate finding in the upstream JPART study. We coded an additional categorical variable, *match type*, to characterize the nature of the match. We set match type to "1 to 1" if the finding in the downstream citation instance mapped to one and only one upstream finding; "1 to N" if it mapped, or drew upon, more than one upstream finding; and "no match" if did not match any of the findings we had coded from the upstream abstracts. The importance of mapping a downstream citation instance to its upstream finding is that it enables us to compare its downstream utilization or interpretation to the finding's underlying logic of inference. Third, whenever we coded the citation as causal and matched it to an upstream finding based on a treatment interaction coefficient, we examined whether the non-randomized or randomized variable was interpreted as causal. This check enabled us to distinguish between the cases where the downstream citation referred to a treatment effect disaggregated by subgroups or implied causal moderation (Bansak, 2021).

Independent of whether we coded a downstream citation instance as a finding, we also assessed whether the citation instance indicated if the upstream study was *experimental*. We coded a citation instance as "experimental" when the citation instance used language indicating that the original work came from an experimental study (e.g., "experimental," "randomly assigned," or "random trial") and as "not experimental" otherwise. If the citation instance was a finding, we made this judgment independent of whether the citation instance used causal language or not.

Six individuals coded the 237 downstream citation instances. As a training exercise, all six individuals independently coded and subsequently discussed and reconciled a batch of 10 downstream citation instances that were not used in the analysis using this scheme. Two coders were then assigned to each citation instance and coded them independently. Upon completion, we estimated the reliability of the two variables that could be coded before any further reconciliation of agreements (Landis & Koch, 1977). For the *finding* variable, pre-agreement was 83.1% (expected chance agreement = 51.3%), kappa = 0.653, SE = 0.065, $z = 10.1$, $p < 0.001$. For the *experiment* variable, pre-agreement was 95.4% (expected chance agreement = 68.9%), kappa = 0.851, SE = 0.065, $z = 13.2$, $p < 0.001$. Although the initial agreement was substantial, the coders reconciled their ratings until reaching a complete agreement with a tie-breaking vote taken among all six coders if a pair of coders could not agree. When reconciliation led to changing a non-finding to a finding, the original coder received the citation instance to further code the *causal* and *match type* variables (which only applied to downstream citation instances coded as findings). Pre-agreement for the *causal* variable (after we reconciled the findings variable) was 82.2% (expected chance agreement = 65.4%), kappa = 0.486, SE = 0.105, $z = 4.61$, $p < 0.001$. For the *match type* variable, pre-agreement was 73.1% (expected chance agreement = 48.3%), kappa = 0.480, SE = 0.076, $z = 6.30$, $p < 0.001$. We repeated the reconciliation process as necessary. The remaining analysis in this paper utilizes all reconciled codes.

Results

Upstream Experimental Study Findings

Our analysis of the upstream articles confirms that there is substantial heterogeneity concerning the logic of inference that underlies the findings reported in public administration scholarship. The 64 upstream articles contained 181 total findings. Figure 2 shows that twenty-three articles (35.9%) reported findings in their abstracts based only on a random logic of inference. Thirty-five articles (54.7%) reported findings in their abstracts from a mix of the different types of logic of inference: random, treatment interaction, non-random, and summary. Six articles (9.4%) reported findings only based on a non-random or summary logic of inference.

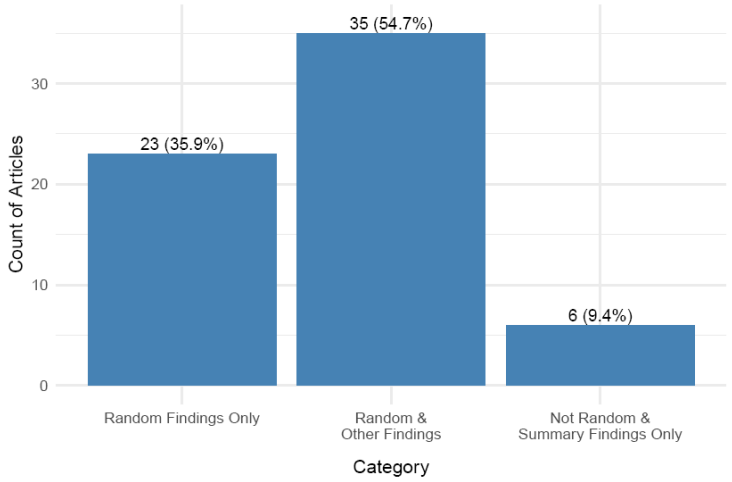

Figure 2. The Logic of Inference Finding Composition of Upstream Articles

We classified most of the 181 findings within the 64 upstream experimental studies as based on a random logic of inference. Still, almost one-quarter of the findings did not rely on variables randomly manipulated or assigned by researchers as seen in Figure 3. We classified 104 (57.5%) of the 181 findings in the upstream studies as relying on a random logic of inference; 34 (18.8%) as relying on a treatment interaction between randomly and non-randomly assigned variables; 20 (11.0%) as findings based on a variable that was not randomly assigned; and 23 (12.7%) of the findings as using a summary logic of inference.

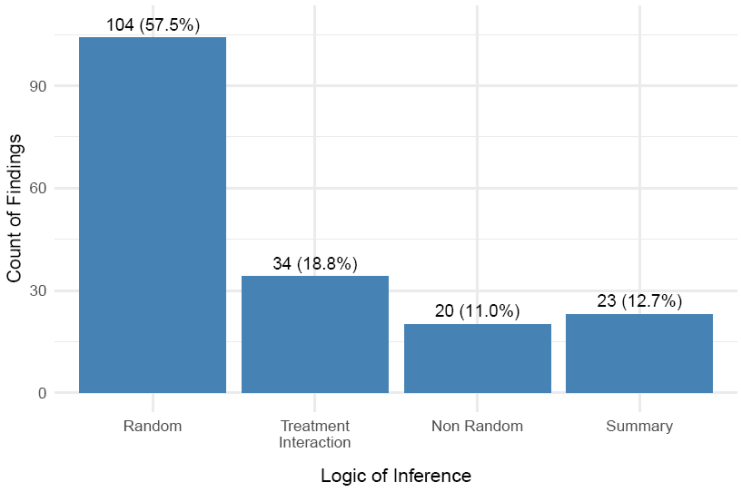

Figure 3. Classification of Upstream Findings Based on Logic of Inference

While most of these findings come from studies published in the last few years, the practice of using these different logics of inference has been established for several years. Figure 4 shows the distribution of each type of logic of inference in upstream findings over time. Given the few articles in many years, it is difficult to discern a clear pattern in the distribution, other than to note that all logics of inference appeared in studies published since 1993.

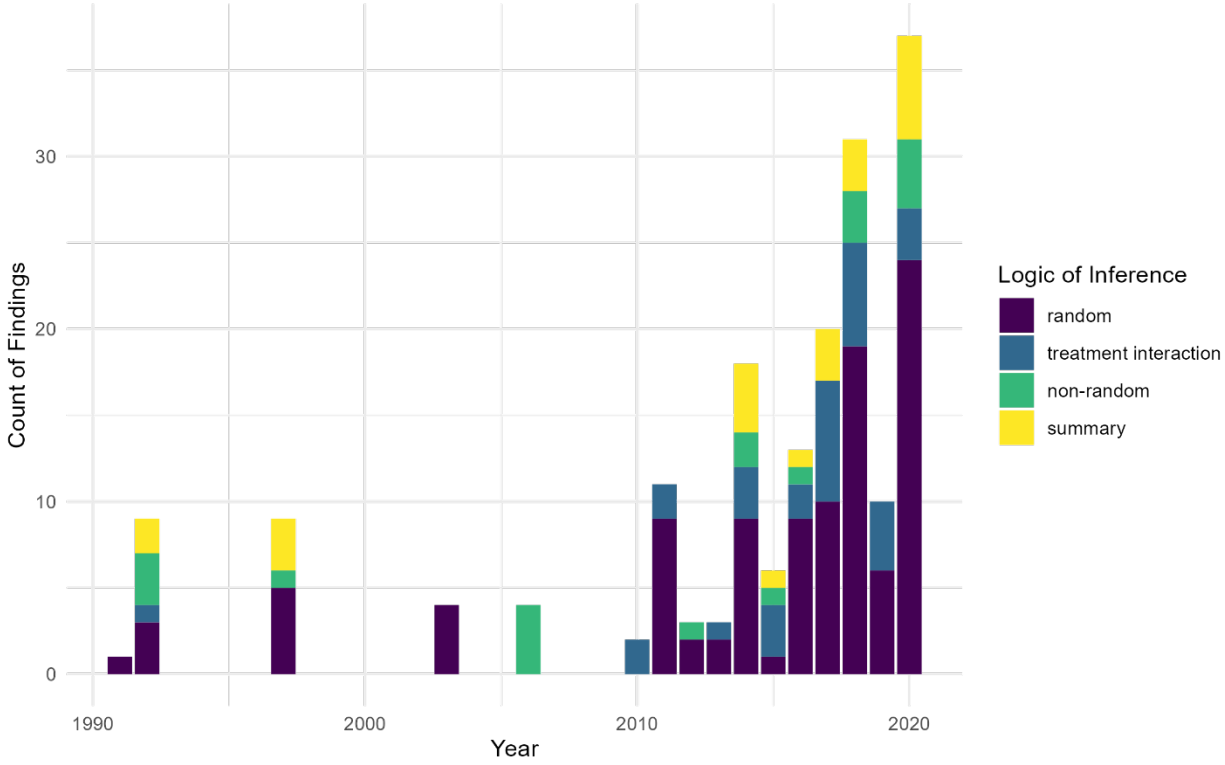

Figure 4. Trends of Findings and Articles

Downstream Citation Analysis

Having identified the primary findings from the experimental studies and coded them for their logic of inference, the next step was to examine how authors treated and interpreted those findings in the downstream studies after publication. Our goal was to assess the alignment between the logic of inference of the upstream finding and the attribution of causality in the downstream citation instance.

We first had to determine whether the author used the downstream citation instance to reference a finding in the upstream article, or to more generally reference the topic addressed or method utilized by the study (e.g., referencing the study's use of an experimental approach to examine discrimination or administrative burdens). Among the 237 downstream citation instances, 93 referenced findings. We further classified 72 of those 93 findings as "causal." We also documented whether the downstream citation instance indicated whether the original study was experimental. The majority did not. Of the 93 downstream citations referencing specific findings, 18 also explicitly indicated that the upstream study was experimental. Of the 144 downstream citations that did not refer to a specific finding, 30 explicitly indicated the upstream study was experimental. It is important to note that our evaluation of "experimental" and "causal" were independent. We did not use the reference to the study as experimental as evidence of a downstream citation instance implying a causal relationship.

The next step was to map the downstream citation to a finding in the upstream article. Of the 93 citations referencing a finding, 58 mapped to precisely one finding, and 16 mapped to more than one finding. We could not map the remaining 19 citation instances to a primary finding in the abstract of the upstream article. One of the 19 citation instances we were able to map to a finding elsewhere in the article, which we

subsequently coded as using a summary logic of inference (for the purpose of including that citation in the analysis below).

Having mapped the downstream citations to the upstream findings, we could assess the alignment of the downstream use and upstream logic of inference. To be clear, we are not assessing whether there is alignment between instances when the upstream article makes claims of causality and instances when the downstream article makes similar claims. Rather than relying on interpretation of the upstream finding by the upstream author, we are instead directly matching its underlying logic of inference to downstream interpretations of the finding.

Table 1 shows the results. Of the 72 findings interpreted as causal downstream, 37 relied on a random logic of inference, two on a non-random logic of inference, and 7 on a summary logic of inference. We could not map 16 cases coded as causal downstream to a statistical result in the upstream experimental study. That does not imply that there was no relation between the text of the downstream citation and the upstream article for these 16 citation instances, but given their lack of a connection to a specific empirical finding, we can confidently categorize them as not being rooted in random logic of inference contained within the upstream study. Additionally, 5 of the 10 causal findings based on a treatment interaction in the upstream citation instance interpreted the non-randomized variable in the interaction as causal. When those 5 are added to the 2 citation instances referencing upstream findings with a non-random logic of inference, the 7 referencing upstream findings with a summary logic of inference, and the 16 we could not map, the result is that 30 out of the 72 findings treated as causal downstream (41.6%) were not rooted in an upstream finding based on the random manipulation or assignment of the causal factor.

Table 1. Alignment of Downstream Use to Upstream Logic of Inference

Upstream Logic of Inference	Downstream Interpretation	
	Causal	Not causal
Random	37	9
Treatment interaction	10†	5
Non-random	2	4
Summary	7††	1
Cannot map to finding	16	2
Total (93 downstream citations)	72	21

†Five out of ten cases interpreted the non-randomized variable in the interaction as causal.
†† Includes one finding that could not be mapped to an upstream abstract but was located in the main body of the upstream article.

Discussion and Conclusion

One of the main reasons behind the proliferation of experiments in public administration is that they provide scholars with an unparalleled foundation for inferring causation. Assigning participants randomly to conditions researchers exogenously manipulate eliminates the risk that preexisting differences between experimental groups confound the treatment estimate (Fischer, 1925; Shadish et al., 2002). Compared to other research methods, the primary advantage of experiments lies in the internal validity achieved through using randomization as the basis for causal inference. Randomization ensures that groups, in the absence of any treatment, are expected to exhibit identical outcomes. Therefore, any differences between the groups on a particular outcome must derive from the variability introduced by the experimenter.

However, a similar amount of internal validity does not imbue all findings from experimental studies – a significant degree of heterogeneity exists in the basis of inference used across the findings of experimental articles in our field. To document and offer insight into the ramifications of this heterogeneity, we conducted a thorough analysis of experimental articles published in JPART. Concerning the composition of findings within a study, we found that 9.4% of the articles had abstracts that reported no findings involving randomization, 35.9% of experimental papers reported findings in their abstract based solely on the principle of randomization, and 54.7% presented at least one finding based on randomization, along with at least one

finding using a different logic of inference. For findings reported across all experimental upstream articles, only 57.5 % were based entirely on a logic of inference utilizing randomization. Another 18.8% relied on an interaction between randomly and non-randomly assigned variables, and a non-trivial 24% of findings reported in the abstracts of the upstream experimental articles did not relate to the random manipulation of a condition in any way.

Of course, it is not necessary, nor even advisable, that experimental studies only report insights from randomized conditions, and experimental researchers are often careful in reporting the results accordingly. Indeed, findings relying on logics of inference other than randomization can hold significant value, whether discovered within or outside the context of an experimental study. However, the danger in not highlighting the difference in the logic of inference used in experimental studies lies in the potential to contribute to a subsequent misinterpretation of those findings by other researchers. In our downstream citation analysis, we frequently encountered discrepancies between the logic of inference of a finding and how researchers interpreted it. More specifically, 41.6% of the downstream citation instances treated as causal by the citing articles were not rooted in an upstream finding based on the random assignment of the causal factor. This discrepancy suggests a potential misalignment between the logic of inference underlying results from experimental studies in public administration and the subsequent interpretation of such studies by other researchers in the field.

In recent years, there has been a notable increase in attention toward the practices employed by experimental researchers, leading to recommendations concerning preregistration, statistical power, and generalizability (Hansen & Tummers, 2020; James et al., 2017; Maroulis, 2016; Mutz, 2011). This study adds to that list a foundational but easy take-for-granted consideration–- the crucial significance of acknowledging differences in the logic of inference underlying specific findings in experimental studies. Without the randomized logic of inference, findings from experimental studies lose their fundamental narrative, which is the high internal validity gained by averaging out unobserved differences when estimating an average effect of a treatment (Angrist & Pischke, 2009). Researchers in the field of public administration must highlight and address the substantial heterogeneity in the logic of inference from the outset to fully leverage the benefits of experimentation.

Like any study, it is essential to acknowledge the limitations associated with this research. Firstly, our analysis and results rely on upstream findings from one journal. This decision was partly made for practical reasons, as coding upstream and downstream citations in such a detailed manner is an extensive undertaking. Choosing a historically impactful journal in our field that has published many experimental studies was a sensible way to circumscribe our analysis. However, norms and practices for reporting results could vary across journals, even within the same field. Future research could examine the extent to which results might vary using upstream experimental findings from other journals and publishers. Secondly, our assumption about what constitutes a primary finding in an upstream article relies on the finding appearing in the article's abstract. The citation process largely relies on reading abstracts (Jin et al., 2021), suggesting that this decision likely impacted our downstream citation analysis less than our upstream analysis. In cases where we could not find a mapping between a downstream citation and an upstream finding, we examined the entire paper for a match to a result. Thirdly, the process of coding citations for how an author is using the citation is inherently a subjective exercise and does not necessarily provide unequivocal evidence. Although we believe it highly unlikely that a different set of coders would reach qualitatively different conclusions than this study, there would likely be some discrepancies in the detailed classification of the upstream and downstream citations. Finally, it is possible that the misalignment we observe between the logic of inference of an upstream finding and characterization of the finding in downstream citations is at least in part due to a general tendency among public administration researchers to cite empirical findings as causal. Relatedly, even when researchers deliberately choose words like "associated with" or "linked to" to describe a finding, such statements may still be implicitly interpreted causally by others. As Haber and colleagues (2022) commented in their study on the use of observational and associational language in medical research, "It is likely that the rhetorical standard of 'just say association' has meant that many researchers no longer fully believe that the word 'association' just means association" (Haber et al., 2022, p. 2092). Future research could investigate this issue through a similar analysis as the one performed here, incorporating upstream articles that were exclusively non-experimental.

These limitations notwithstanding, our findings have significant implications for both experimental researchers and those citing such research. First, regardless of whether there exists a general tendency among researchers to characterize empirical findings as causal, it is crucial for experimentalists to transparently state the logic of their inference, separating findings based on randomization from those that are not. By clarifying this distinction, researchers promote transparency and enable a more informed assessment of the reliability and robustness of their findings. This practice facilitates an interpretation of their findings in future work aligned with the initial rigor of the claims. It is a fundamental prerequisite for building a robust, cumulative body of knowledge on the causal interrelationships and implications of concepts core to public administration research. The current trend towards a practice of pre-registration can help in this regard, but it is not a panacea. There will, and should, always be experimental studies utilizing multiple logics of inference to gain insight from their work.

Second, researchers who reference and cite experimental studies must be mindful of the logic of inference employed, especially when using the citation as evidence supporting a causal claim. Some experiments may present findings based on a logic of inference that differ from what researchers expect from an experimental study. As a result, it becomes imperative for researchers who aim to cite experimental findings to carefully evaluate the study in question before furthering its central claims within the field. This need highlights the importance of critically assessing the research before incorporating it into the scientific discourse.

Third, when contingent or moderating effects are of primary interest, researchers might also consider randomizing their moderating variables, enabling a causal interpretation of the moderator (Acharya et al., 2018). However, researchers must pay careful attention to number of observations available for statistical inference from an interaction. Findings that hinge on interactions between variables may require up to 16 times more statistical power than findings centered on main effects (Gelman et al., 2020). Even if the results appear to be statistically significant, weakly robust inferences can be identified using measures that quantify the sensitivity of a causal inference to changes in a sample (Frank et al., 2023).

In summary, our findings underscore the importance of explicit acknowledgment of the logic of inference when presenting results within an experimental study, and careful evaluation of the logic used when citing findings from experimental studies. They are also consistent with existing calls for better consideration of statistical power in relation to intended analyses, and explicit disclosure of preregistration and post hoc analyses.[1] Adhering to these practices will help bolster our ability to advance the field of public administration, contribute to the advancement of behavioral research, enhance the credibility of experimental findings, and facilitate the accumulation of knowledge within the public administration community.

Notes

1. Of the 64 experimental articles analyzed in this study through 2020, only 2 explicitly mentioned pre-registration and ethics approval. This is not surprising since the practice of pre-registration has only recently started to gain attention in the field of public administration. It is also very likely that more than 2 of the studies received ethics approval. All research involving human subjects at U.S. institutions is required to obtain Institutional Review Board (IRB) approval before data collection. Moreover, since most public administration research is classified as minimal risk, there has been no explicit norm or requirement to include this information in manuscripts.

References

Acharya, A., Blackwell, M., & Sen, M. (2018). Analyzing Causal Mechanisms in Survey Experiments. *Political Analysis*, *26*(4), 357–378. https://doi.org/10.1017/PAN.2018.19

Alon-Barkat, S. (2020). Can government public communications elicit undue trust? Exploring the interaction between symbols and substantive information in communications. *Journal of Public Administration Research and Theory*, *30*(1), 77–95. https://doi.org/10.1093/jopart/muz013

Alon-Barkat, S., & Gilad, S. (2017). Compensating for poor performance with promotional symbols: Evidence from a survey experiment. *Journal of Public*

Administration Research and Theory, *27*(4), 661–675. https://doi.org/10.1093/jopart/mux013

Andersen, S. C. (2017). From passive to active representation-experimental evidence on the role of normative values in shaping white and minority bureaucrats' policy attitudes. *Journal of Public Administration Research and Theory*, *27*(3), 400–414. https://doi.org/10.1093/jopart/mux006

Andersen, S. C., & Guul, T. S. (2019). Reducing minority discrimination at the front line-combined survey and field experimental evidence. *Journal of Public Administration Research and Theory*, *29*(3), 429–444. https://doi.org/10.1093/jopart/muy083

Andersen, S. C., & Hjortskov, M. (2016). Cognitive Biases in Performance Evaluations. *Journal of Public Administration Research and Theory*, *26*(4), 647–662. https://doi.org/10.1093/jopart/muv036

Andersen, S. C., & Moynihan, D. P. (2016). How leaders respond to diversity: The moderating role of organizational culture on performance information use. *Journal of Public Administration Research and Theory*, *26*(3), 448–460. https://doi.org/10.1093/jopart/muv038

Anderson, D. M., & Stritch, J. M. (2016). Goal Clarity, Task Significance, and Performance: Evidence from a Laboratory Experiment. *Journal of Public Administration Research and Theory*, *26*(2), 211–225. https://doi.org/10.1093/jopart/muv019

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Asseburg, J., Hattke, J., Hensel, D., Homberg, F., & Vogel, R. (2020). The tacit dimension of public sector attraction in multi-incentive settings. *Journal of Public Administration Research and Theory*, *30*(1), 41–59. https://doi.org/10.1093/jopart/muz004

Avellaneda, C. N. (2013). Mayoral decision-making: Issue salience, decision context, and choice constraint? An experimental study with 120 Latin American mayors. *Journal of Public Administration Research and Theory*, *23*(3), 631–661. https://doi.org/10.1093/jopart/mus041

Baekgaard, M., Baethge, C., Blom-Hansen, J., Dunlop, C. A., Esteve, M., Jakobsen, M., Kisida, B., Marvel, J., Moseley, A., Serritzlew, S., Stewart, P., Thomsen, M. K., & Wolf, P. J. (2015). Conducting Experiments in Public Management Research: A Practical Guide. *International Public Management Journal*, *18*(2), 323–342. https://doi.org/10.1080/10967494.2015.1024905

Baekgaard, M., & George, B. (2018). Equal access to the top? Representative bureaucracy and politicians' recruitment preferences for top administrative staff. *Journal of Public Administration Research and Theory*, *28*(4), 535–550. https://doi.org/10.1093/jopart/muy038

Baekgaard, M., & Serritzlew, S. (2016). Interpreting Performance Information: Motivated Reasoning or Unbiased Comprehension. *Public Administration Review*, *76*(1), 73–82. https://doi.org/10.1111/PUAR.12406

Bansak, K. (2021). Estimating Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(1), 65–86. https://doi.org/10.1111/RSSA.12614

Barrows, S., Henderson, M., Peterson, P. E., & West, M. R. (2016). Relative performance information and perceptions of public service quality: Evidence from American school districts. *Journal of Public Administration Research and Theory*, *26*(3), 571–583. https://doi.org/10.1093/jopart/muw028

Battaglio, R. P., Belardinelli, P., Bellé, N., & Cantarelli, P. (2019). Behavioral Public Administration ad fontes: A Synthesis of Research on Bounded Rationality, Cognitive Biases, and Nudging in Public Organizations. *Public Administration Review*, *79*(3), 304–320. https://doi.org/10.1111/PUAR.12994

Bellé, N. (2014). Leading to make a difference: A field experiment on the performance effects of transformational leadership, perceived social impact, and public service motivation. *Journal of Public Administration Research and Theory*, *24*(1), 109–136. https://doi.org/10.1093/jopart/mut033

Berg, M., & Johansson, T. (2020). Building institutional trust through service experiences - Private versus public provision matter. *Journal of Public Administration Research and Theory*, *30*(2), 290–306. https://doi.org/10.1093/jopart/muz029

Bretschneider, S., & Straussman, J. (1992). Statistical Laws of Confidence versus Behavioral Response: How Individuals Respond to Public Management Decisions under Uncertainty. *Journal of Public Administration Research and Theory*, *2*(3), 333–345.

Brewer, G. A. (2011). Parsing public/private differences in work motivation and performance: An experimental study. *Journal of Public Administration Research and Theory*, *21*(SUPPL. 3). https://doi.org/10.1093/jopart/mur030

Christensen, J., Dahlmann, C. M., Mathiasen, A. H., Moynihan, D. P., & Petersen, N. B. G. (2018). How do elected officials evaluate performance? Goal preferences, governance preferences, and the process of goal reprioritization. *Journal of Public Administration Research and Theory*, *28*(2), 197–211. https://doi.org/10.1093/jopart/muy001

Coursey, D. H. (1992). Information Credibility and Choosing Policy Alternatives: An Experimental Test of Cognitive-Response Theory. *Journal of Public Administration Research and Theory*, *2*(3), 315–331.

Deslatte, A. (2020). Positivity and Negativity Dominance in Citizen Assessments of Intergovernmental Sustainability Performance. *Journal of Public Administration Research and Theory*, *30*(4), 563–578. https://doi.org/10.1093/jopart/muaa004

Feeney, M. K. (2012). Organizational red tape: A measurement experiment. *Journal of Public Administration Research and Theory*, *22*(3), 427–444. https://doi.org/10.1093/jopart/mus002

Fischer, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd.

Frank, K. A., Lin, Q., Xu, R., Maroulis, S., & Mueller, A. (2023). Quantifying the robustness of causal inferences: Sensitivity analysis for pragmatic social science. *Social Science Research*, *110*, 102815. https://doi.org/10.1016/j.ssresearch.2022.102815

Gelman, A., Hill, J., & Vehtari, A. (2020). Interactions are harder to estimate than main effects. In *Regression and Other Stories*. Cambridge University Press.

Grimmelikhuijsen, S. G., & Meijer, A. J. (2014). Effects of transparency on the perceived trustworthiness of a government organization: Evidence from an online experiment. *Journal of Public Administration Research and Theory*, *24*(1), 137–157. https://doi.org/10.1093/jopart/mus048

Grimmelikhuijsen, S., Jilke, S., Olsen, A. L., & Tummers, L. (2017). Behavioral Public Administration: Combining Insights from Public Administration and Psychology. *Public Administration Review*, *77*(1), 45–56. https://doi.org/10.1111/PUAR.12609

Haber, N. A., Wieten, S. E., Rohrer, J. M., Arah, O. A., Tennant, P. W. G., Stuart, E. A., Murray, E. J., Pilleron, S., Lam, S. T., Riederer, E., Howcutt, S. J., Simmons, A. E., Leyrat, C., Schoenegger, P., Booman, A., Dufour, M.-S. K., O'Donoghue, A. L., Baglini, R., Do, S., … Fox, M. P. (2022). Causal and Associational Language in Observational Health Research: A Systematic Evaluation. *American Journal of Epidemiology*, 191(12), 2084–2097. https://doi.org/10.1093/aje/kwac137

Hansen, J. A., & Tummers, L. (2020). A Systematic Review of Field Experiments in Public Administration. *Public Administration Review*, *80*(6), 921–931. https://doi.org/10.1111/PUAR.13181

Herian, M. N., Hamm, J. A., Tomkins, A. J., & Pytlik Zillig, L. M. (2012). Public participation, procedural Fairness, and evaluations of local governance: The moderating role of uncertainty. *Journal of Public Administration Research and Theory*, *22*(4), 815–840. https://doi.org/10.1093/jopart/mur064

Hvidman, U. (2019). Citizens' Evaluations of the Public Sector: Evidence from Two Large-Scale Experiments. *Journal of Public Administration Research and Theory*, *29*(2), 255–267. https://doi.org/10.1093/jopart/muy064

Jacobsen, C. B., Andersen, L. B., Bøllingtoft, A., & Eriksen, T. L. M. (2022). Can Leadership Training Improve Organizational Effectiveness? Evidence from a Randomized Field Experiment on Transformational and Transactional Leadership. *Public Administration Review*, *82*(1), 117–131. https://doi.org/10.1111/PUAR.13356

Jakobsen, M. (2013). Can Government Initiatives Increase Citizen Coproduction? Results of a Randomized Field Experiment. *Journal of Public Administration Research and Theory*, *23*(1), 27–54. https://doi.org/10.1093/jopart/mus036

Jakobsen, M., Jacobsen, C. B., & Serritzlew, S. (2019). Managing the Behavior of Public Frontline Employees through Change-Oriented Training: Evidence from a Randomized Field Experiment. *Journal of Public Administration Research and Theory*, *29*(4), 556–571. https://doi.org/10.1093/jopart/muy080

James, O. (2011). Performance measures and democracy: Information effects on citizens in field and laboratory experiments. *Journal of Public Administration Research and Theory*, *21*(3), 399–418. https://doi.org/10.1093/jopart/muq057

James, O., Jilke, S. R., & Van Ryzin, G. G. (2017). *Experiments in Public Management Research*. Cambridge University Press. https://doi.org/10.1017/9781316676912

James, O., & Van Ryzin, G. G. (2017). Motivated reasoning about public performance: An experimental study of how citizens judge the affordable care act. *Journal of Public Administration Research and Theory*, *27*(1), 197–209. https://doi.org/10.1093/jopart/muw049

Jankowski, M., Prokop, C., & Tepe, M. (2020). Representative Bureaucracy and Public Hiring Preferences: Evidence from a Conjoint Experiment among German Municipal Civil Servants and Private Sector Employees. *Journal of Public Administration Research and Theory*, *30*(4), 596–618. https://doi.org/10.1093/jopart/muaa012

Jensen, D. C., & Pedersen, L. B. (2017). The impact of empathy-explaining diversity in street-level decision-making. *Journal of Public Administration Research and Theory*, *27*(3), 433–449. https://doi.org/10.1093/jopart/muw070

Jensen, U. T., Andersen, L. B., & Jacobsen, C. B. (2019). Only When We Agree! How Value Congruence Moderates the Impact of Goal-Oriented Leadership on Public Service Motivation. *Public Administration Review*, *79*(1), 12–24. https://doi.org/https://doi.org/10.1111/puar.13008

Jilke, S., & Baekgaard, M. (2020). The political psychology of citizen satisfaction: Does functional responsibility matter? *Journal of Public Administration Research*

and Theory, *30*(1), 130–143. https://doi.org/10.1093/jopart/muz012

Jilke, S., Lu, J., Xu, C., & Shinohara, S. (2019). Using Large-Scale Social Media Experiments in Public Administration: Assessing Charitable Consequences of Government Funding of Nonprofits. *Journal of Public Administration Research and Theory*, *29*(4), 627–639. https://doi.org/10.1093/jopart/muy021

Jilke, S., & Tummers, L. (2018). Which clients are deserving of help? A theoretical model and experimental test. *Journal of Public Administration Research and Theory*, *28*(2), 226–238. https://doi.org/10.1093/jopart/muy002

Jilke, S., Van Dooren, W., & Rys, S. (2018). Discrimination and administrative burden in public service markets: Does a public-private difference exist? *Journal of Public Administration Research and Theory*, *28*(3), 423–439. https://doi.org/10.1093/jopart/muy009

Jilke, S., Van Ryzin, G. G., & Van De Walle, S. (2016). Responses to decline in marketized public services: An experimental evaluation of choice overload. *Journal of Public Administration Research and Theory*, *26*(3), 421–432. https://doi.org/10.1093/jopart/muv021

Jin, T., Duan, H., Lu, X., Ni, J., & Guo, K. (2021). Do research articles with more readable abstracts receive higher online attention? Evidence from Science. *Scientometrics*, *126*(10), 8471–8490. https://doi.org/10.1007/S11192-021-04112-9/TABLES/7

Knott, J. H. (2003). Adaptive Incrementalism and Complexity: Experiments with Two-Person Cooperative Signaling Games. *Journal of Public Administration Research and Theory*, *13*(3), 341–365. https://doi.org/10.1093/jopart/mug023

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Landsbergen, D., Bozeman, B., & Bretschneider, S. (1992). "Internal Rationality" and the Effects of Perceived Decision Difficulty: Results of a Public Management Decisionmaking Experiment. *Journal of Public Administration Research and Theory: J-PART*, *2*(3), 247–264.

Landsbergen, D., Coursey, D. H., Loveless, S., & Shangraw, R. F. (1997). Decision Quality, Confidence, and Commitment with Expert Systems: An Experimental Study. *Journal of Public Administration Research and Theory*, *7*, 1.

Linos, E. (2018). More than public service: A field experiment on job advertisements and diversity in the police. *Journal of Public Administration Research and Theory*, *28*(1), 67–85. https://doi.org/10.1093/jopart/mux032

Linos, E., Ruffini, K., & Wilcoxen, S. (2022). Reducing Burnout and Resignations among Frontline Workers: A Field Experiment. *Journal of Public Administration Research and Theory*, *32*(3), 473–488. https://doi.org/10.1093/JOPART/MUAB042

Maroulis, S. (2016). Interpreting School Choice Treatment Effects: Results and Implications from Computational Experiments. *Journal of Artificial Societies and Social Simulation*, *19*(1). https://doi.org/10.18564/jasss.3002

Marvel, J. D. (2016). Unconscious Bias in Citizens Evaluations of Public Sector Performance. *Journal of Public Administration Research and Theory*, *26*(1), 143–158. https://doi.org/10.1093/jopart/muu053

Meyer-Sahling, J. H., Mikkelsen, K. S., & Schuster, C. (2019). The causal effect of public service motivation on ethical behavior in the public sector: Evidence from a large-scale survey experiment. *Journal of Public Administration Research and Theory*, *29*(3), 445–459. https://doi.org/10.1093/jopart/muy071

Mutz, D. C. (2011). *Population-Based Survey Experiments*. Princeton University Press. https://doi.org/10.1515/9781400840489

Nielsen, P. A., & Baekgaard, M. (2015). Performance information, blame avoidance, and politicians' attitudes to spending and reform: Evidence from an experiment. *Journal of Public Administration Research and Theory*, *25*(2), 545–569. https://doi.org/10.1093/jopart/mut051

Nielsen, P. A., & Moynihan, D. P. (2017). How do politicians attribute bureaucratic responsibility for performance? Negativity bias and interest group advocacy. *Journal of Public Administration Research and Theory*, *27*(2), 269–283. https://doi.org/10.1093/jopart/muw060

Nutt, P. C. (2006). Comparing public and private sector decision-making practices. *Journal of Public Administration Research and Theory*, *16*(2), 289–318. https://doi.org/10.1093/jopart/mui041

Olsen, A. L. (2017). Compared to what? How social and historical reference points affect citizens' performance evaluations. *Journal of Public Administration Research and Theory*, *27*(4), 562–580. https://doi.org/10.1093/jopart/mux023

Pedersen, M. J., Stritch, J. M., & Thuesen, F. (2018). Punishment on the frontlines of public service delivery: Client ethnicity and caseworker sanctioning decisions in a Scandinavian welfare state. *Journal of Public Administration Research and Theory*, *28*(3), 339–354. https://doi.org/10.1093/jopart/muy018

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

Petersen, N. B. G., Laumann, T. V., & Jakobsen, M. (2019). Acceptance or disapproval: Performance information in the eyes of public frontline employees. *Journal of Public Administration Research and Theory*, *29*(1), 101–117. https://doi.org/10.1093/jopart/muy035

Porter, E., & Rogowski, J. C. (2018). Partisanship, Bureaucratic Responsiveness, and Election Administration: Evidence from a Field Experiment. *Journal of Public Administration Research and Theory*, *28*(4), 602–617. https://doi.org/10.1093/jopart/muy025

Riccucci, N. M., Van Ryzin, G. G., & Jackson, K. (2018). Representative Bureaucracy, Race, and Policing: A Survey Experiment. *Journal of Public Administration Research and Theory*, *28*(4), 506–518. https://doi.org/10.1093/jopart/muy023

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

Ruder, A. I., & Woods, N. D. (2020). Procedural fairness and the legitimacy of agency rulemaking. *Journal of Public Administration Research and Theory*, *30*(3), 400–414. https://doi.org/10.1093/jopart/muz017

Schwartz-Shea, P. (1991). Understanding Subgroup Optimization: Experimental Evidence on Individual Choice and. *Journal of Public Administration Research and Theory*, *1*(1), 49–74.

Scott, P. G. (1997). Assessing Determinants of Bureaucratic Discretion: An Experiment in Street-Level Decision Making. *Journal of Public Administration Research and Theory*, 35–57.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2nd ed.). Cengage Learning.

Silvia, C. (2018). Picking the team: A preliminary experimental study of the activation of collaborative network members. *Journal of Public Administration Research and Theory*, *28*(1), 120–137. https://doi.org/10.1093/jopart/mux026

Teodoro, M. P., & An, S. H. (2018). Citizen-based brand equity: A model and experimental evaluation. *Journal of Public Administration Research and Theory*, *28*(3), 321–338. https://doi.org/10.1093/jopart/mux044

Tepe, M., & Prokop, C. (2018). Are future bureaucrats more risk averse? The effect of studying public administration and PSM on risk preferences. *Journal of Public Administration Research and Theory*, *28*(2), 182–196. https://doi.org/10.1093/jopart/muy007

Thomsen, M. K., Baekgaard, M., & Jensen, U. T. (2020). The Psychological Costs of Citizen Coproduction. *Journal of Public Administration Research and Theory*, *30*(4), 656–673. https://doi.org/10.1093/jopart/muaa001

Thomsen, M. K., & Jensen, U. T. (2020). Service professionals' response to volunteer involvement in service production. *Journal of Public Administration Research and Theory*, *30*(2), 220–239. https://doi.org/10.1093/jopart/muz028

Thurmaier, K. (1992). Budgetary Decisionmaking in Central Budget Bureaus: An Experiment. *Journal of Public Administration Research and Theory*, *2*(4), 463–487.

Valant, J., & Newark, D. A. (2020). The Word on the Street or the Number from the State? Government-Provided Information and Americans' Opinions of Schools. *Journal of Public Administration Research and Theory*, *30*(4), 674–692. https://doi.org/10.1093/jopart/muaa010

Vogel, D., & Willems, J. (2020). The effects of making public service employees aware of their prosocial and societal impact: A microintervention. *Journal of Public Administration Research and Theory*, *30*(3), 485–503. https://doi.org/10.1093/jopart/muz044

Vogel, D., & Xu, C. (2021). Everything hacked? What is the evidential value of the experimental public administration literature? *Journal of Behavioral Public Administration*, *4*(2). https://doi.org/10.30636/JBPA.42.239

Weibel, A., Rost, K., & Osterloh, M. (2010). Pay for performance in the public sector - Benefits and (Hidden) costs. *Journal of Public Administration Research and Theory*, *20*(2), 387–412. https://doi.org/10.1093/jopart/mup009

Wittmer, D. (1992). Ethical Sensitivity and Managerial Decisionmaking: An Experiment. *Journal of Public Administration Research and Theory*, *2*(4), 443–462.

Worthy, B., John, P., & Vannoni, M. (2017). Transparency at the parish pump: A field experiment to measure the effectiveness of Freedom of Information requests in England. *Journal of Public Administration Research and Theory*, *27*(3), 485–500. https://doi.org/10.1093/jopart/muw063

Yackee, S. W. (2015). Participant voice in the bureaucratic policymaking process. *Journal of Public Administration Research and Theory*, *25*(2), 427–449. https://doi.org/10.1093/jopart/muu007.

Appendix

Table A1. List of experimental studies included in the upstream analysis

	Authors / Year	Title
1	Asseburg et al., 2020	The Tacit Dimension of Public Sector Attraction in Multi-Incentive Settings
2	Alon-Barkat, 2020	Can Government Public Communications Elicit Undue Trust? Exploring the Interaction between Symbols and Substantive Information in Communications
3	Thomsen & Jensen, 2020	Service Professionals' Response to Volunteer Involvement in Service Production
4	Berg & Johansson, 2020	Building Institutional Trust Through Service Experiences—Private Versus Public Provision Matter
5	Ruder & Woods, 2020	Procedural Fairness and the Legitimacy of Agency Rulemaking
6	Vogel & Willems, 2020	The Effects of Making Public Service Employees Aware of Their Prosocial and Societal Impact: A Microintervention
7	Jilke & Baekgaard, 2020	The Political Psychology of Citizen Satisfaction: Does Functional Responsibility Matter?
8	Petersen, 2020	Whoever Has Will be Given More: The Effect of Performance Information on Frontline Employees' Support for Managerial Policy Initiatives
9	Deslatte, 2020	Positivity and Negativity Dominance in Citizen Assessments of Intergovernmental Sustainability Performance
10	Jankowski et al., 2020	Representative Bureaucracy and Public Hiring Preferences: Evidence from a Conjoint Experiment among German Municipal Civil Servants and Private Sector Employees
11	Thomsen et al., 2020	The Psychological Costs of Citizen Coproduction
12	Valant & Newark, 2020	The Word on the Street or the Number from the State? Government-Provided Information and Americans' Opinions of Schools
13	Hvidman, 2019	Citizens' Evaluations of the Public Sector: Evidence From Two Large-Scale Experiments
14	Andersen & Guul, 2019	Reducing Minority Discrimination at the Front Line—Combined Survey and Field Experimental Evidence
15	Meyer-Sahling et al., 2019	The Causal Effect of Public Service Motivation on Ethical Behavior in the Public Sector: Evidence from a Large-Scale Survey Experiment

16	Jakobsen et al., 2019	Managing the Behavior of Public Frontline Employees through Change-Oriented Training: Evidence from a Randomized Field Experiment
17	Jilke et al., 2019	Using Large-Scale Social Media Experiments in Public Administration: Assessing Charitable Consequences of Government Funding of Nonprofits
18	Petersen et al., 2019	Acceptance or Disapproval: Performance Information in the Eyes of Public Frontline Employees
19	Linos, 2018	More Than Public Service: A Field Experiment on Job Advertisements and Diversity in the Police
20	Christensen et al., 2018	How Do Elected Officials Evaluate Performance? Goal Preferences, Governance Preferences, and the Process of Goal Reprioritization
21	Jilke et al., 2018	Discrimination and Administrative Burden in Public Service Markets: Does a Public–Private Difference Exist?
22	Baekgaard & George, 2018	Equal Access to the Top? Representative Bureaucracy and Politicians' Recruitment Preferences for Top Administrative Staff
23	Silvia, 2018	Picking the Team: A Preliminary Experimental Study of the Activation of Collaborative Network Members
24	Tepe & Prokop, 2018	Are Future Bureaucrats More Risk Averse? The Effect of Studying Public Administration and PSM on Risk Preferences
25	Jilke & Tummers, 2018	Which Clients are Deserving of Help? A Theoretical Model and Experimental Test
26	Teodoro & An, 2018	Citizen-Based Brand Equity: A Model and Experimental Evaluation
27	Riccucci et al., 2018	Representative Bureaucracy, Race, and Policing: A Survey Experiment
28	Pedersen et al., 2018	Punishment on the Frontlines of Public Service Delivery: Client Ethnicity and Caseworker Sanctioning Decisions in a Scandinavian Welfare State
29	Porter & Rogowski, 2018	Partisanship, Bureaucratic Responsiveness, and Election Administration: Evidence from a Field Experiment
30	Nielsen & Moynihan, 2017	How Do Politicians Attribute Bureaucratic Responsibility for Performance? Negativity Bias and Interest Group Advocacy
31	Andersen, 2017	From Passive to Active Representation—Experimental Evidence on the Role of Normative Values in Shaping White and Minority Bureaucrats' Policy Attitudes
32	Jensen & Pedersen, 2017	The Impact of Empathy—Explaining Diversity in Street-Level Decision-Making

33	Worthy et al., 2017	Transparency at the Parish Pump: A Field Experiment to Measure the Effectiveness of Freedom of Information Requests in England
34	Alon-Barkat & Gilad, 2017	Compensating for Poor Performance with Promotional Symbols: Evidence from a Survey Experiment
35	Olsen, 2017	Compared to What? How Social and Historical Reference Points Affect Citizens' Performance Evaluations
36	James & Van Ryzin, 2017	Motivated Reasoning about Public Performance: An Experimental Study of How Citizens Judge the Affordable Care Act
37	Marvel, 2016	Unconscious Bias in Citizens' Evaluations of Public Sector Performance
38	Andersen & Moynihan, 2016	How Leaders Respond to Diversity: The Moderating Role of Organizational Culture on Performance Information Use
39	Barrows et al., 2016	Relative Performance Information and Perceptions of Public Service Quality: Evidence From American School Districts
40	Andersen & Hjortskov, 2016	Cognitive Biases in Performance Evaluations
41	Anderson & Stritch, 2016	Goal Clarity, Task Significance, and Performance: Evidence From a Laboratory Experiment
42	Jilke et al., 2016	Responses to Decline in Marketized Public Services: An Experimental Evaluation of Choice Overload
43	Yackee, 2015	Participant Voice in the Bureaucratic Policymaking Process
44	Nielsen & Baekgaard, 2015	Performance Information, Blame Avoidance, and Politicians' Attitudes to Spending and Reform: Evidence from an Experiment
45	Grimmelikhuijsen & Meijer, 2014	Effects of Transparency on the Perceived Trustworthiness of a Government Organization: Evidence from an Online Experiment
46	Bellé, 2014	Leading to Make a Difference: A Field Experiment on the Performance Effects of Transformational Leadership, Perceived Social Impact, and Public Service Motivation
47	Riccucci et al., 2014	Representative Bureaucracy in Policing: Does It Increase Perceived Legitimacy?
48	Jakobsen, 2013	Can Government Initiatives Increase Citizen Coproduction? Results of a Randomized Field Experiment
49	Avellaneda, 2013	Mayoral Decision-Making: Issue Salience, Decision Context, and Choice Constraint? An Experimental Study with 120 Latin American Mayors
50	Feeney, 2012	Organizational Red Tape: A Measurement Experiment

51	Herian et al., 2012	Public Participation, Procedural Fairness, and Evaluations of Local Governance: The Moderating Role of Uncertainty
52	James, 2011	Performance Measures and Democracy: Information Effects on Citizens in Field and Laboratory Experiments
53	Brewer, 2011	Parsing Public/Private Differences in Work Motivation and Performance: An Experimental Study
54	Weibel et al., 2010	Pay for Performance in the Public Sector—Benefits and Hidden Costs
55	Nutt, 2006	Comparing Public and Private Sector Decision-Making Practices
56	Knott, 2003	Adaptive Incrementalism and Complexity: Experiments with Two-Person Cooperative Signaling Games
57	Landsbergen et al., 1997	Decision Quality, Confidence, and Commitment with Expert Systems: An Experimental Study
58	Scott, 1997	Assessing Determinants of Bureaucratic Discretion: An Experiment in Street-Level Decision Making
59	Thurmaier, 1992	Budgetary Decisionmaking in Central Budget Bureaus: An Experiment
60	Bretschneider & Straussman, 1992	Statistical Laws of Confidence versus Behavioral Response: How Individuals Respond to Public Management Decisions under Uncertainty
61	Wittmer, 1992	Ethical Sensitivity and Managerial Decisionmaking: An Experiment
62	Landsbergen et al., 1992	"Internal Rationality" and the Effects of Perceived Decision Difficulty: Results of a Public Management Decisionmaking Experiment
63	Coursey, 1992	Information Credibility and Choosing Policy Alternatives: An Experimental Test of Cognitive-Response Theory
64	Schwartz-Shea, 1991	Understanding Subgroup Optimization: Experimental Evidence on Individual Choice and Group Processes